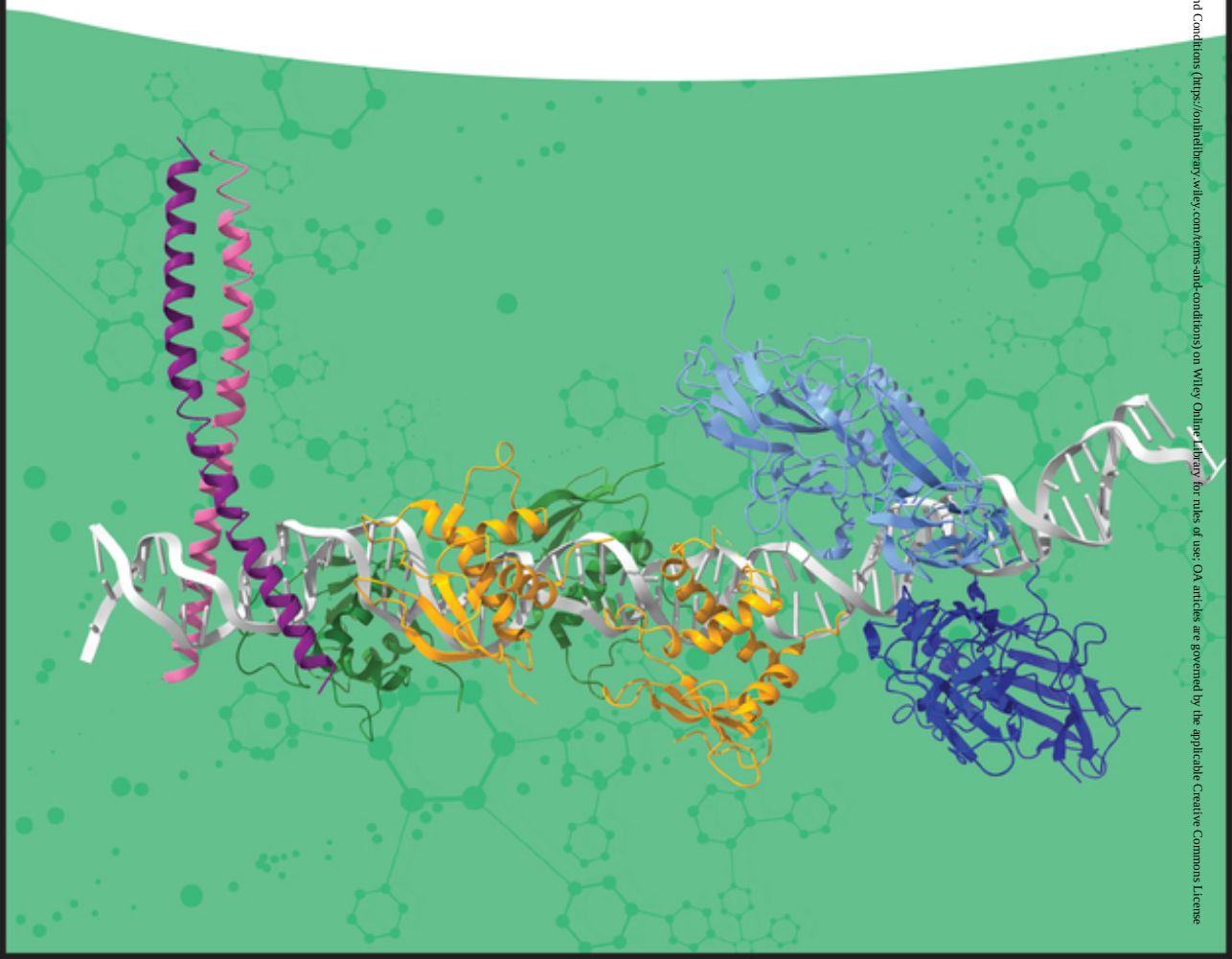
WILEY-VCH

Edited by
Volkhard Helms and Olga V. Kalinina

Protein Interactions

The Molecular Basis of Interactomics



Protein Interactions

Protein Interactions

The Molecular Basis of Interactomics

Edited by Volkhard Helms and Olga V. Kalinina

WILEY-VCH

Editors

Prof. Volkhard Helms
Saarland University
Center for Bioinformatics
Saarbrücken
Germany

Prof. Olga V. Kalinina
Helmholtz Institute for Pharmaceutical Research Saarland (HIPS) / Helmholtz Centre for Infection Research (HZI);
Medical Faculty, Saarland University; and
Center for Bioinformatics,
Saarland University
Saarbrücken
Germany

Cover Image:
Foreground image © Volkhard Helms
Background © Shutterstock

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

Print ISBN: 978-3-527-34864-0
ePDF ISBN: 978-3-527-83051-0
ePub ISBN: 978-3-527-83052-7
oBook ISBN: 978-3-527-83050-3

Typesetting Straive, Chennai, India

Contents

Preface

Proteins are biochemical machines that participate in virtually all key processes of biological systems. For example, enzymes catalyze biochemical reactions, and hence control the metabolic state of the cell. Transcription factors guide RNA polymerase to read off the needed parts of the genome. Initiation factors tell ribosomes what mRNAs to process. Receptors in the cell membrane sense signals from the outside. Transporters and channel proteins mediate exchange of substances across the cell and organelle membranes, etc. In these processes, proteins exert their biological function by interacting with other proteins, nucleic acids, membranes, low-molecular-weight ligands such as substrates and drugs, and so forth. In essence, all these interactions are governed by nonbonded interactions between protein atoms (backbone as well as side-chain atoms) and atoms of the corresponding interaction partners, whereby the scale of these interactions differs from a handful of involved atoms for low-molecular-weight ligands to thousands of atoms in large protein complexes. Consequently, methods that allow studying and predicting these interactions differ in scale.

The research field that discovers and analyzes this myriad of interactions is termed interactomics and builds on contributions from experiments and computation. The technologies used in this field are constantly being refined, but still lag somehow behind the level at which individual pairwise interactions can be resolved and predicted in terms of three-dimensional structure, binding thermodynamics, specificity, etc. One important aspect of interactomics is to integrate data from different sources. Practically ignored so far are the effects of post-translational modifications and alternative splicing on interactomics (which are addressed in detail in this book, see below). It is the aim of this book to capture the state-of-art of cellular interactomics involving proteins and to describe existing technical and conceptual challenges that need to be overcome in the future.

This book presents an overview of protein interactions, experimental techniques and findings, computational tools and resources that have been developed to study them. In its first part, we introduce the molecular basics of protein structure (Chapter 1) and properties of protein–protein binding interfaces (Chapter 2). Recently, new-generation sequencing methods yielded large amounts of protein sequence data that can also be leveraged to predict protein interactions, in particular, pairwise protein–protein interactions (Chapter 3). In parallel, classical methods of protein–protein docking (reconstruction of pairwise protein complexes from isolated structures of their components) have matured and nowadays successfully

manage to incorporate additional experimental constraints and may even account for protein conformational changes (Chapter 4). A systemic view of protein pairwise interactions is provided by protein interaction networks, which can comprise both experimentally resolved and computationally predicted interactions (Chapter 5).

Large protein complexes present an additional challenge due to numerous ways in which individual protomers can interact with each other. Such complexes can be extracted from protein interaction networks (Chapter 5) or predicted in a combinatorial fashion or using experimental constraints (Chapter 6). Systematic integration of many experimental constraints with structural data can provide exciting insights into the structure and evolution of very large and complex protein assemblies (Chapter 7).

The physics of protein interactions with different partners takes place at different scales, since the size of the partners and hence the number of individual non-covalent interactions differ considerably. The second part of this book analyzes these different interactions and ways to model them in detail. We start with computational techniques to examine the kinetics and thermodynamics of interactions between pairs of proteins (Chapter 8), followed by a chapter on Markov-state models that statistically evaluate all transitions along association and dissociation pathways (Chapter 9). We continue with protein–DNA interactions exemplified by transcription factor binding to DNA (Chapter 10) and chromatin (Chapter 11), followed by a chapter on the emerging field of protein–RNA interactions, e.g. during the preprocessing stage of pre-mRNA and with noncoding RNAs (Chapter 12). As many signaling and transport processes involve cellular membranes, protein–membrane interactions are then covered in Chapter 13, followed by a discussion of how proteins interact with low-molecular-weight ligands such as drugs (Chapter 14).

All these different kinds and instances of protein interactions crucially contribute to the flow of matter and information in living cells. If such interactions are modulated, this can obviously alter many cellular processes. In the third part of this book, three important types of modulating effects are addressed, namely the effects of genetic mutations (Chapter 15), of alternative splicing (Chapter 16), and those of posttranslational modifications (Chapter 17). There, the main focus is again placed on how protein–protein interactions are affected. The impact of these types of protein alterations on other types of interactions (e.g. with small molecules) is less well understood, although prominent examples, such as drug resistance-associated mutations, exist. Computational methods for systematic assessment of such changes are still to be developed.

The individual chapters were written by experts in their fields, and we are extremely grateful to them for their time and effort they invested in this. We hope that this book paints a complex, but versatile and instructive picture of all different kinds of interactions that proteins engage in. Interactomics, building on combined experimental and computational work, is an emerging discipline that bears great promise to better understand the molecular mechanisms of life. In our view, protein interactions hold the key to it.

Volkhard Helms
Olga V. Kalinina

1

Protein Structure and Conformational Dynamics

Volkhard Helms

Saarland University, Center for Bioinformatics, Saarland Informatics Campus, Postfach 15 11 50, 66041 Saarbrücken, Germany

1.1 Structural and Hierarchical Aspects

1.1.1 Size of Proteins

The size of proteins ranges from very small proteins, such as the 20-amino acid miniprotein Trp cage, to the largest protein in the human body, titin, which consists of about 27 000 amino acids and has a molecular weight of 3 million Dalton. Generally, when speaking of typical proteins, we refer to compact proteins of about 80 to 500 amino acids (residues) in size. Tiessen et al. reported that archaeal proteins had the smallest average size (283 aa), followed by bacterial proteins (320 aa) and eukaryotic proteins (472 aa) [1]. Among eukaryotes, plant proteins (392 aa) had a smaller size, whereas animal proteins (486 aa) and proteins from fungi (487 aa) were larger.

1.1.2 Protein Domains

The larger a single protein gets, the higher is the chance that it will be composed of multiple structurally distinct "domains." These are typically sequential parts of the protein sequence with a characteristic length between 100 and 200 amino acids [2]. For example, the protein Src kinase consists of an SH3 domain (that binds to proline-rich peptides), an SH2 domain (that binds to phosphorylated tyrosine residues), and the catalytic kinase domain, see Figure 1.1. In the inactive state, the SH3 domain will hold on to the linker connecting SH2 and catalytic domain that contains several prolines, and the SH2 domain will hold on to a phosphorylated tyrosine in the C-terminal tail of the catalytic domain. Thereby, all three domains are locked in a conformationally restricted state. Once activated by dephosphorylation of the tyrosine, these contacts are released, and the catalytic domain can undergo the characteristic Pacman-type opening/closing motion of protein kinases, enabling the binding of adenosine triphosphate (ATP). In the closed conformation, the active site residues catalyze transfer of the terminal γ-phosphate of ATP to a nearby tyrosine of a substrate protein bound on the Src kinase surface. The catalytic

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

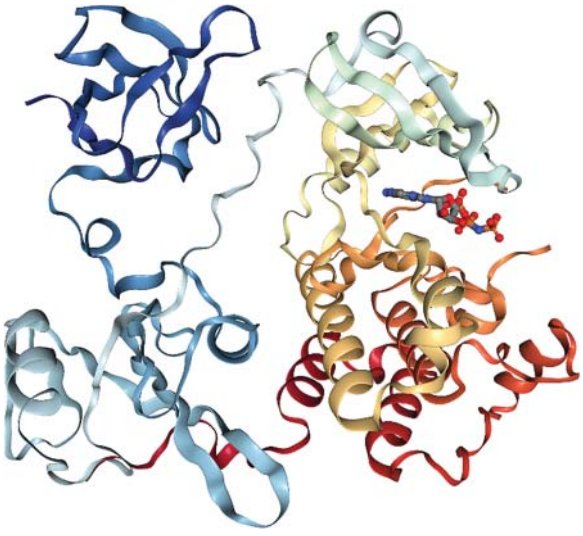


Figure 1.1 X-ray structure (PDB code 1AD5) of human Src kinase. The peptide sequence starts with an SH3 domain (top left), followed by an SH2 domain (bottom left) and then leads to the catalytic kinase domain (right). ATP is bound between small (top) and large lobe (bottom) of the kinase domain. Source: Figure generated with NGL viewer.

domain of kinases itself consists of two domain-like "lobes," a smaller N-terminal lobe (of about 80 aa) and a larger C-terminal lobe (of about 180 aa).

Although multi-domain proteins exist in all life forms, more complex organisms (having a larger number of unique cell types) contain more unique domains and a larger fraction of multi-domain proteins: eukaryotes have more multi-domain proteins than prokaryotes, and animals have more multi-domain proteins than unicellular eukaryotes [3].

1.1.3 Protein Composition

The composition of a protein depends on its environment and its posttranslational modifications, such as phosphorylation and sumoylation. For example, extracellular domains of most cell membrane proteins are often extensively glycosylated. Here, we will focus on the varying mixture of the 20 commonly occurring amino acids that make up most of all existing proteins. Water-soluble proteins possess a rather hydrophobic core and a polar surface that is in contact with the cytoplasm. This clear organizational principle provides the main driving force for the folding of water-soluble domains via the "hydrophobic effect."

Prokaryotic proteins contain more than 10% of leucine and about 9% of alanine residues, but rather few (only 1–2%) cysteine, tryptophan, histidine, and methionine residues [4]. Brüne et al. compared the amino acid composition of prokaryotic and eukaryotic proteins [5]. Eukaryotes have the highest variability for proline, cysteine, and asparagine. Amino acids showing high variability across species are lysine, alanine, and isoleucine, whereas histidine, tryptophan, and methionine vary the least. Cysteine is more common in eukaryotes than in archaea and bacteria, whereas isoleucine is less abundant in eukaryotes. The authors also analyzed the differential usage of amino acids in domains and linkers. Proline and glutamine, but to a smaller extent, polar and charged amino acids, are more common in linkers

that are rather exposed to surrounding water. Globular domains contain larger fractions of hydrophobic amino acids, such as leucine and valine, and aromatic ones, such as phenylalanine and tyrosine.

1.1.4 Secondary Structure Elements

Folded proteins contain two types of secondary structure elements, α-helices and β-sheets. α-Helices have lengths between 9 and 37 residues with a peak at 11 amino acids [6]. β-Sheets are considerably shorter, being 2–17 residues long with a peak at 5 residues [7]. The secondary structure content of proteins ranges from purely helical proteins, such as myoglobin, containing six α-helices (see Figure 1.2) over mixed α/β proteins to so-called β-barrels, such as green fluorescent protein (GFP), see Figure 1.3, or Omp membrane pores in the outer membranes of gram-negative bacteria. Secondary structure elements provide stability to the protein structure and serve, e.g to anchor the catalytic residues of the active site at precise positions from each other (see below). α-Helices are also the structural basis of coiled coils, see Figure 1.4, because the helices can nicely pack against each other. α-Helices are frequently used by transcription factors, such as GCN4, at the DNA-binding interface, where the α-helices can intercalate in the major or minor grooves of the DNA double helix.

1.1.5 Active Sites

Active sites of enzymes are locations where bound substrate molecules undergo chemical modifications while being bound to the enzyme. Figure 1.5 shows the active site of the serine protease chymotrypsinogen A with the characteristic catalytic residues serine, histidine, and aspartic acid. In principle, discussing enzymatic mechanisms is out of scope for this book, which mostly deals with interactions that proteins engage in. Some multienzyme complexes having multiple active sites assemble to enable the product of one reaction to be passed from

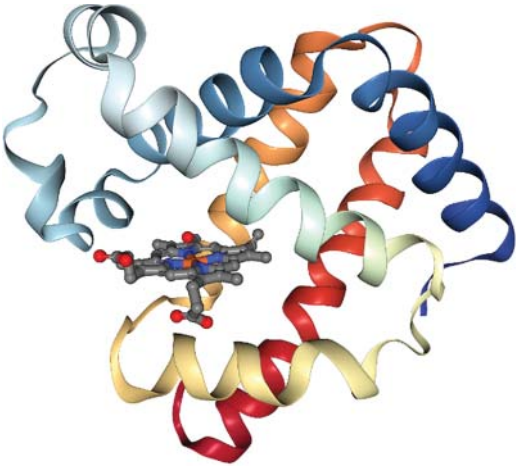


Figure 1.2 X-ray structure (PDB code 1MBN) of myoglobin from *Physeter catodon*. The porphyrin cofactor is anchored between six alpha helices. Source: Figure generated with NGL viewer.

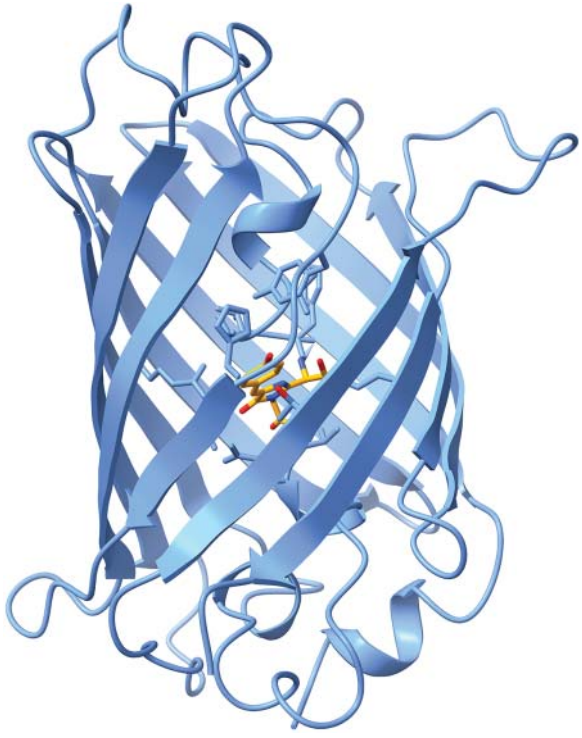


Figure 1.3 X-ray structure of the green fluorescent protein from *Aequorea victoria* (PDB code 1EMA). The barrel-shaped structure is formed by 11 beta-strands surrounding a central alpha-helix holding the chromophore. Source: Figure generated with UCSF Chimera.

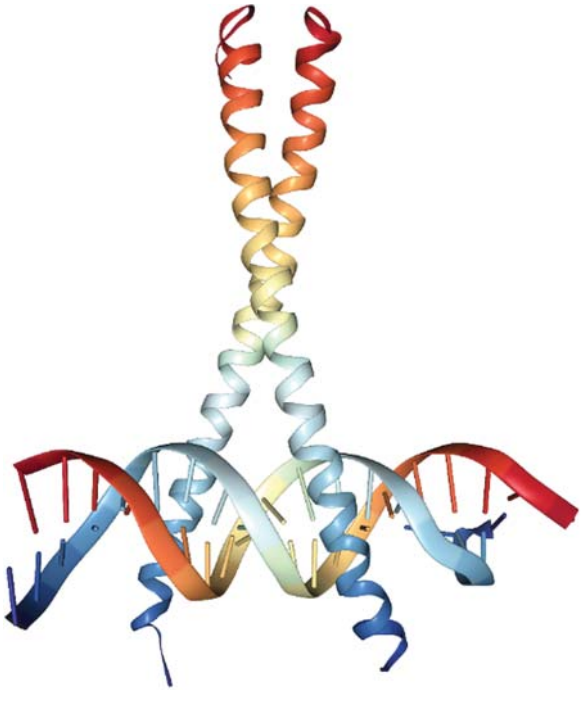


Figure 1.4 X-ray structure of GCN4 dimer from *S. cerevisiae* forming a so-called coiled coil and bound here to DNA (PDB code 1YSA). Source: Figure generated with NGL viewer.

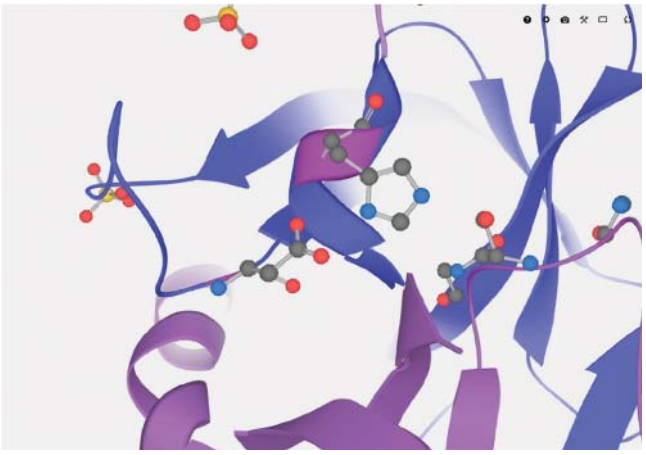


Figure 1.5 Catalytic triad – aspartic acid, histidine, serine – in the active site of a serine protease. Source: European Molecular Biology Laboratory (EMBL).

one active site to other, where it becomes the substrate of a follow-up chemical reaction. Generally, access to active sites should not be precluded by binding to other interaction partners, although, in some cases, binding patches need to be close to the active site, e.g. when a kinase binds its substrate on a patch on the surface of the large lobe so that a phosphate group can be transferred from bound ATP to a serine residue of the bound substrate as mentioned before.

Often, the active sites of enzymes are located on the protein surface, so that substrates can easily bind while remaining partially solvent exposed. A frequent structural motif is a flexible protein loop that reaches over the bound substrate, e.g. in HIV protease, see Figure 1.6. In other cases, the active site is located inside the protein, such as for cytochrome P450 enzymes or acetylcholine esterase. There, substrates need to pass into the protein structure through a channel that may be up to several nanometers long, see Figure 1.7. The main purpose of such an arrangement is to place the substrate in a low-dielectric cavity that enables complicated chemical reactions to take place. Note that the strength of electrostatic interactions is inversely

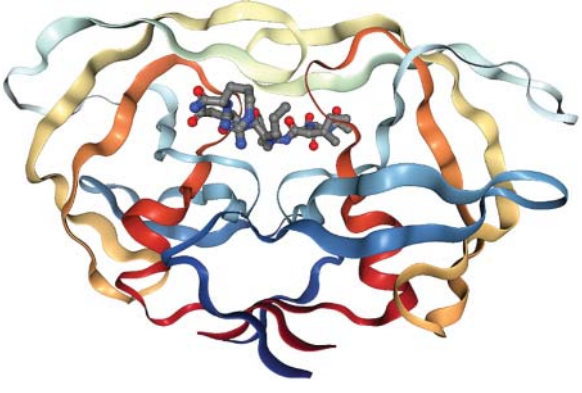


Figure 1.6 X-ray structure of an HIV protease dimer (PDB code 4HVP). A substrate peptide is bound in the active site. Access to the active site is controlled by opening/closing transitions of two flexible loops above the peptide (flaps). Figure generated with NGL viewer.

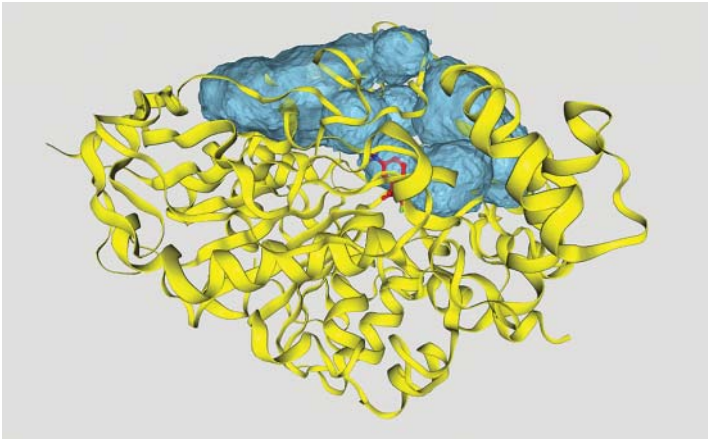


Figure 1.7 Trimethyl ammonio trifluoroacetophenone ligand bound in the active site of acetylcholinesterase from *tetronarce californica* (PDB code 1AMN). The surface contours illustrate several pores and cavities that make up tunnels leading to the internal active site. Source: The figure was generated with the ProPores2 web server (https://service.bioinformatik.uni-saarland.de/propores) [8].

proportional to the dielectric constant of the environment. In a low dielectric environment, charged protein residues can exert stronger electron-pulling or pushing effects on the substrate. Enzyme active sites, ligand binding sites, or translocation pores of ion channels can either reside in individual protein units or in between the interfaces of multimers.

1.1.6 Membrane Proteins

Integral transmembrane proteins are integrated into cellular membranes whereby their amino acid chain crosses the hydrophobic bilayer once or multiple times. While their soluble domains have the same composition as water-soluble proteins, the membrane-spanning parts have a so-called "inside-out" composition. These membrane regions are very hydrophobic on the outside that is in contact with the aliphatic lipid chains of the phospholipid bilayer and have a partially polar interior that often contains a water-filled translocation channel for substrate molecules. When the peptide chain crosses the bilayer, no hydrogen bonding is possible with the aliphatic lipid chains that are in strong contrast to the situation in the water phase. To satisfy the hydrogen bonding capacity of its backbone atoms, the chain thus adopts either an α-helical conformation or a β-sheet conformation in the membrane. Beta barrels consist of 8–22 β-sheets [9] but are only found in the outer membranes of gram-negative bacteria, mitochondria, and chloroplasts. Helical transmembrane proteins possess between 1 and around 20 alpha helices [10] that are between 10 and 30 residues long. The majority of helical membrane proteins possess only 1 transmembrane domain (TMD), followed by those having 2 TMDs and smaller fractions with 3, 4, 7, and 12 TMDs [10]. Oligomerization is frequently

found among helical transmembrane proteins, whereby their binding interfaces consist of roughly perpendicular α-helices. Many receptors on cell surfaces form functional dimers. Ion channels form tetra- and hexamers, with the ion-conducting pore between the monomers. Interactions between proteins and membranes are further discussed in Chapter 13.

1.1.7 Folding of Proteins

Predicting the folded structure of a protein from its sequence has long been a holy grail. In the meantime, scientists have been able to put many pieces of this puzzle together. Important contributions to this were, e.g. the phi-value analysis experiments by Fersht and coworkers that quantify the degree of native folded structure around mutated residues in the folding transition state [11] and the theoretical work by Wolynes, Onuchic, and others, who drew an analogy between the folding of biopolymers and relaxation processes in spin glasses [12]. According to this "new view" of protein folding, a polypeptide chain folds on a rugged funnel-shaped energy landscape where the entropy is plotted on the x-axis and the enthalpy on the y-axis. A protein reaches the lowest free energy point, its folded state, by trading entropy for enthalpy. In this model, protein chains are not able to fold properly either above the folding temperature (where adopting a compact folded structure is entropically unfavorable) or below the glass-transition temperature (where the protein dynamics essentially freeze before reaching the folded state). The David Baker group has been leading the protein structure prediction field for many years using their Rosetta simulation method that extensively samples the combinatorial structural manifold made up of small structural fragments [13]. A further important advance was the brute-force molecular dynamics simulations by the D.E. Shaw group, who were able to simulate the repeated folding and unfolding of small globular proteins at the folding temperature [14]. Recently, the company DeepMind successfully applied deep-learning methods to tackle the problem of protein structure prediction [15, 16]. They trained a neural network to make accurate predictions of the distances between pairs of residues. In the latest Critical Assessment of protein Structure Prediction (CASP), their method termed AlphaFold2 created highly accurate structure predictions with a median backbone accuracy of 0.96 Å root mean square deviation (RMSD) and all-atom accuracy of 1.5 Å RMSD.

Proteins are synthesized by ribosomes either in the cytosol, close to the membrane of the endoplasmic reticulum, or close to the bacterial plasma membrane [17]. It is becoming more and more clear that portions of the nascent peptide chains may already start adopting alpha-helical conformations while passing through the ribosomal exit tunnel. All proteins of the secretory pathway and all membrane proteins are passed from the ribosome to the Sec translocon, an integral membrane channel in the endoplasmic reticulum (ER) membrane. The peptide sequences of membrane proteins are able to exit the Sec complex sideways into the membrane via a so-called lateral gate. Proteins targeted for the secretory pathway need to translocate into the ER, and often get glycosylated by a nearby oligosaccharyltransferase enzyme.

1.2 Conformational Dynamics

Thermal motion of atoms implies that proteins are not rigid objects. Yet, they can still be fairly stiff and have a pure scaffolding function. Examples of this are the proteins of virus capsids or the cytoskeleton. Most proteins, however, undergo some type of conformational transition either during their catalytic cycle, when they bind and unbind ligands, or if they are part of a signaling cascade.

1.2.1 Large-Scale Domain Motions

Proteins consisting of multiple domains or lobes (such as kinases) can undergo large-scale conformational transitions by characteristic domain movements. Prototypes for this are kinases and lysozyme. The first normal mode typically describes a Pacman-type opening–closing transition of the two domains relative to each other, see Figure 1.8. The second normal mode would then be a scissor-like motion perpendicular to the first mode. Often, these movements are connected to biological functions and facilitate either ligand binding and unbinding or help in catalyzing the enzymatic reaction. Membrane transporters, such as the leucine transporter LeuT, undergo a conformational transition between an inward-facing conformation and an outward-facing conformation, see Figure 1.9.

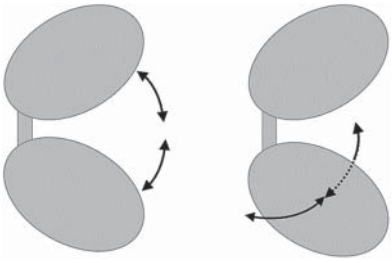


Figure 1.8 Schematic illustration of the first (lowest energy) normal mode of a two-domain protein, such as protein kinases (left), and the second normal mode (right).

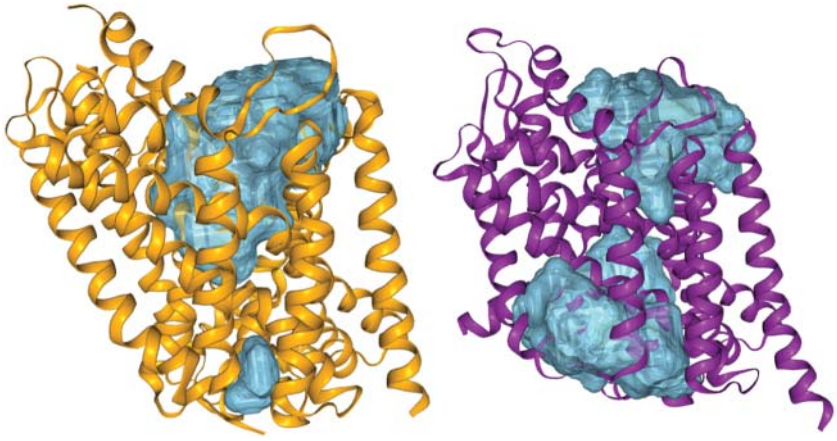


Figure 1.9 X-ray structures of the bacterial leucine transporter LeuT in the outward-facing conformation (left, PDB code 3TT1) and in the inward-facing conformation (right, PDB code 3TT3). The figures were again generated with ProPores2 (cf. Figure 1.7).

Besides such large-scale dynamics, the rest of the protein structure is of course not rigid but undergoes constant thermal motion as well. Since the 1970s, time-resolved IR spectroscopy was used to characterize the dynamics of laser-induced CO dissociation from the internal porphyrin ring of myoglobin [18]. The observed multi-exponential kinetics of the time needed for CO to rebind to the porphyrin was interpreted to reflect the intrinsic dynamics of the myoglobin matrix. Subsequently, Halle and coworkers showed, by NMR, that water molecules buried in the protein bovine pancreatic trypsin inhibitor (BPTI) exchanged with bulk solvent on time scales of milliseconds [19]. This proved that even compact globular protein structures undergo continuous conformational breathing transitions that are large enough to allow the passage of water molecules in and out of a folded protein.

1.2.2 Dynamics of N-Terminal and C-Terminal Tails

N-terminus and C-terminus of a protein chain are typically located on its protein surface, where they often stretch out into solution and have substantial conformational flexibility. Probably, the functionally most important N-terminal tails are those of histone proteins. They undergo posttranslational modifications in many ways, and this strongly affects their interaction with double-stranded DNA that winds around histone proteins. The C-terminal tails of proteins can function, e.g. as recognition sites for PDZ adaptor domains.

1.2.3 Surface Dynamics

Amino acid side chains on the surface of proteins often also show considerable conformational dynamics [20]. Frequently, transient pockets open and close on protein surfaces on a timescale of tens of picoseconds. Thus, the protein surface rather resembles the surface of a sponge. Another type of functionally relevant conformational motions are loop movements on the protein surface, e.g. lipases possess a loop termed “lid” that controls access to the active site beneath. The same is the case for HIV protease as mentioned before. Interestingly, it has been argued that disease-associated mutations in proteins often result in flexibility changes even at positions distal from mutational sites, particularly in the modulation of active-site dynamics [21].

1.2.4 Disordered Proteins

X-ray crystallography and Cryo-EM are perfect structural techniques to resolve precise conformational details of well-ordered portions of proteins. Obviously, N-terminus, C-terminus, and surface loops extend into the solvent, and their conformational dynamics may sometimes not yield precise electron density that can be detected against the background. Furthermore, it came as a surprise when NMR experiments showed in the mid 1990s that there exist numerous “disordered” proteins that do not adopt a well-folded conformation at all. Sometimes, they may refold when they bind to other proteins, or when they undergo a phenotypic order-to-disorder transition, such as the prion protein that is more folded in the non-disease state and is thought to be the origin of mad cow disease. All of us

contain prion proteins and we are usually just fine. According to the "protein-only" hypothesis, the key event in the prion disease pathogenesis occurs when the cellular prion protein (PrPC) undergoes a conformational transition from a mainly α-helix-rich folded structure into an infectious and pathogenic β-sheet-rich conformer (PrPSc). PrPSc possesses abnormal physiological properties, such as resistance to proteolytic degradation, relative insolubility, and the propensity to polymerize into scrapie agents [22].

Monzon et al. distinguished short and disordered regions (between 5 and 30 residues long) that are usually associated with flexible linkers or loops in folded proteins and so-called long disorder regions (LDRs) that have at least 30 consecutive disordered residues. These LDRs were found to be enriched in charged and hydrophilic amino acids and depleted in hydrophobic ones [23], such as the linker segments discussed before in the context of protein domains. Disordered regions may also have important roles in mediating protein interactions. For example, so-called eukaryotic linear motifs (ELMs) are located in disordered regions of proteins and mediate interactions between proteins [24].

1.3 From Structure to Function

1.3.1 Evolutionary Conservation

One important principle of evolutionary biology is that functionally important protein regions tend to be conserved between related organisms whereas unimportant regions are subject to considerable variation. Functionally important regions include, of course, active site residues. Mutations of catalytic residues may render enzymes nonfunctional and are, therefore, rarely tolerated. Furthermore, conservation also extends to structural elements, such as disulfide bridges and residues in short turns.

In general, structure is better conserved than sequence. Therefore, functionally related pairs of proteins may sometimes show very low sequence similarity, but fairly high structural similarity. Assuming that both proteins were derived from a distant common ancestor protein, it came about that their structures were conserved during evolution, but their sequences were not, except for a few crucial positions.

1.3.2 Binding Interfaces

Many proteins carry out their function by binding to other proteins, small molecules, membranes, or nucleic acids. This is actually what all of this book is about. Usually, this involves one or more binding patches on the surface of the proteins. Binding interfaces of two proteins have sizes ranging from 500 to 3000 $Å^2$ [25]. Small interfaces are preferred for transient contacts of small hydrophilic proteins, e.g. those of redox proteins such as the electron carrier cytochrome *c*. In contrast, antibodies bind to their antigens with rather large and hydrophobic interfaces that support permanent or at least long-lasting contacts. Also, permanent dimers tend to have rather

hydrophobic interfaces. How much of the protein surface is part of an interface depends on the total size of the complex. An internal protein, e.g. in the ribosome may even be fully shielded from solvent and all of its surfaces are in contact with other biomolecules. Protein–protein interactions and large protein complexes are discussed in Chapters 2–7.

DNA and RNA are strongly negatively charged due to their phosphate backbones. Hence, proteins need to possess complementary, positively charged surface patches, to be able to bind to DNA or RNA. Such patches are typically not suitable for binding to other proteins. However, there are certain proteins that are able to mimic nucleotide polymers. One example is the intracellular inhibitor protein barstar that binds to the RNAse barnase and prevents it from chewing up all mRNA and other RNA molecules inside the cell. Thus, barnase only acts extracellularly. Barstar has a strongly negative binding patch to mimic the natural substrate RNA. Chapters 10–12 give a deeper insight into protein interactions with nucleic acids.

The topology and composition of binding interfaces will be discussed in detail in Chapter 2.

1.3.3 Surface Loops

Surface loops are used, for example by antibodies, to bind to their antigens via complementarity-determining regions (CDRs). As mentioned, surface loops can also regulate the access to the active site of proteins, and they may contain cleavage sites for restriction enzymes. Note that cleavage is almost as frequently observed in α-helices as in regions without secondary structure, such as loops, but less in β-strands [26].

1.3.4 Posttranslational Modifications

Often, the activity of proteins is determined by the proper placement of posttranslational modifications to surface residues. For example, about 75% of all human proteins get phosphorylated, often at multiple positions [27]. Other modifications are glycosylation, farnesylation (e.g. of the Ras protein), etc. Ubiquitination often ends the life of proteins because this modification targets them for transport to the proteasome that shreds peptide sequences into small components. The modification sites are usually located on the protein surface and the modifications are placed by other enzymes, again involving protein interactions. Posttranslational modifications are important markers for binding partners and may also affect protein conformation (see Chapter 17 for further discussion).

1.4 Summary

The characterization of protein structure has become fairly routine these days. For about 70% of all human proteins, there exist structural models either from experimental determination or from homology modeling [28]. In fact, DeepMind, in cooperation with European Bioinformatics Institute (EBI), recently published structural

models produced with AlphaFold for all human proteins and proteins of several other model organisms [29]. Some believe that even the protein folding problem has been, at least partially, solved. Despite all the accumulated knowledge, we still do not know the function of a considerable fraction of the human proteins, and it is very hard to rationalize the functional effects of posttranslational modifications or to even predict them. We have a limited understanding of what determines protein interactions, and we are rarely able to correctly predict the structures of protein assemblies from scratch, without additional experimental evidence.

References

1 Tiessen, A., Pérez-Rodríguez, P., and Delaye-Arredondo, L.J. (2012). Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res. Notes* 5: 85. https://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-5-85.

2 Wheelan, S.J. et al. (2000). Domain size distributions can predict domain boundaries. *Bioinformatics* 16: 613–618.

3 Yu, L., Tanwar, D.K., Penha, E.D.S. et al. (ed.) (2019). Grammar of protein domain architectures. *Proc. Natl. Acad. Sci.* 116: 3636–3645. https://www.pnas.org/content/116/9/3636.

4 Hormoz, S. (2013). Amino acid composition of proteins reduces deleterious impact of mutations. *Sci. Rep.* 3: 2919.

5 Brüne, D., Andrade-Navarro, M.A., and Mier, P. (2018). Proteome-wide comparison between the amino acid composition of domains and linkers. *BMC Res. Notes* 11: 117.

6 Kumar, S. and Bansal, M. (1998). Geometrical and sequence characteristics of α-helices in globular proteins. *Biophys. J.* 75: 1935–1944.

7 Penel, S. et al. (2003). Length preferences and periodicity in β-strands. Antiparallel edge β-sheets are more likely to finish in non-hydrogen bonded rings. *Protein Eng. Des. Sel.* 16: 957–961.

8 Hollander, M., Rasp, D., Aziz, M., and Helms, V. (2021). ProPores2: web service and stand-alone tool for identifying, manipulating and visualizing pores in protein structures. *J. Chem. Inf. Model.* 61: 1555–1559.

9 Tian, W., Lin, M., Tang, K. et al. (2018). High-resolution structure prediction of β-barrel membrane proteins. *Proc. Natl. Acad. Sci.* 115: 1511–1516.

10 Reeb, J., Kloppmann, E., Bernhofer, M., and Rost, B. (2015). Evaluation of transmembrane helix predictions in 2014. *Proteins* 83 (3): 473–484.

11 Matouschek, A., Kellis, J.T. Jr., Serrano, L., and Fersht, A.R. (1989). Mapping the transition state and pathway of protein folding by protein engineering. *Nature* 340: 122–126.

12 Onuchic, J.N. and Wolynes, P.G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.* 14: 70–75.

13 Yang, J., Anishchenko, I., Park, H. et al. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* 117: 1496–1503.

14 Robustelli, P., Piana, S., and Shaw, D.E. (2018). Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* 115: E4758–E4766.

15 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589.

16 Senior, A.W., Evans, R., Jumper, J. et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577: 706–710.

17 Bornemann, T., Jöckel, J., Rodnina, M.V., and Wintermeyer, W. (2008). Signal sequence–independent membrane targeting of ribosomes containing short nascent peptides within the exit tunnel. *Nat. Struct. Mol. Biol.* 15: 494–499.

18 Austin, R.H., Beeson, K.W., Eisenstein, L. et al. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry* 14: 5355–5373.

19 Denisov, V.P., Peters, J., Hörlein, H.D., and Halle, B. (1996). Using buried water molecules to explore the energy landscape of proteins. *Nat. Struct. Biol.* 3: 505–509.

20 Helms, V. (2007). Protein dynamics tightly connected to the dynamics of surrounding and internal water molecules. *ChemPhysChem* 8: 23–33.

21 Campitelli, P., Modi, T., Kumar, S., and Ozkan, S.B. (2020). The role of conformational dynamics and allostery in modulating protein evolution. *Annu. Rev. Biophys.* 49: 267–288.

22 Baral, P.K., Yin, J., Aguzzi, A., and James, M.N.G. (2019). Transition of the prion protein from a structured cellular form (PrPC) to the infectious scrapie agent (PrPSc). *Protein Sci.* 28: 2055–2063.

23 Monzon, A.M., Necci, M., Quaglia, F. et al. (2020). Experimentally determined long intrinsically disordered protein regions are now abundant in the protein data bank. *Int. J. Mol. Sci.* 21: 4496.

24 Tompa, P., Davey, N.E., Gibson, T.J., and Babu, M.M. (2014). A million peptide motifs for the molecular biologist. *Mol. Cell* 55: 161–169.

25 Janin, J., Bahadur, R.P., and Chakrabarti, P. (2008). Protein–protein interaction and quaternary structure. *Q. Rev. Biophys.* 41: 133–180.

26 Timmer, J.C., Zhu, W., Pop, C. et al. (2009). Structural and kinetic determinants of protease substrates. *Nat. Struct. Mol. Biol.* 16: 1101–1108.

27 Sharma, K., D'Souza, R.C.J., Tyanova, S. et al. (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* 8: 1583–1594.

28 Somody, J.C., MacKinnon, S.S., and Windemuth, A. (2017). Structural coverage of the proteome for pharmaceutical applications. *Drug Discovery Today* 22: 1792–1799.

29 Varadi, M., Anyango, S., Deshpande, M. et al. (2022). AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50: D439–D444.

2

Protein–Protein-Binding Interfaces

Zeynep Abali[1], Damla Ovek[2], Simge Senyuz[1], Ozlem Keskin[3], and Attila Gursoy[2]

[1] *Koc University, Computational Science and Engineering Program, Istanbul, 34450, Turkey*
[2] *Koc University, Computer Engineering, Istanbul, 34450, Turkey*
[3] *Koc University, Chemical and Biological Engineering, Istanbul, 34450, Turkey*

2.1 Definition and Properties of Protein–Protein Interfaces

The surface regions where proteins interact with other molecules are called **protein-binding sites**. If the interaction occurs between two proteins, then interacting binding sites form a **protein–protein interface**. Interfaces involve amino acids from each side forming mainly non-covalent bonds. Interfaces might also contain covalent bonds, such as disulfide bridges, but with lower frequency.

The physical proximity of residues from two protein chains determines the interface residues in each protein. Interfaces can be described using a variety of computational methods [1]. These methods use structures of protein–protein complexes and various metrics, such as distance between the atoms belonging to different subunits (protein chain), or accessible surface area (ASA). Interface residues do not need to be continuous in sequence but should be close to each other in 3D space. Here, we present some of the commonly used methods. A distance-based approach is one of them. Residues of an interface can be defined by the distance between their atoms. A threshold distance is defined, usually ranging between 4 and 6 Å. If two residues of opposing chains have heavy atoms (non-hydrogen) within the defined threshold distance, then these residues are categorized as **interface residues** [2, 3]. Some other studies consider only the distances between Cα atoms to identify interface residues. When Cα atoms are used, the threshold distance is usually greater than the ones used with heavy-atom approaches, ranging from 8 to 12 Å [4–6]. Another distance-based method defines the distance between two atoms using the van der Waals (VDW) radii of the individual atoms. Two residues are defined as interface residues if they have atoms within a distance that is smaller than the sum of their VDW radii plus a threshold distance (usually 0.5 Å) [7, 8].

Distance-based methods are not the only ones for identifying interface regions in protein complexes. Alternatively, ASA or relative accessible surface area (rASA)

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

of individual residues can be used to find interface residues. ASA is the area of a molecule that is accessible to a solvent. In ASA calculations, usually, a sphere with the radius of a water molecule (1.4 Å) is rolled around the protein to probe its surface. There are several available tools for calculating the ASA of residues in a protein, such as NACCESS ([9]), POPScomp [10], or FreeSASA [11]. rASA is calculated by taking the ratio of two states of a residue: (i) when it is in the most solvent-exposed state (in Ala-X-Ala or Gly-X-Gly tripeptide where X is the residue of interest) and (ii) when it is in the folded conformation of the protein.

Interface regions on complexes can be identified by considering the change in ASA (ΔASA). The residue ASAs are calculated when the protein is in its monomeric form and in complex form. If the difference between monomeric ASA and the complex ASA is larger than a threshold, then the residue is identified as an interface residue. A threshold value of 1 $Å^2$ is generally used [12]. SPPIDER [13] is one of the available tools that uses rASA values for identification of interface residues. It uses a 4% threshold of rASA change between the monomer and the complex and $\Delta ASA > 5$ $Å^2$. Another study uses a threshold of 25% for rASA and $\Delta ASA > 0$ $Å^2$ to define interface residues [14].

There are other methods to define interfaces that are not as common as the mentioned ones. For example, Voronoi diagrams are used as a geometric approach for identifying interfaces and specifying the boundaries of a given interface [15]. There are also some studies that embrace graph-based approaches to define interface regions [16].

Methods for defining protein–protein interfaces can be used on their own as a single method, or as a combination of multiple methods. For example, Hadarovich et al. defined interface residues by a 12 Å atom–atom distance cutoff between the interacting monomers and then eliminated small interfaces that have buried surface area <200 $Å^2$ per chain [4]. Since distance-based calculations are compute-intensive, Cukuroglu et al. defined interface regions first using $\Delta ASA > 1$ $Å^2$ and then by distance criteria. They defined interface residues as **contacting** (Figure 2.1a) if the distance between any two atoms of the two residues from different chains is less than the sum of their corresponding VDW radii plus a threshold of 0.5 Å [17].

A more continuous interface structure is usually preferable. Besides the interface residues that are in contact, the nearby (neighbor) residues can also be included in the interface regions to make it more continuous and to preserve the secondary structures [7, 17, 18]. After identifying contacting residues, nearby residues are defined based on the contacting residues. If a residue has a Cα atom at most 6 Å away from the Cα of a contacting residue, then it is defined as a nearby residue (Figure 2.1b). Nearby residues provide a supporting scaffold for contacting residues in interface regions [7].

Interface regions can be divided into **core** and **rim areas** similar to regions in protein globular structures. Interface cores are similar to protein cores, and interface rims are similar to protein surfaces. Core residues contribute more to the binding affinity and specificity [14, 19–21]. Core and rim regions are defined by the change of ASA of residues upon complex formation. If a surface residue becomes solvent inaccessible after complex formation, it is part of the interface core; on the other

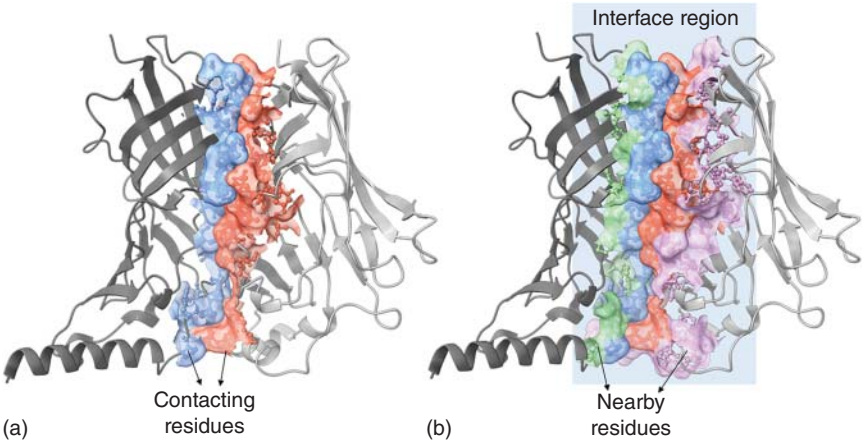


Figure 2.1 This figure shows the structure of A and B chains of the protein with PDB ID 2WNJ. (a) Contacting residues are highlighted with blue and red on the left-hand side, and (b) nearby residues are added with green and pink on the right-hand side of the figure. It can be seen that the addition of nearby residues provides a more complete representation of the interface region between these two chains.

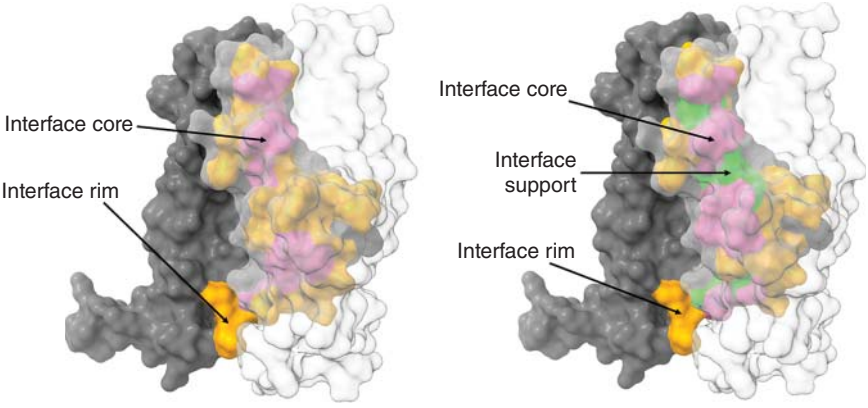


Figure 2.2 (a) Shows the partition of an interface into two regions, core and rim. (b) shows the partition of an interface into three regions, core, rim, and support. Regions shown in pink are **core**, orange regions are **rim**, and green regions are **support**.

hand, if a surface residue remains partially solvent accessible, then it is a part of the interface rim. Figure 2.2a shows a model interface with core and rim regions. The comparison of interface core and rim residues shows that core residues are more likely to be conserved, and their side chains are less flexible [20, 22]. In addition to these regions, Levy also defined a support region [14]. Figure 2.2b shows the three regions on the same interface.

An overview of how to partition the residues in an interface according to both methods is given in the table below (Table 2.1).

Table 2.1 Definition of core and rim regions in interfaces.

	Two-region (core-rim)	Three-region (core-rim-support)
Core	$\Delta ASA > 0$ & $ASA_{Complex} = 0$	$\Delta rASA > 0$ & $rASA_{Monomer} > 25\%$ & $rASA_{Complex} < 25\%$
Rim	$\Delta ASA > 0$ & $ASA_{Complex} > 0$	$\Delta rASA > 0$ & $rASA_{Complex} > 25\%$
Support	–	$\Delta rASA > 0$ & $rASA_{Monomer} < 25\%$

Source: Adapted from Levy [14].

2.2 Growing Number of Known Protein–Protein Interface Structures

More than 170 000 structures are deposited to Protein Data Bank (PDB) [23] as of February 2021, ranging from small monomer structures, like ubiquitin, to considerably large complex structures, such as the entire HIV-1 capsid. Especially advanced imaging methods, such as Cryo-EM, enable structural determination of large proteins, and with better resolution, thanks to current experimental improvements [24]. This advancement enables obtaining large and multi-protein complex structures. The average number of chains per deposited structure in PDB increased from 2 to 7 in the last ten years; and currently, the largest available protein complex has 1356 chains (PDB ID: 3J3Q). The number of structurally available interfaces is related to the chain count in a given structure. The increase in the number of available structures, as well as the growth in size of the structures deposited, in turn, helps in determining more interface structures each year.

Figure 2.3 shows the increase in the number of interfaces identified in the PDB throughout the years. In a previous study by Cukuroglu et al., 130 209 protein–protein interfaces were identified in the PDB in 2014 [17]. A more recent analysis of the PDB for interfaces revealed that this number is 449 169 in 2020 [25]. Our previous interface definition is used in Figure 2.3 [7, 17, 26]. There is

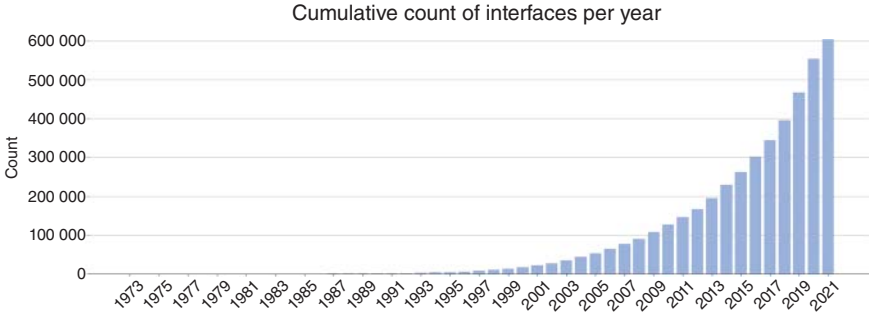


Figure 2.3 Cumulative number of interfaces in PDB through years from 1971 to 2021.

Table 2.2 Available interface related datasets.

Tool/data set	Web server	Interface identification method	Input	Description
COCOMAPS Vangone et al. [28]	https://www.molnac.unisa.it/BioTools/cocomaps/	Atomistic Distance Threshold	PDB ID or PDB File	COCOMAPS is a web application to analyze and visualize the interface of biological complexes including protein–protein, protein–DNA, and protein–RNA complexes. The output of a query includes contact maps of the interface, a table about the interacting residues, and a 3D visualization of the complex
PDBParam Nagarajan et al. [29]	https://www.iitm.ac.in/bioinfo/pdbparam/compute-new.html	CA or CB Distance Threshold	PDB ID or PDB File	PDBparam is an online tool for identifying binding sites, inter-residue interactions between chains, secondary structure propensities of the complex, and various physicochemical properties, such as ASA, surface hydrophobicity, and normalized flexibility parameters
PDBePISA Krissinel et al. [30]	https://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver	RSA Change	PDB ID	PDBePISA is an online tool for calculating structural and chemical properties of macromolecular interfaces. It provides information on ASA
PDBSum Laskowski et al. [31]	https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html	RSA Change	PDB ID	PDBsum provides image-based structural information on the structures available in PDB. Prot-prot tab available on the results page of a search shows the schematic diagrams of all protein–protein interfaces that the structure has. It can visualize the 3D structure of the complex, as well as it shows the residue–residue interactions between the chains
Piface Cukuroglu et al. [17]	https://interactome.ku.edu.tr/piface/	VDW Distance Threshold and CA Distance Threshold	PDB ID, PDB File, Pfam Domain	PIFACE is a data set of nonredundant unique protein–protein interfaces from PDB that are clustered according to their structural similarity. Users can find information about size and ASA of interfaces and chains, taxonomy information, experimental method, and resolution. Search by Pfam domains is also available. Full cluster information is also available for download

(*continued*)

Table 2.2 (Continued)

Tool/data set	Web server	Interface identification method	Input	Description
PIMA Kaleeckal Mathew and Sowdhamini [32]	http://caps.ncbs.res.in/pima/	CA–CA distance threshold	PDB File or URL to PDB ID	PIMA is an online tool for analyzing interactions between chains in a protein complex. It identifies the interactions based on features, such as total stabilizing energy, hydrogen bonds, salt bridges, and interface contacts, and provides a graphical representation of the interactions, along with the energy values and queried interface residues
ProtinDB Jordan et al. [33]	http://ailab-projects2.ist.psu.edu/protInDb/DataSetsNew.py	User Selected	PDB ID	ProtinDB is a database of protein–protein interfaces extracted from protein complexes available in PDB. The web server provides a visualization interface for residues that are on the interface. It is possible to construct a data set of protein–protein interfaces using a customized list of PDB IDs
ProtCID Xu and Dunbrack et al. [34]	http://dunbrack2.fccc.edu/ProtCiD/Default.aspx	CB Distance Threshold	PDB code, Pfam ID sequence, or UniProt IDs	ProtCID provides structural information about the interaction of proteins and individual protein domains with other molecules. It aims to identify and cluster homodimeric and heterodimeric interfaces seen in multiple crystal forms of homologous proteins and their interactions with peptides and ligands. The results of a search query include the number of crystal forms that contain a common interface, the number of PDB and PISA biological assembly annotations that have the same interface, the average surface area, and the minimum sequence identity of proteins that have the interface
3did Mosca et al. [35]	https://3did.irbbarcelona.org/		Domain name, Pfam access code, PDB ID, motif name, or GO term	3did provides structural templates for domain–domain interactions of high-resolution structures in PDB. It includes template information between globular domains and also domain–peptide interactions. Results provided for a search term include a graph that shows the chains of the structure with Pfam domains, and interactions of domains between chains, visualization of protein structure with Jmol, and a list of the domain architectures of each chain, which also gives detailed information about the location of the domain on the chain, and a list of interactions that involve the given chain

almost a four-fold increase in the number of structurally available interfaces in the PDB in six years, and this is still only a fraction of all protein–protein interactions *in vivo*, since not every complex can be identified structurally with our current experimental capabilities.

This large, and fast-growing, set of 3D structures of protein–protein interfaces is an invaluable resource to better analyze the properties of these interfaces, such as geometrical properties of size and shape, structural as well as sequence conservation, residue propensities, or complementarity of the interfaces. In addition to providing a better analysis set, a larger number of interface structures can provide better templates for structural prediction of interfaces without known structure [27]. To take advantage of this growing resource, several datasets are created. In Table 2.2, we present some of the currently available interface-related datasets that can be used for analysis at large and small scales.

2.3 Surface Areas of Protein–Protein Interfaces

The surface of a protein–protein interface is the buried area upon complex formation. The buried surface areas range from 300 to 6000 Å^2 [36]. The size of one side of an interface generally ranges between 200 and 2800 Å^2 [37]. The majority of the interfaces are within the 600–1200 Å^2 range. The average size of an interface is found as 1227 Å^2. Figure 2.4 shows the distribution of interface sizes (one side) extracted from the PDB. The number of interface residues is correlated with the buried surface areas. There are on average 56.9 contacting residues in an interface, while the largest interface has 803 residues on one side [25].

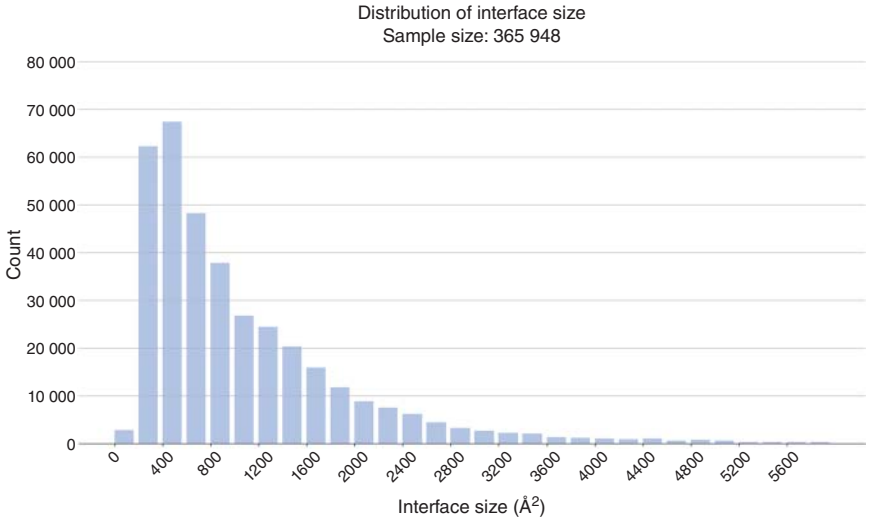


Figure 2.4 Distribution of interface sizes in structures from PDB as of January 2021.

2.4 Gap Volume of Protein–Protein Interfaces

Protein surfaces are not flat. This also applies to protein–protein interfaces, although to a lesser extent. There are cavities and protruding regions that affect protein binding. The size and shape of these cavities, for example, are important for ligand binding; therefore, they have a key role in modulating protein-protein interfaces. Cavities are one of the reasons interacting surfaces on monomers cannot fit perfectly, but they leave gaps in some regions. The total volume of these unfilled spaces between two interacting surfaces is termed the gap volume. Gap volume can be used as a measure of the complementarity and the closeness of packing between the two binding sites of an interface. The SURFNET [38] package is one of the tools that enable the investigation of cavities on proteins and gap volume of interfaces. Gap volume index is usually used to classify tightly and loosely packed interfaces. It is the ratio of the gap volume to the interface buried surface area. The mean gap volume index is shown to be higher in heterodimers than in homodimers [39].

2.5 Amino Acid Composition of Interfaces

The frequencies of different residues in the interfaces may provide information about the hydrophobic/hydrophilic character of the interface [40]. When interface residues are compared with core residues of proteins, interfaces have more polar and charged residues [41]. Previous studies imply that the properties of residues, such as hydrophobicity, polarity, and electrostatics, can be used to identify interface regions on protein surfaces [42]. Figure 2.5 shows the amino acid frequencies evaluated on three different regions – protein core, interface, and non-interface surface regions of protein complexes, using 22 604 interfaces from 16 181 nonredundant PDB complex structures. Interface regions are defined as the collection of contacting and nearby

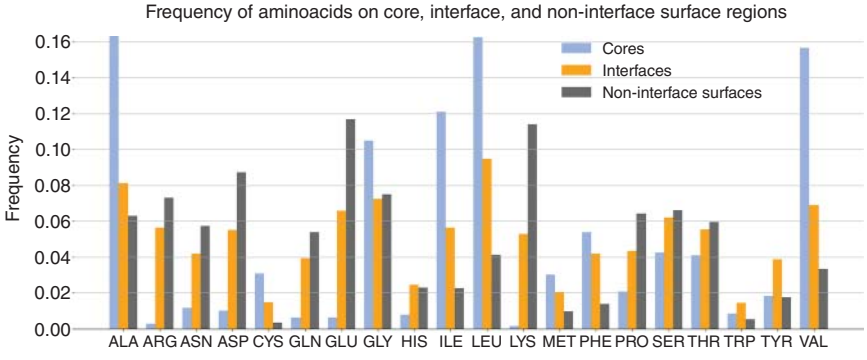


Figure 2.5 Frequencies of amino acids on protein core, interface, and non-interface surface residues of protein complexes.

residues [17]. Core residues are non-interface residues with an ASA value of 0, and non-interface surface residues are the ones with an rASA >25%. Frequencies are calculated as follows:

$$\text{Frequency of an aminoacid} = \frac{\{\text{Count of the aminoacid in a given region}\}}{\{\text{Total number of aminoacids in the given region}\}}$$

Many groups analyzed the frequencies of different amino acids in three regions of protein–protein complexes: core, interface, and non-interface surface [18, 37, 43]. When evaluated under three different regions as core, interface, and non-interface surfaces, hydrophobic residues, Ala, Cys, Ile, Leu, Met, Phe, and Val, are more likely to be found in the core region of a protein, whereas hydrophilic residues, Asn, Asp, Glu, Lys, and Arg, are more likely to be found on non-interface surfaces of a structure. The frequencies of amino acids at interface regions are usually between the frequencies for core and non-interface surface regions [37]. These analyses are limited to a fraction of all possible interactions found in nature, since knowing the three-dimensional structure of a complex, and the interface region is a must for analyzing amino acid propensities. Even though our knowledge of the structure of protein complexes is not complete, these analyses consistently show us that hydrophobic/hydrophilic property of an amino acid significantly affects its frequency in different regions of a protein complex.

2.6 Secondary Structure of Interfaces

The secondary structure of proteins is an essential factor in how chains interact with each other. One of widely used tools to assign secondary structural elements (SSE) to protein structures is DSSP [44, 45]. SSEs on protein–protein interfaces can be grouped into five categories: α–α, β–β, α–αβ, β–αβ, and αβ–αβ. α–α defines interfaces that have only α-helices as SSE in both chains, apart from coils. Likewise, β–β defines interfaces with only β-sheets as SSE in both chains. α–β and β–αβ define interfaces that have only α-helices or β-sheets in one chain and both in the other chain, and lastly, αβ–αβ defines interfaces with both SSEs in both chains. A recent analysis of all interfaces available in PDB revealed that 52.5% of all interfaces are in the α–α category, and around 11% of interfaces are in the β–β category [25]. The rest of the interfaces have at least one αβ chain. α interfaces and αβ interfaces are more frequent in homodimers (34% and 47%, respectively) than they are in heterodimers (22% and 31%, respectively). β interfaces are almost equally common in both homodimers and heterodimers, 11% and 15%, respectively [46].

Previously, we showed that there are recurring architectural motifs in protein–protein interfaces [18, 47]. We observed that although there are many motifs based on SSEs in interfaces, some architectures are more favorable and frequently used, and these are the ones that are also preferred in single-chain protein cores [47].

2.7 Protein–Protein-Binding Energy

Protein–protein-binding free energy (ΔG) can be used to assess the binding affinity of two proteins. Both the enthalpic and entropic terms contribute to ΔG. The formula, $\Delta G = RT\ln(K_d)$, relates the binding free energy to the dissociation constant (K_d) and thus the binding affinity. K_d gives information about protein-binding/unbinding processes at equilibrium. Among these terms, K_d is often reported to describe the stability of a protein–protein complex [48]. Experimentally measured dissociation constants of some biological complexes deposited in the PDB are curated in PDBbind and updated annually [49]. PDBbind aims to show the correlation between the structural and energetic properties of protein–protein, protein–nucleic acid, and protein–ligand complexes. SKEMPI 2.0 is another manually curated database which presents binding affinity and other thermodynamic properties of protein–protein complexes with a focus on the changes in the binding energy upon mutations [50].

The shape complementarity, physicochemical properties, including electrostatic interactions, salt bridges, hydrogen bonds, and Van der Waals interactions, contribute to the binding energy/affinity between proteins [51]. Therefore, the change in the electrostatic environment of the protein–protein-binding interface – resulting from solvation and individual interface residue charges – contributes to the binding free energy [52, 53]. Even though both hydrophobic and electrostatic contributions are important in binding, hydrophobic forces are discussed to be the main driving forces in binding [54]. The binding affinity is related to polarity of the interacting residues and the number of charged residues on the binding interface [55, 56]. High-affinity complexes exhibit more polar–polar and polar–nonpolar interactions [55]. On a structural level, protein–protein-binding affinity is also correlated with buried surface area of an interface [57].

2.8 Interfaces of Homo- and Hetero-Dimeric Complexes

Protein–protein interactions may occur between two or more identical or non-identical monomers. If the monomers in a complex are the same, then it is named as homooligomer. A complex with non-identical subunits is named as heterooligomer. Furthermore, if each subunit in a homooligomer is contacting through the same surface, it is more specifically named as isologous homooligomer. The same naming convention also holds for complexes with two chains: homodimers and heterodimers. Homodimers are the most common protein complexes in nature [58]. As previously discussed, interfaces formed in homodimers or heterodimers show different SSEs [46], complementarity, and amino acid composition. It has been observed that when compared with heterodimers, homodimers have a larger surface area, they have more hydrogen bonds at the interface, and the interface surfaces are more hydrophobic [59, 60]. Homodimer binding sites have an average size of 1311 $Å^2$, while the average size of heterodimers is 1112 $Å^2$ [37].

2.9 Interfaces of Non-obligate and Obligate Complexes

Complex structures can be categorized as non-obligate or obligate complexes. In obligate complexes, the subunits of the complex, the individual proteins that form the complex, are not stable as independent structures *in vivo*. In contrast, non-obligate complexes are formed by subunits that are also stable as separate structures. Obligate structures are usually obligate functionally as well. Whereas complexes that function in receptor–ligand, antibody–antigen, enzyme–inhibitor, or intracellular signaling are usually non-obligate; therefore, they are independently stable as well, since the subunits of such complexes may not be co-localized when they are not interacting [61, 62].

An analysis of obligate and non-obligate interfaces for amino acid propensities shows that amino acids such as Ile, Val, Pro, His, Gln, and Leu have a higher propensity in obligate interfaces compared to non-obligate ones, whereas Cys, Tyr, Asn, Glu, Asp, and Lys have a higher propensity in non-obligate interfaces. The difference in amino acid propensities reveals that non-obligate interfaces are more likely to be polar compared to obligate interfaces. The core and peripheral regions, rim, or support, of obligate and non-obligate interfaces, show different characteristics as well. Compared to the peripheral regions, the core region of non-obligate interfaces has been shown to have a higher frequency of short non-polar residues Ile, Val, Leu, Cys, Ala, Gly, Pro and of aromatic residues, such as Trp and Tyr [63].

Obligatory interfaces are larger than non-obligatory interfaces. The average interface area (one side) for obligatory interfaces is 492.74 Å and for non-obligatory complexes is 279.55 Å [63]. Also, the number of contacts in obligatory interfaces is higher than that of non-obligatory complexes, with 20 and 13 contacts per chain, respectively. Obligatory interfaces are more evolutionarily conserved, they have a higher geometric complementarity, and larger interface-to-surface ratio [64]. There are various tools for identifying interfaces as obligate or non-obligate. DynaFace [65] uses the dynamic motion of the protein complex for discriminating between obligatory and non-obligatory protein–protein interactions, whereas NOXclass [66] uses a support vector machine (SVM) classifier to differentiate between obligate/non-obligate interfaces depending on interface properties, such as interface area, amino acid composition, and residue conservation.

2.10 Interfaces of Transient and Permanent Complexes

Another way to categorize complexes is based on their lifetime. Transient interactions may associate and dissociate *in vivo*, while permanent interactions are generally stable, and subunits involved in permanent interactions usually only exist in the complex. The lifetime of a complex, whether the complex is transient or permanent, depends on interaction strength between the subunits of the complex. The interaction strength is highly affected by hydrophobic interactions, hydrogen bonds, salt bridges, and disulfide bonds that take part in forming the complex

structure. Complexes that are structurally or functionally obligate are generally permanent, while non-obligate interactions can be either transient or permanent.

Protein–protein interactions may not be categorized distinctly into one of these types [61]. Usually, there is a continuum between obligate and non-obligate states of interactions for complex structures, where the stability of the complexes depends on the physiological conditions and the environment of the interaction [67]. An interaction that is mainly transient may become permanent under different environmental conditions. Even though it is not always possible to determine for certain, the location of the subunit of the complex and the function may indicate the interaction type between the subunit. To give an example, complex interactions that take part in intracellular signaling are usually expected to be transient, since they need to associate and dissociate to function [68].

In comparison with transient interfaces, permanent interfaces are more conserved, and they have a higher tendency of having more hydrophobic residues, whereas transient interfaces are shown to have more polar residues [64, 69]. When the size of the interfaces is compared, permanent complexes have interfaces that are usually larger than transient interfaces [70].

2.11 Biological vs. Crystal Interfaces

X-ray crystallography, nuclear magnetic resonance (NMR), electron microscopy, and neutron diffraction are the most frequently used methods to determine structures. As of February 2021, 88.2% of all structures deposited in PDB have been obtained by using X-ray crystallography [71]. Even though it is the most commonly used method, X-ray crystallography poses a challenge in determining which structures are biologically relevant, and which are artifacts from the crystal packing [72] of proteins. The nonbiological interfaces that are formed as a result of the crystallization process are called crystal packing contacts, or in short crystal contacts [73]. Since these crystal contacts are not biologically relevant, they cause a noise in the analysis of protein–protein interfaces.

The increasing number of available three-dimensional structures of protein complexes enabled the identification of different physicochemical properties between biological and crystal interfaces [72]. These properties are beneficial for differentiating the biologically relevant interfaces. Generally, biological interfaces are more conserved in terms of amino acid composition, and they are thermodynamically more stable. Whereas crystal interfaces are usually formed as a result of kinetically driven associations, and they are usually nonspecific [74].

There are several tools available to help with differentiating between biological and crystal interfaces. These tools are energy based [30, 75], empirical knowledge based [76, 77], and machine learning based [66, 78, 79]. Correctly identifying the biological interfaces is important for correctly identifying the biologically relevant properties of protein–protein interfaces.

2.12 Type I, Type II, and Type III Interfaces

Interfaces can be divided into three types according to the structural similarity of the global structures of their monomers [18]. Usually, if two interfaces have similar structures in two different complexes, they are derived from globally similar protein chains. These interfaces are called Type I interfaces. Sometimes, the interfaces of complexes may be similar, but the global folds of the proteins that form the interfaces differ. These interfaces are called Type II interfaces. These proteins usually have different functions, and they may be good candidates for structural/functional studies. The last one, Type III interfaces represent a group of interfaces with only one side similar. This interface type suggests that proteins with different geometries can bind to the same site [80]. Figure 2.6 presents examples of each type. The left panel shows two complexes where the interfaces are similar. The two pairs of proteins interacting in the complexes are homologous (i.e. 1A7Q_H is similar to 1AP2_A and 1A7Q_L is similar to 1AP2_B). The middle panel represents two complexes where the proteins are non-homologous, yet the interface architectures are similar. The right panel shows that one protein (1AZZ_A is homologous to 1SR5_A) can bind to different proteins using a similar interface architecture.

When compared with Type III interfaces, Type I interfaces have larger interface areas. The average interface area for Type I interfaces is 1967 $Å^2$, whereas it is 1235 $Å^2$ for Type III interfaces. In addition, Type III interfaces have an average gap volume index of 3.21 Å, whereas Type I interfaces have a much smaller gap volume index of 1.98 Å [36]. Type I interfaces are more hydrophobic compared to Type III interfaces.

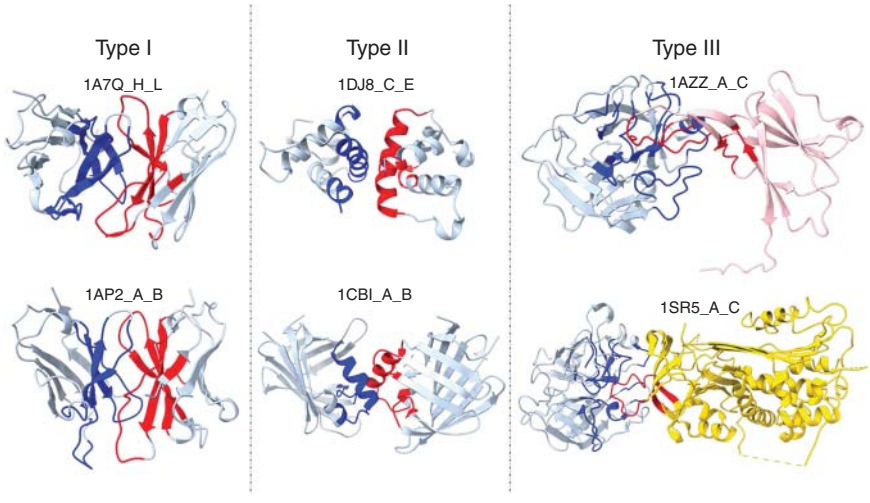


Figure 2.6 In this figure two members for each type of interface are shown. As can be seen, in Type I and Type II interfaces, interface regions of both interfaces are similar; in Type III interfaces, only one side of the interface is similar. Both for Type II and Type III interfaces, the globular folds of the proteins differ from each other.

2.13 Conserved Residues and Hot Spots in Interfaces

Protein interfaces are more conserved than the rest of the protein surfaces [81, 82]. Residues in obligate interfaces evolve at a relatively slower rate, allowing the residues in two interacting proteins to coevolve. On the other hand, residues in transient interfaces exhibit an increased rate of evolution thus with no evidence of correlated mutations across the interfaces [64]. Therefore, evolutionary conservation is an important feature to identify interfaces.

The conformational dynamics of conserved residues in unbound monomers illustrate significantly lower flexibility, suggesting that already before binding they are constrained in a bound-like configuration [83]. Backbone torsional angle distributions of conserved residues correspond to restricted regions of space and the most visited conformations in the bound and unbound trajectories are similar, suggesting that conserved residues are preorganized before binding.

When two proteins bind to each other, some critical residues, called hot spots, contribute more to the binding free energy [84]. Only a small portion of the interface residues are hot spots, and they are essential for protein interactions [8]. Hot spots are not randomly spread along with the protein–protein interfaces; rather, they tend to be clustered as hot regions [85, 86].

Several studies have tried to identify and characterize hot spots on protein–protein interfaces. Alanine Scanning Mutagenesis experiments are usually used to find hot spots. In these experiments, every interface residue is mutated to alanine and the corresponding changes in the binding affinity ($\Delta\Delta G$) are observed. The residues, which result in a significant reduction of the binding energy ($\geq$2 kcal) upon alanine mutagenesis, are considered hot spots. As this experimental procedure is resource intensive, computational methods are used frequently for hot spot prediction [87].

Computational methods might exploit the physicochemical properties of amino acids to find hot spots. These properties are mostly hydrophobicity, hydrophilicity, polarity, and average ASA. A previous study revealed an inverse correlation between binding energy and the ASA of individual residues upon complexation [20]. Previous studies have also demonstrated that the amino acid composition of hot spots is not random. Trp (21%), Arg (13.3%), and Tyr (12.3%) are the most frequent amino acids found as hot spots [88].

Some studies suggested energy-based methods to predict hot spots [89]. Over the last decade, various computational tools have been developed, including graph-based algorithms [85] and machine-learning-based approaches [90]. These methods use structure-based, sequence-based, and energy-based features. A comparison study has shown that one of the ensemble-learning algorithms, gradient tree boosting (GTB), and combining ASA-related properties and the position-specific scoring matrix (ASA + PSSM) achieved state-of-the-art results [91].

Figure 2.7 illustrates the importance of hot spots in pharmaceutical studies. Interleukin-2 (IL2) bound to its receptor (ILR2) is shown in the left part of the figure. IL2 is a cytokine that functions as a growth factor and central regulator in the immune system and mediates its effects through ligand-induced hetero-trimerization of the receptor subunits α, β, and γ. There are three hot regions

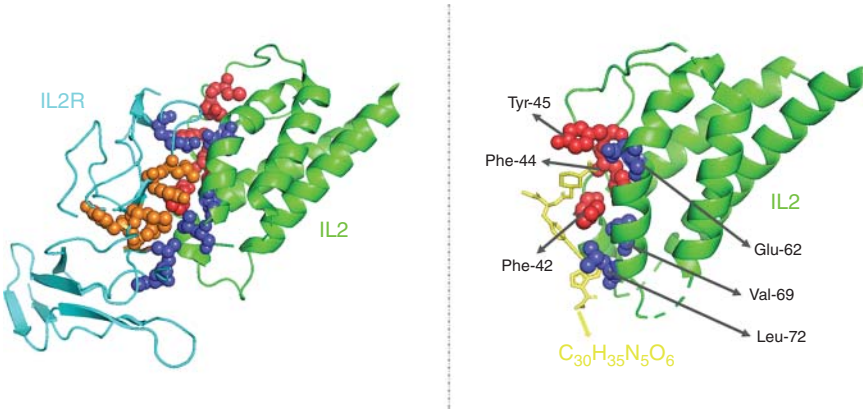


Figure 2.7 In the left figure, hot regions of Interleukin 2 (IL2) bound to its receptor (ILR2) are shown. (PDB ID: 1Z92) In the right figure, a small molecule binds to hot spots of IL2 and interferes with the PPI between IL2 and IL2R. Source: Adapted from Cukuroglu et al. [92]. (PDB ID: 1M49) Hot region information is obtained via the HotRegion server. Source: Adapted from Arkin et al. [85].

in the complex (red, blue, and orange). A previous study has shown that a small molecule binds to the IL2 binding site involving the residues Tyr-45, Phe-44, Phe-42, Glu-62, Val-69, and Leu-72 and thereby blocks the interaction of IL2 and its receptor [92].

2.14 Conclusion and Future Implications

Characterization of protein-binding interfaces is essential to understand protein–protein interactions. The accumulated knowledge in protein–protein interfaces is useful in various disease-related areas, such as drug discovery, phenotypic effects of single amino acid variations (SAVs) at interfaces, and prediction of host–pathogen interactions using interface mimicry.

More than 645 000 disease-associated PPIs in the human interactome have been reported [93]. Therefore, the identification of drug-like small molecules that disrupt disease-related protein–protein interactions and targeting interfaces between these disease-related proteins might have significant therapeutic potential. However, designing drug-like small molecules is a challenge as the PPI interfaces are usually flat and do not contain deep cavities. On the other side, drug-like small molecules tend to target specifically hot spot residues [94]. Therefore, computational prediction of hot spots might help significantly to identify druggable sites on the interfaces.

Missense mutations, which result in SAVs, lead to various diseases. SAVs affect the function of the protein and protein–protein interactions, thereby causing diseases [95, 96]. The studies on SAVs show that the disease-causing mutations tend to be located at the protein–protein interfaces rather than in other regions on the protein surface [97–101]. The location of SAVs within the binding interface is also

shown to be important because disease-causing SAVs tend to be located buried in the binding interface (e.g. interface core), rather than at the not buried and relatively solvent-accessible parts of the interface [99]. Disease-causing SAVs are also likely to be located at the hot spots rather than in hot regions, whereas benign SAVs are located at the non-hot spots [102]. On the other hand, for the hub proteins, cancer mutations are located at patches and are not singletons [103].

In the course of a disease caused by a pathogen (i.e. viral, or microbial infections), the pathogen proteins and host proteins compete to bind to the same binding partners [104, 105]. Interface mimicry is one of the strategies that pathogens use to compete with host proteins. Here, the binding interface of a pathogen protein has a high similarity to the binding interface of the competed human protein. Interface mimicry appears in both endogenous (between a human protein pair) and exogenous (between a pathogen protein and a human protein) interactions [106, 107]. Structurally similar binding interfaces permit proteins to interact with the same binding partners [36, 108]. Therefore, by mimicking the binding interface of a human protein, pathogen proteins might interact with the binding partners of those mimicked proteins even though their global structures are different. Having a comprehensive understanding of host–pathogen interactions and the role of binding interfaces in this respect is crucial to develop and advance antipathogenic therapies or drugs [104].

References

1 Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. (2008). Principles of protein–protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.* 108 (4): 1225–1244. https://doi.org/10.1021/cr040409x.

2 Minhas, F., Geiss, B.J., and Ben-Hur, A. (2014). PAIRpred: partner-specific prediction of interacting residues from sequence and structure. *Proteins* 82 (7): 1142–1155. https://doi.org/10.1002/prot.24479.

3 Xue, L.C., Dobbs, D., and Honavar, V. (2011). HomPPI: a class of sequence homology based protein–protein interface prediction methods. *BMC Bioinf.* 12: 244. https://doi.org/10.1186/1471-2105-12-244.

4 Hadarovich, A., Chakravarty, D., Tuzikov, A.V. et al. (2021). Structural motifs in protein cores and at protein–protein interfaces are different. *Protein Sci.* 30 (2): 381–390. https://doi.org/10.1002/pro.3996.

5 Ofran, Y. and Rost, B. (2003). Analysing six types of protein–protein interfaces. *J. Mol. Biol.* 325 (2): 377–387. https://doi.org/10.1016/s0022-2836(02)01223-8.

6 Xue, L.C., Dobbs, D., Bonvin, A.M., and Honavar, V. (2015). Computational prediction of protein interfaces: a review of data driven methods. *FEBS Lett.* 589 (23): 3516–3526. https://doi.org/10.1016/j.febslet.2015.10.003.

7 Tsai, C.J., Lin, S.L., Wolfson, H.J., and Nussinov, R. (1996a). A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.* 260 (4): 604–620. https://doi.org/10.1006/jmbi.1996.0424.

8 Tuncbag, N., Gursoy, A., and Keskin, O. (2009). Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics* 25 (12): 1513–1520. https://doi.org/10.1093/bioinformatics/btp240.

9 Hubbard, S. and Thornton, J. (1993). *NACCESS*. London: Department of Biochemisty Molecular Biology, University College.

10 Kleinjung, J. and Fraternali, F. (2005). POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic Acids Res.* 33 (Web Server issue): W342–W346. https://doi.org/10.1093/nar/gki369.

11 Mitternacht, S. (2016). FreeSASA: an open source C library for solvent accessible surface area calculations. *F1000Res* 5: 189. https://doi.org/10.12688/f1000research.7931.1.

12 Jones, S. and Thornton, J.M. (1997). Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* 272 (1): 121–132. https://doi.org/10.1006/jmbi.1997.1234.

13 Porollo, A. and Meller, J. (2007). Prediction-based fingerprints of protein–protein interactions. *Proteins* 66 (3): 630–645. https://doi.org/10.1002/prot.21248.

14 Levy, E.D. (2010). A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.* 403 (4): 660–670. https://doi.org/10.1016/j.jmb.2010.09.028.

15 Headd, J.J., Ban, Y.E., Brown, P. et al. (2007). Protein–protein interfaces: properties, preferences, and projections. *J. Proteome Res.* 6 (7): 2576–2586. https://doi.org/10.1021/pr070018+.

16 Lim, J., Ryu, S., Park, K. et al. (2019). Predicting drug-target interaction using a novel graph neural network with 3D structure-embedded graph representation. *J. Chem. Inf. Model.* 59 (9): 3981–3988. https://doi.org/10.1021/acs.jcim.9b00387.

17 Cukuroglu, E., Gursoy, A., Nussinov, R., and Keskin, O. (2014). Non-redundant unique interface structures as templates for modeling protein interactions. *PLoS One* 9 (1): e86738. https://doi.org/10.1371/journal.pone.0086738.

18 Keskin, O., Tsai, C.J., Wolfson, H., and Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci.* 13 (4): 1043–1055. https://doi.org/10.1110/ps.03484604.

19 Chakrabarti, P. and Janin, J. (2002). Dissecting protein–protein recognition sites. *Proteins* 47 (3): 334–343. https://doi.org/10.1002/prot.10085.

20 Guharoy, M. and Chakrabarti, P. (2005). Conservation and relative importance of residues across protein–protein interfaces. *PNAS* 102 (43): 15447–15452.

21 Schreiber, G. (2020). Protein–protein interaction interfaces and their functional implications. In: *Protein–Protein Interaction Regulators* (ed. S. Roy and H. Fu), 1–24. Royal Society of Chemistry https://doi.org/10.1039/9781788016544-00001.

22 Lin, J.J., Lin, Z.L., Hwang, J.K., and Huang, T.T. (2015). On the packing density of the unbound protein–protein interaction interface and its implications in dynamics. *BMC Bioinf.* 16 (Suppl 1): S7. https://doi.org/10.1186/1471-2105-16-S1-S7.

23 Berman, H.M., Westbrook, J., Feng, Z. et al. (2000). The protein data bank. *Nucleic Acids Res.* 28 (1): 235–242. https://doi.org/10.1093/nar/28.1.235.

24 Callaway, E. (2020). Revolutionary cryo-EM is taking over structural biology. *Nature* 578: 201. https://doi.org/10.1038/d41586-020-00341-9.

25 Abali, Z. (2021). *A Data-Centric Approach for Investigation of Protein–Protein Interfaces in Protein Data Bank*, 46–47. Turkey: Koç University.

26 Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011). Predicting protein–protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat. Protoc.* 6 (9): 1341–1354. https://doi.org/10.1038/nprot.2011.367.

27 Baspinar, A., Cukuroglu, E., Nussinov, R. et al. (2014). PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. *Nucleic Acids Res.* 42 (Web Server issue): W285–W289. https://doi.org/10.1093/nar/gku397.

28 Vangone, A., Spinelli, R., Scarano, V. et al. (2011). COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. *Bioinformatics* 27 (20): 2915–2916. https://doi.org/10.1093/bioinformatics/btr484.

29 Nagarajan, R., Archana, A., Thangakani, A.M. et al. (2016). PDBparam: online resource for computing structural parameters of proteins. *Bioinf. Biol. Insights* 10: 73–80. https://doi.org/10.4137/BBI.S38423.

30 Krissinel, E. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* 372 (3): 774–797. https://doi.org/10.1016/j.jmb.2007.05.022.

31 Laskowski, R.A., Jablonska, J., Pravda, L. et al. (2018). PDBsum: Structural summaries of PDB entries. *Protein Sci.* 27 (1): 129–134. https://doi.org/10.1002/pro.3289.

32 Kaleeckal Mathew, O. and Sowdhamini, R. (2016). PIMA: protein–protein interactions in macromolecular assembly – a web server for its analysis and visualization. *Bioinformation* 12 (1): 9–11. https://doi.org/10.6026/97320630012009.

33 Jordan, R. A., Wu, F., Dobbs, D., & Honavar, V. (In preparation). ProtinDb: A data base of protein–protein interface residues. http://protindb.cs.iastate.edu/

34 Xu, Q. and Dunbrack, R.L. Jr. (2020). ProtCID: a data resource for structural information on protein interactions. *Nat. Commun.* 11 (1): 711. https://doi.org/10.1038/s41467-020-14301-4.

35 Mosca, R., Ceol, A., Stein, A. et al. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* 42 (Database issue): D374–D379. https://doi.org/10.1093/nar/gkt887.

36 Keskin, O. and Nussinov, R. (2007). Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure* 15: 341–354.

37 Yan, C., Wu, F., Jernigan, R.L. et al. (2008). Characterization of protein–protein interfaces. *Protein J.* 27 (1): 59–70. https://doi.org/10.1007/s10930-007-9108-x.

38 Laskowski, R.A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics* 13 (5): 323, 307–330, 328. https://doi.org/10.1016/0263-7855(95)00073-9.

39 Sowmya, G., Anita, S., and Kangueane, P. (2011). Insights from the structural analysis of protein heterodimer interfaces. *Bioinformation* 6 (4): 137–143. https://doi.org/10.6026/97320630006137.

40 Jones, S. and Thornton, J.M. (1996). Principles of protein–protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* 93 (1): 13–20. https://doi.org/10.1073/pnas.93.1.13.

41 Tsai, C.J., Lin, S.L., Wolfson, H.J., and Nussinov, R. (1997). Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* 6 (1): 53–64. https://doi.org/10.1002/pro.5560060106.

42 Lazar, T., Guharoy, M., Schad, E., and Tompa, P. (2018). Unique Physicochemical Patterns of Residues in Protein–Protein Interfaces. *J. Chem. Inf. Model.* 58 (10): 2164–2173. https://doi.org/10.1021/acs.jcim.8b00270.

43 Savojardo, C., Manfredi, M., Martelli, P.L., and Casadio, R. (2020). Solvent accessibility of residues undergoing pathogenic variations in humans: from protein structures to protein sequences. *Front. Mol. Biosci.* 7 (626): 363. https://doi.org/10.3389/fmolb.2020.626363.

44 Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (12): 2577–2637. https://doi.org/10.1002/bip.360221211.

45 Touw, W.G., Baakman, C., Black, J. et al. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* 43 (Database issue): D364–D368. https://doi.org/10.1093/nar/gku1028.

46 Guharoy, M. and Chakrabarti, P. (2007). Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics* 23 (15): 1909–1918. https://doi.org/10.1093/bioinformatics/btm274.

47 Tuncbag, N., Gursoy, A., Guney, E. et al. (2008). Architectures and functional coverage of protein–protein interfaces. *J. Mol. Biol.* 381 (3): 785–802. https://doi.org/10.1016/j.jmb.2008.04.071.

48 Kastritis, P.L. and Bonvin, A.M. (2013). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J. R. Soc. Interface* 10 (79): 20120835. https://doi.org/10.1098/rsif.2012.0835.

49 Su, M., Yang, Q., Du, Y. et al. (2019). Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.* 59 (2): 895–913. https://doi.org/10.1021/acs.jcim.8b00545.

50 Jankauskaite, J., Jiménez-García, B., Dapkunas, J. et al. (2019). SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* 35 (3): 462–469. https://doi.org/10.1093/bioinformatics/bty635.

51 Erijman, A., Rosenthal, E., and Shifman, J.M. (2014). How structure defines affinity in protein–protein interactions. *PLoS One* 9 (10): e110085. https://doi.org/10.1371/journal.pone.0110085.

52 Sheinerman, F.B. and Honig, B. (2002). On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.* 318 (1): 161–177. https://doi.org/10.1016/s0022-2836(02)00030-x.

53 Xu, D., Lin, S.L., and Nussinov, R. (1997). Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *J. Mol. Biol.* 265 (1): 68–84. https://doi.org/10.1006/jmbi.1996.0712.

54 Kumar, S. and Nussinov, R. (2002). Close-range electrostatic interactions in proteins. *ChemBioChem* 3 (7): 604–617. https://doi.org/10.1002/1439-7633(20020703)3:7<604::AID-CBIC604>3.0.CO;2-X.

55 Kulandaisamy, A., Lathi, V., ViswaPoorani, K. et al. (2017). Important amino acid residues involved in folding and binding of protein–protein complexes. *Int. J. Biol. Macromol.* 94 (Pt A): 438–444. https://doi.org/10.1016/j.ijbiomac.2016.10.045.

56 Yugandhar, K. and Gromiha, M.M. (2014). Protein–protein binding affinity prediction from amino acid sequence. *Bioinformatics* 30 (24): 3583–3589. https://doi.org/10.1093/bioinformatics/btu580.

57 Chen, J., Sawyer, N., and Regan, L. (2013). Protein–protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.* 22 (4): 510–515. https://doi.org/10.1002/pro.2230.

58 Mou, Y., Huang, P.S., Hsu, F.C. et al. (2015). Computational design and experimental verification of a symmetric protein homodimer. *Proc. Natl. Acad. Sci. U.S.A.* 112 (34): 10714–10719. https://doi.org/10.1073/pnas.1505072112.

59 Bahadur, R.P., Chakrabarti, P., Rodier, F., and Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. *Proteins* 53 (3): 708–719. https://doi.org/10.1002/prot.10461.

60 Zhanhua, C., Gan, J.G., Lei, L. et al. (2005). Protein subunit interfaces: heterodimers versus homodimers. *Bioinformation* 1 (2): 28–39. https://doi.org/10.6026/97320630001028.

61 Acuner Ozbabacan, S.E., Engin, H.B., Gursoy, A., and Keskin, O. (2011). Transient protein–protein interactions. *Protein Eng. Des. Sel.* 24 (9): 635–648. https://doi.org/10.1093/protein/gzr025.

62 Perkins, J.R., Diboun, I., Dessailly, B.H. et al. (2010). Transient protein–protein interactions: structural, functional, and network properties. *Structure* 18 (10): 1233–1243. https://doi.org/10.1016/j.str.2010.08.007.

63 De, S., Krishnadev, O., Srinivasan, N., and Rekha, N. (2005). Interaction preferences across protein–protein interfaces of obligatory and non-obligatory components are different. *BMC Struct. Biol.* 5: 15. https://doi.org/10.1186/1472-6807-5-15.

64 Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* 102 (31): 10930–10935. https://doi.org/10.1073/pnas.0502667102.

65 Soner, S., Ozbek, P., Garzon, J.I. et al. (2015). DynaFace: discrimination between obligatory and non-obligatory protein–protein interactions based on the complex's dynamics. *PLoS Comput. Biol.* 11 (10): e1004461. https://doi.org/10.1371/journal.pcbi.1004461.

66 Zhu, H., Domingues, F.S., Sommer, I., and Lengauer, T. (2006). NOXclass: prediction of protein–protein interaction types. *BMC Bioinf.* 7: 27. https://doi.org/10.1186/1471-2105-7-27.

67 Stein, A., Pache, R.A., Bernado, P. et al. (2009). Dynamic interactions of proteins in complex networks: a more structured view. *FEBS J.* 276 (19): 5390–5405. https://doi.org/10.1111/j.1742-4658.2009.07251.x.

68 Nooren, I.M. and Thornton, J.M. (2003). Diversity of protein–protein interactions. *EMBO J.* 22 (14): 3486–3492. https://doi.org/10.1093/emboj/cdg359.

69 Block, P., Paern, J., Hullermeier, E. et al. (2006). Physicochemical descriptors to discriminate protein–protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins* 65 (3): 607–622. https://doi.org/10.1002/prot.21104.

70 Ansari, S. and Helms, V. (2005). Statistical analysis of predominantly transient protein–protein interfaces. *Proteins* 61 (2): 344–355. https://doi.org/10.1002/prot.20593.

71 *RCSB PDB Statistics* (2021). https://www.rcsb.org/stats/summary (accessed February 2021).

72 Elez, K., Bonvin, A., and Vangone, A. (2020). Biological vs. crystallographic protein interfaces: an overview of computational approaches for their classification. *Crystals* 10: 114. https://doi.org/10.3390/cryst10020114.

73 Capitani, G., Duarte, J.M., Baskaran, K. et al. (2016). Understanding the fabric of protein crystals: computational classification of biological interfaces and crystal contacts. *Bioinformatics* 32 (4): 481–489. https://doi.org/10.1093/bioinformatics/btv622.

74 Taudt, A., Arnold, A., and Pleiss, J. (2015). Simulation of protein association: kinetic pathways towards crystal contacts. *Phys. Rev. E: Stat. Nonlinear Soft Matter Phys.* 91 (3): 033311. https://doi.org/10.1103/PhysRevE.91.033311.

75 Yueh, C., Hall, D.R., Xia, B. et al. (2017). ClusPro-DC: dimer classification by the cluspro server for protein–protein docking. *J. Mol. Biol.* 429 (3): 372–381. https://doi.org/10.1016/j.jmb.2016.10.019.

76 Bliven, S., Lafita, A., Parker, A. et al. (2018). Automated evaluation of quaternary structures from protein crystals. *PLoS Comput. Biol.* 14 (4): e1006104. https://doi.org/10.1371/journal.pcbi.1006104.

77 Tsuchiya, Y., Kinoshita, K., Ito, N., and Nakamura, H. (2006). PreBI: prediction of biological interfaces of proteins in crystals. *Nucleic Acids Res.* 34 (Web Server issue): W320–W324. https://doi.org/10.1093/nar/gkl267.

78 Fukasawa, Y. and Tomii, K. (2019). Accurate classification of biological and non-biological interfaces in protein crystal structures using subtle covariation signals. *Sci. Rep.* 9 (1): 12603. https://doi.org/10.1038/s41598-019-48913-8.

79 Jimenez-Garcia, B., Elez, K., Koukos, P.I. et al. (2019). PRODIGY-crystal: a web-tool for classification of biological interfaces in protein complexes. *Bioinformatics* 35 (22): 4821–4823. https://doi.org/10.1093/bioinformatics/btz437.

80 Keskin, O. and Nussinov, R. (2005). Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng. Des. Sel.* 18 (1): 11–24. https://doi.org/10.1093/protein/gzh095.

81 Caffrey, D.R., Somaroo, S., Hughes, J.D. et al. (2004). Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* 13 (1): 190–202. https://doi.org/10.1110/ps.03323604.

82 Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *PNAS* 100 (10): 5772–5777.

83 Yogurtcu, O.N., Erdemli, S.B., Nussinov, R. et al. (2008). Restricted mobility of conserved residues in protein–protein interfaces in molecular simulations. *Biophys. J.* 94 (9): 3475–3485. https://doi.org/10.1529/biophysj.107.114835.

84 Clackson, T. and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science* 267 (5196): 383–386.

85 Cukuroglu, E., Gursoy, A., and Keskin, O. (2012). HotRegion: a database of predicted hot spot clusters. *Nucleic Acids Res.* 40 (D1): D829–D833.

86 Keskin, O., Ma, B., and Nussinov, R. (2005). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.* 345 (5): 1281–1294. https://doi.org/10.1016/j.jmb.2004.10.077.

87 DeLano, W.L. (2002). Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.* 12 (1): 14–20. https://doi.org/10.1016/s0959-440x(02)00283-x.

88 Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2007). Hot spots – A review of the protein–protein interface determinant amino-acid residues. *Proteins Struct. Funct. Bioinf.* 68 (4): 803–812.

89 Gonzalez-Ruiz, D. and Gohlke, H. (2006). Targeting protein–protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. *Curr. Med. Chem.* 13 (22): 2607–2625. https://doi.org/10.2174/092986706778201530.

90 Lin, X. and Zhang, X. (2018). Prediction of hot regions in PPIs based on improved local community structure detecting. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 15 (5): 1470–1479.

91 Liu, S., Liu, C., and Deng, L. (2018). Machine learning approaches for protein–protein interaction hot spot prediction: progress and comparative assessment. *Molecules* 23 (10): 2535.

92 Arkin, M.R., Randal, M., DeLano, W.L. et al. (2003). Binding of small molecules to an adaptive protein–protein interface. *PNAS* 100 (4): 1603–1608.

93 Heifetz, A., Sladek, V., Townsend-Nicholson, A., and Fedorov, D.G. (2020). Characterizing protein–protein interactions with the fragment molecular orbital method. In: *Quantum Mechanics in Drug Discovery* (ed. A. Heifetz), 187–205. Springer.

94 Arkin, M.R. and Wells, J.A. (2004). Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nat. Rev. Drug Discovery* 3 (4): 301–317. https://doi.org/10.1038/nrd1343.

95 Stefl, S., Nishi, H., Petukh, M. et al. (2013). Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.* 425: 3919–3936.

96 Yates, C.M. and Sternberg, M.J. (2013). The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein–protein interactions. *J. Mol. Biol.* 425: 3949–3963.

97 Butler, B.M., Gerek, Z.N., Kumar, S., and Ozkan, S.B. (2015). Conformational dynamics of nonsynonymous variants at protein interfaces reveals disease association. *Proteins* 83 (3): 428–435. https://doi.org/10.1002/prot.24748.

98 David, A., Razali, R., Wass, M.N., and Sternberg, M.J. (2012). Protein–protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.* 33 (2): 359–363. https://doi.org/10.1002/humu.21656.

99 David, A. and Sternberg, M.J. (2015). The contribution of missense mutations in core and rim residues of protein–protein interfaces to human disease. *J. Mol. Biol.* 427 (17): 2886–2898. https://doi.org/10.1016/j.jmb.2015.07.004.

100 Gao, M., Zhou, H., and Skolnick, J. (2015). Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure* 23 (7): 1362–1369. https://doi.org/10.1016/j.str.2015.03.028.

101 Jubb, H.C., Pandurangan, A.P., Turner, M.A. et al. (2017). Mutations at protein–protein interfaces: Small changes over big surfaces have large impacts on human health. *Prog. Biophys. Mol. Biol.* 128: 3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002.

102 Ozdemir, E.S., Gursoy, A., and Keskin, O. (2018). Analysis of single amino acid variations in singlet hot spots of protein–protein interfaces. *Bioinformatics* 34 (17): i795–i801. https://doi.org/10.1093/bioinformatics/bty569.

103 Dincer, C., Kaya, T., Keskin, O. et al. (2019). 3D spatial organization and network-guided comparison of mutation profiles in Glioblastoma reveals similarities across patients. *PLoS Comput. Biol.* 15 (9): e1006789. https://doi.org/10.1371/journal.pcbi.1006789.

104 Brito, A.F. and Pinney, J.W. (2017). Protein–protein interactions in virus-host systems. *Front. Microbiol.* 8: 1557.

105 Guven-Maiorov, E., Hakouz, A., Valjevac, S. et al. (2020). HMI-PRED: a web server for structural prediction of host–microbe interactions based on interface mimicry. *J. Mol. Biol.* 432 (11): 3395–3403. https://doi.org/10.1016/j.jmb.2020.01.025.

106 Franzosa, E.A. and Xia, Y. (2011). Structural principles within the human-virus protein–protein interaction network. *Proc. Natl. Acad. Sci. U.S.A.* 108 (26): 10538–10 543.

107 Muratcioglu, S., Guven-Maiorov, E., Keskin, Ö., and Gursoy, A. (2015). Advances in template-based protein docking by utilizing interfaces towards completing structural interactome. *Curr. Opin. Struct. Biol.* 35: 87–92.

108 Tsai, C.J., Lin, S.L., Wolfson, H.J., and Nussinov, R. (1996b). Protein–protein interfaces: architectures and interactions in protein–protein interfaces and in protein cores. Their similarities and differences. *Crit. Rev. Biochem. Mol. Biol.* 31: 127–152.

3

Correlated Coevolving Mutations at Protein–Protein Interfaces

Alexander Schug[1,2]

[1] *Jülich Supercomputing Centre, Forschungszentrum Jülich, Wilhelm-Johnen-Straβe, Jülich 52428, Germany*
[2] *Department of Biology, Faculty of Biology, University of Duisburg, Universitätsstraβe 5, Essen 45141, Germany*

3.1 Introduction

Life is organized hierarchically from the molecular level, where life is based on biomolecules such as DNA, proteins, or RNA and their interactions, over cells and their compartments [1] up to organs, organisms, or even whole ecosystems. At the molecular level, biomolecules are the key players. Despite their simple buildup from nucleic acids or amino acids biomolecules realize an incredible diversity of functions in living organisms. Examples include the storage, handling, and readout of genetic information in DNA, enzymatic function, molecular sensing and signaling, motion (e.g. muscles, cell motion) or structural stability (e.g. collagen, hair, or spider silk). To mechanistically understand biomolecular function, however, one must know the unique biomolecular structure, i.e. the three-dimensional arrangement of all atoms inside the biomolecule. Common experimental techniques used in structure determination have made incredible progress but also have their limits and are often quite involved. One of the oldest and best known method, X-ray crystallography, requires the growth of crystals of the investigated biomolecules and subsequent interpretation of scattering data. Nuclear magnetic resonance (NMR), in contrast, does not require such crystals and can directly be applied to biomolecules in solution but relies on the correct assignment of NMR shifts, which gets increasingly difficult for larger systems. The use of cryo election microscopy (cryoEM) has skyrocketed in the last decade, but still relies on automatized interpretation of large data sets and, thus, involves highly optimized workflows [2]. Small angle X-ray scattering (SAXS) is experimentally quite simple, but only provides low-resolution information which has to be carefully interpreted [3–5]. So are there possible theoretical complements to these experimental approaches?

In silico protein structure prediction has a long history [6–20] and can complement experimental work. Commonly summed up under "structure prediction tools", there are many approaches tackling the challenge of predicting biomolecular structures from protein sequence alone. Homology modeling tools rely on the

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

structural similarity of evolutionary related biomolecules and use experimentally resolved known structures as templates upon which unknown structures can be build. If no evolutionary similar structures are known, one could predict a biomolecular structure from its sequence alone by, e.g. identifying the global free-energy minimum in a suitable physics-based force field. Such a global search is challenging due to the gigantic search space. Any guidance toward the global minimum would support the search by reducing this search space. In 2009, a methods coined direct coupling analysis (DCA) provided such guidance by investigating the mutational patterns of coevolution in protein-protein interactions [21] and applied this approach to blind prediction of a protein complex [22]. As highlighted in Figure 3.1, coevolving residue pairs are considered spatially adjacent or contacts, as evolution puts constraints on mutations due to the need to maintain structure and function. While the general idea was already proposed in the 1990s, earlier methods [23–25] based on mutual information were plagued by a high number of false positive (FP) contact predictions due to only accounting for strictly pairwise correlations while disregarding the global context of other residues. DCA [21, 22, 26, 27] considers this global context and is based on inverse problems in statistical physics, so-called inverse Potts models. In short, DCA mimics fitness landscapes of proteins and drastically improves signal-to-noise ratios [28–31]. DCA has inspired similar approaches [32–37] for tracing coevolution. In a typical interpretation, such coevolving residue pairs are considered spatially adjacent contacts and exploited as structural constraints in molecular modeling tools for proteins [22, 32, 34, 37–44] but also for RNA [45–47]. Remarkably, the Hamiltonian can be regarded as a fitness landscape, and thus one may infer biomolecular function such as biological signaling [48], antibiotics resistance [30], or protein–protein interactions [49, 50].

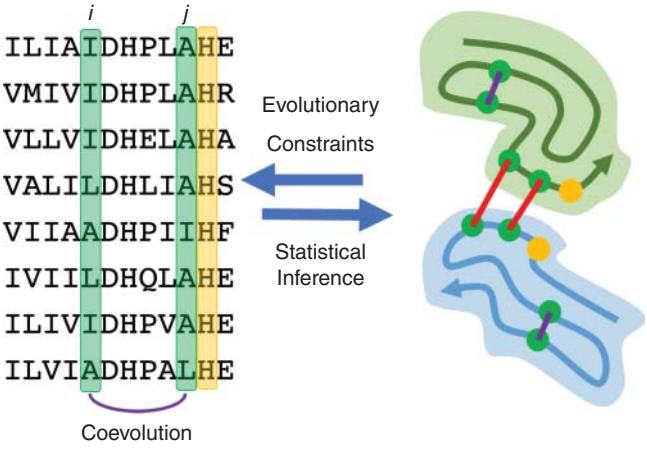


Figure 3.1 Coevolutionary analysis builds on the premise that a biomolecular 3D structure leaves an evolutionary imprint on the sequences of a protein family or in both protein families that form a complex, as a mutation at site *i* affects mutations at the spatially adjacent sites *j*. Statistical analysis can therefore infer such pairs of coevolving residues, both intramolecular (purple) or intermolecular (red). Specific functionally relevant residues (orange) are conserved across a protein family and will not show coevolutionary signals.

3.2 A Short Introduction into Biomolecular Modeling

A realistic theoretical description of biomolecules based on quantum mechanical (QM) *ab initio* approaches to accurately model electronic properties and atomic interactions is computationally extremely demanding. Therefore, the most common atomistic description of biomolecules is based on classical or Newtonian force fields and simplifies the QM interactions coarsely into *molecular mechanics*. Typical energy terms are divided into short ranged bonded and long-ranged nonbonded interactions.

Bonded interactions are named by counting the involved number of atoms as 1–2 or bond, 1–3 or angle, and 1–4 or dihedral interactions. The 1–2 interaction is a harmonic potential $V_{\mathrm{B}} = \epsilon_{\mathrm{B}}(x - x_0)^2$ (bond constant ϵ_{B}, distance x of involved atoms 1, 2 and their equilibrium distance x_0). The 1–3-interaction is also harmonic $V_{\mathrm{A}} = \epsilon_{\mathrm{A}}(\theta - \theta_0)^2$ (angle constant ϵ_{A}, angle between bonds of atoms (1,2) and (2, 3), θ_0 equilibrium angle). The 1–4-interaction is provided by $V_{\mathrm{D}} = \sum_{z=1,3} \epsilon_{z,\mathrm{D}} \left(1 - \cos n(\phi - \phi_0)\right)$ (dihedral constant $\epsilon_{z,\mathrm{D}}$, ϕ the angle or dihedral between the respective planes formed by atoms (1, 2, 3) and (2, 3, 4), equilibrium dihedral θ_0 and the multiplicity n).

In addition, there are typically two types of *nonbonded interactions*. The short-ranged *Lennard–Jones* potential can be written as $V_{\mathrm{LJ}} = \epsilon_{\mathrm{LJ}} \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma}{r_{ij}} \right)^{6} \right]$ (ϵ_{LJ} the potential strength, σ the equilibrium distance, r_{ij} interatomic distance of atoms i and j). Finally, the *electrostatics* term represents interactions resulting from two point charges $V_{\mathrm{ES}} = \epsilon_{\mathrm{ES}} \frac{q_i q_j}{4\pi\epsilon_0 \epsilon_{\mathrm{R}} r_{ij}}$ (ϵ_{ES} potential strength, q_i point charge of atom i, ϵ_0 electric constant, ϵ_{R} dielectric constant, r_{ij} distance between charged atoms i and j).

The total sum of all terms for all atoms is the molecular mechanics potential or *force field*. Some common force fields for biomolecular simulations are AMBER [51] or CHARMM [52]. Given the importance of water to biomolecules [53] and their interactions [54, 55], the solvent interactions have to be modeled as well, either explicitly or implicitly [56–63]. In structure prediction, the global minimum of the potential should represent the native fold and can be identified by, for example, stochastic global optimization methods such as simulated annealing and its variants for this task [9, 64–68].

3.3 Statistical Inference of Coevolution

3.3.1 Limitations of Local Statistical Inference

Protein interactions are the main actuator in biological signaling. Proteins need to interact specifically to prevent unwanted cross-talk, interact sufficiently strong to accommodate transfer of signaling molecules, and, at the same time, interact sufficiently weak or transient to allow dissociation after functional interactions. The main element of protein interactions is the interaction interface being stabilized

by the properties of the involved amino acids (see Chapter 2). If specific amino acids enable chemical functions (e.g. catalytic sites), these amino acids tend to be conserved in evolution. All other involved amino acids can more freely mutate in evolution but are still constrained by the need to maintain the overall interaction interface.

These general considerations led to the development of statistical methods to infer such mutational constraints, e.g. by scoring substitution patterns [23] or comparing single $f_i(\alpha)$ and pairwise amino acid frequencies $f_{ij}(\alpha, \beta)$ ($\alpha,\beta\in\{1,\ldots,q\}$ are typically the q naturally occurring amino acids plus gap) [25, 69, 70]. One can calculate the f_i, f_{ij} from a multiple sequence alignment (MSA) for a protein family or from a joint MSA for a complex. The sequences of such a protein family in a MSA are assumed to undergo selective pressure. Commonly, mutual information (MI) is then used to quantify coevolution of sites i, j

$$\mathrm{MI}_{ij} = \sum_{\alpha,\beta\in\{1,\ldots,q\}} f_{ij}(\alpha,\beta) \ln\left(\frac{f_{ij}(\alpha,\beta)}{f_i(\alpha) f_j(\beta)}\right) \tag{3.1}$$

with the sum running over all the possible amino acids. Here, high values of MI correlate with biological function, but MI is plagued by high numbers of false positive signals when interpreted, e.g., as spatial adjacency when above a threshold. How can we improve this statistical analysis?

3.3.2 Direct-Coupling Analysis – A Potts Model Based on Multiple Sequence Alignments

Direct-coupling analysis (DCA) [21, 22, 27] frames coevolution as an inverse problem based on statistical mechanics (cf. Figure 3.2). As above, the sequences in a protein family as found in the MSA are assumed to undergo selective pressure. Thus, a MSA should allow inferring the evolutionary dynamics based on the marginal distributions of single sites and pairs by, e.g., a maximum-entropy approach to derive a Boltzmann-type distribution.

$[q] = \{i \in \mathbb{N} | 1 \leq i \leq q\}$ is a q-letter alphabet of amino acids or nucleic acids for RNA plus the gap position in a MSA. For the rest of this section, I will focus on proteins, but the inference for RNA is analogous. L-tuples formed from $[q]$ provide protein sequences $\boldsymbol{\sigma} = \{\sigma_\nu\}_{\nu=1}^{L}$, where σ_ν is the amino acid in position ν for a protein of length L. An MSA is viewed as a random sampling of possible sequences $\boldsymbol{\sigma}$ of the entire protein family Γ (i.e. Γ are the set of possible L-tuples formed from $[q]$) and we want to infer the probability distribution. According to the maximum-entropy principle, the distribution P that best represents the data given prior knowledge maximizes the entropy function

$$S(P) = -\sum_{\boldsymbol{\sigma}\in\Gamma} P(\boldsymbol{\sigma}) \ln P(\boldsymbol{\sigma}) \tag{3.2}$$

The distribution maximizing the entropy is of the form

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\{-\mathcal{H}(\boldsymbol{\sigma})\} \tag{3.3}$$

with the Hamiltonian function $\mathcal{H}(\boldsymbol{\sigma}) = -\sum_{k=1}^{N} \lambda_k g_k(\boldsymbol{\sigma})$, the Lagrange multipliers $\{\lambda_k\}_{k=1}^{N}$, and the partition function $Z = \sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\mathcal{H}(\boldsymbol{\sigma})}$. We now need to find the Lagrange multipliers best describing our data.

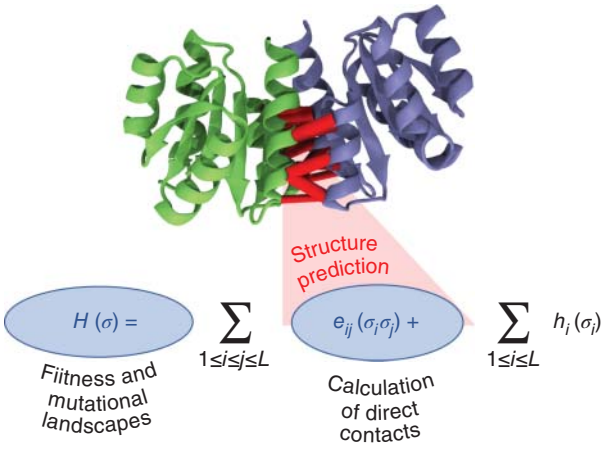


Figure 3.2 Once the inverse problem is solved, the DCA Hamiltonian can be interpreted. In the context of structure prediction, typically the coupling parameters e_{ij} are projected on a scalar such as the direction interaction score and high values interpreted as spatial adjacency of the involved residues i and j. The entire Hamiltonian can also be interpreted as a fitness landscape.

In DCA [21, 22, 27], we assume the pair-wise coupling i, j (e.g. stabilizing interactions) and single-site i behavior (e.g. active sites) to be reflected in the MSA:

$$\left\langle \delta_{\sigma_i,\alpha} \right\rangle = \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}|\sigma_i = \alpha) \tag{3.4}$$

$$\left\langle \delta_{\sigma_i,\alpha}\delta_{\sigma_j,\beta} \right\rangle = \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}|\sigma_i = \alpha, \sigma_j = \beta) \tag{3.5}$$

Ignoring numerical stability [21, 22, 27], marginal probabilities can be estimated from the MSA by direct frequency counts of the single sites $f_i(\alpha)$ and pairs $f_{ij}(\alpha, \beta)$ and we arrive at the Hamiltonian:

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_{1\le i\le L} h_i(\sigma_i) - \sum_{1\le i<j\le L} e_{ij}(\sigma_i, \sigma_j) \tag{3.6}$$

The matrix e of pairwise interactions is called the couplings matrix and the single site components h_i local fields. In statistical physics, this mathematical description is called a Potts model, a generalized model of the Ising model. The Potts model has $\binom{N}{2} q^2 + Nq$ inferred parameters, or $\binom{N}{2}(q-1)^2 + N(q-1)$ non-redundant constraints upon normalization, i.e. couplings and local fields are not uniquely defined. A common constraint is to impose gauge-fixation to reduce the parameter space.

3.4 Solving the Inverse Potts Model

Inferring the Hamiltonian from available sequence data requires solving an inverse problem, which was for biological sequence data first solved by Weigt and coworkers [21, 22] by DCA. Due to the finite nature of any MSA and the fact, that

sequences in the MSA are not a random subsample of possible sequences,[1] we can only approximately solve the inverse problem. It is also common to improve numerical robustness by including pseudo-count corrections λ in the restrictions of the marginals [21, 22]:

$$f_v(\alpha) = \frac{1}{\lambda q + M}\left[\lambda + \sum_{a=1}^{M} \delta_{\sigma_v,\alpha}\right] \tag{3.7}$$

$$f_{v\xi}(\alpha,\beta) = \frac{1}{\lambda q + M}\left[\frac{\lambda}{q} + \sum_{a=1}^{M} \delta_{\sigma_v,\alpha}\delta_{\sigma_\xi,\beta}\right] \tag{3.8}$$

(M is the size in sequences of the MSA) with minimal impact for large sample size ($M \gg \lambda q$), sampling re-weighting [27] or other refinements of input data. Several approaches have been developed to solve the inverse statistical problem. The original message-passing DCA [21, 22] is based on susceptibility propagation and computationally quite expensive as it scales as $O(L^4q^2)$. An improvement was mean-field direct-coupling analysis (mfDCA) [27], which considerably lowered computational cost. In mfDCA, the DCA Hamiltonian is decomposed into a noninteracting part $\mathcal{H}_0(\boldsymbol{\sigma}) = -\sum_{1\le i\le L} h_i(\sigma_i)$ and a couplings sector $\mathcal{H}(\boldsymbol{\sigma}) = \mathcal{H}_0(\boldsymbol{\sigma}) + \Delta\mathcal{H}(\boldsymbol{\sigma})$. Then, one introduces a trial noninteracting Hamiltonian $\mathcal{H}_0(\boldsymbol{\sigma}) + \langle\Delta\mathcal{H}(\boldsymbol{\sigma})\rangle_0$ (here $\langle X\rangle$ stands for the average of X over the canonical ensemble defined by the noninteracting part). Due to the Bogoliubov inequality $\mathcal{F} \le \mathcal{F}_0 + \langle\Delta\mathcal{H}(\boldsymbol{\sigma})\rangle_0$ mfDCA optimizes the local fields to ensure that the trial noninteracting model approximates the closest free energy to the actual system. In the first mfDCA [27] *Ursell functions* are calculated from the empirical frequencies and the corresponding matrix is inverted to recover the mean-field couplings.

While mfDCA is already computationally quite efficient, a subsequent approach based on pseudo-likelihood maximization DCA (plmDCA) [33, 71] presents another alternative. To infer the values of couplings and of single-site fields, the likelihood is substituted by the product of conditional probabilities of observing the variable σ_i^n given the ensemble of all the others ($\sigma_1^n \cdots \sigma_{i-1}^n \sigma_{i+1}^n \cdots \sigma_L^n$). In plmDCA, a maximization step proves to be the computational bottleneck, and different gradient descent algorithms are able to tackle this challenge. The most common one is the limited-memory Broyden–Fletcher–Goldfarb-Shanno (BFGS) algorithm [72] used as default by many plmDCA implementations [46, 73–76]. The large redundancy of parameters is solved by regularization [77]. A l_1-block regularization was first employed in [78]. Many plmDCA implementations use a l_2-regularization by adding $l_2 = \lambda_h \sum_{i=1}^{N} \|\mathbf{h}_i\|^2 + \lambda_J \sum_{1\le i\le j\le N} \|\mathbf{e}_{ij}\|^2$ to the pseudo-likelihood, which leads to an Ising-type gauge [75].

Finally, one typically takes the coupling matrices e_{ij} and condenses them into a scalar for scoring, e.g. by the Frobenius norm [35, 46, 73, 74]:

$$\mathrm{FN}_{ij} = \|e_{ij}\| = \sqrt{\sum_{k,l=1}^{q} e_{ij}(\alpha,\beta)^2} \tag{3.9}$$

1 Databases tend to focus on sequences that are either experimentally easy to access as they are from biologically relevant organisms, such as those of particular medical relevance. This leads to phylogenetic and other biases in the sequence data.

often combined with a averaged product correction [79] $\mathrm{APC}_{ij} = \mathrm{FN}_{ij} - \frac{\sum_i \mathrm{FN}_{ij} \sum_j \mathrm{FN}_{ij}}{\sum_i \sum_j \mathrm{FN}_{ij}}$. The highest scoring pairs of residues are then assumed to be spatially adjacent. Alternative scores such as direct information [21] exist. Typically, for proteins plmDCA provides higher accuracies than mfDCA at elevated computational costs, while for RNA both plmDCA and mfDCA provide similar results [47, 80].

3.5 Contact Guided Protein and RNA Structure Prediction

Experimental measurement of protein and RNA 3D structures is often quite involved, while the sequence databases grow exponentially and can be exploited by DCA. Here, one typically condenses the coupling matrices into a scalar S_{ij} (see above) and ranks or sorts them by value. These top ranked site-pairs are then inserted as distance constraints into molecular modeling tools to predict protein [22, 32, 38–40, 81, 82] or RNA [45, 46, 80] systems. Specific examples include all-atom models of globular proteins [32], membrane proteins [40, 41], proteins with multiple conformations [39, 40, 81], structural pattern in disordered proteins [83], or combined with NMR data [82]. For RNA, DCA has improved both secondary and tertiary structure prediction [45], which was quickly corroborated [46].

But what are the challenges? One big challenge is building a high quality MSA, as one needs to account for phylogenetic bias, nonrandom sampling of sequence space, etc. Also, while the top scoring contacts tend to be correct or true positive (TP), lower scoring contacts are more likely to be false negatives and only a fraction of all contacts can be predicted with good signal-to-noise ratios. Typically, the top L or $2L$ contacts are used. Lastly, the integration of predicted contacts into molecular modeling software is not unique and needs to be error tolerant. Also, the used modeling force fields are not perfect, i.e. the lowest energy might not be the native state of a protein.

3.6 Inter-Monomer Interaction and Signaling

Residue pair coevolution occurs also at inter-protein interfaces. Here, one typically performs a DCA analysis of possible contacts between the interacting proteins. These contacts are then used as constraints in docking the interacting proteins. The abundance of sequence data makes two-component signal transduction system (TCS) a common target of coevolutionary analysis [21, 22, 27, 48, 50]. This abundance is likely rooted in its wide-spread appearance as signal response systems in gram-negative and cyanobacteria. They are less frequent in eukaryotes and archaea. In fact, the first application of DCA was a blind prediction of a specific TCS [22, 26]. As predicting TCS exemplifies the general approach for predicting protein complexes nicely, I will quickly go over the crucial steps in the latter study. TCS are ubiquitous signal transduction systems in bacteria, hence even in 2009 there were many sequences available. To study this heterodimer, it was necessary

to build a concatenated MSA data of the *interacting* protein partners. Due to the possible presence of paralogs, the identification of the correct non-crosstalking interacting partner is challenging. Luckily for TCS, the two interacting partners, sensor histidine kinase (HK) and response regulator (RR), can be usually found adjacently within the same operon, which greatly simplifies building the common MSA. The HK receives an extracellular signal which affects its autophosphorylation rate. The chemical signal, i.e. the phosphoryl group, is then transduced between a highly conserved His of the HK and an Asp of the RR. This conserved His-Asp pair is invisible to DCA due to its immutability but provides an additional spatial constraint for docking HK and RR. Based on the DCA contacts at the HK–RR interfaces and this additional contact, is was possible to blindly predict the TCS complex within about 3.5 Å RMSD of an independently measured crystal structure [22].

For other classes of protein complexes, there are other challenges. The difficulty to build a joint MSA is greatly diminished for homodimeric complexes where a protein interacts with itself. Here, the challenge is to distinguish between inter-monomeric and intra-monomeric contacts, as it is unknown whether the coevolving contact pairs are formed within each monomer, between the copies of the monomer, or even simultaneously within and between. Also, contacts could be formed only in additional conformations, e.g. in conformational transitions or even in domain-swapping. One possibility to address this challenge is assuming that intra-monomeric contacts are not participating in the inter-monomer interaction and signaling taking place at the interface. Thus, one can simply rank all contacts by their score and exclude all contacts already formed within the (typically known) monomer. The remaining contacts can then be formed at the interface [37, 44, 84]. A problem with this approach is the large number of false positive contact predictions at the interface [44], which needs to be addressed by the modeling tools. A large scale study of ≈ 2000 homodimers [44] systematically identified several main results for homodimeric interfaces.

- Higher quality MSAs lead to significantly improved signal-to-noise ratios. Large protein families from the database could contain subfamilies with different binding modes, strongly distorting the statistical analysis.
- Larger interacting surface regions are better detected by DCA. Smaller interacting surface regions are more difficult to detect. This is not trivial, as one could assume that smaller interacting surfaces have stronger coevolutionary signals.
- The majority of predicted false positive contacts in the monomeric structure are in fact true positive contacts of the homodimeric interface. Most predicted contacts are thus formed intramonomeric, intermonomeric, or both, supporting the thesis of spatial adjacency contributing strongly to coevolution.

3.7 Summary

Coevolutionary analysis is a powerful toolkit to quantify evolutionary effects on biomolecular structures. Physics-driven methods such as DCA can be directly integrated into molecular modeling tools to predict a large variety of structures, ranging

from globular proteins to protein complexes and structures of RNA. Considering the ongoing growth of both sequence data and raw computational power, these and similar methods based on machine learning will continue to impact structural biology and complement advances in the experimental techniques.

References

1 Helms, V. (2018). *Principles of Computational Cell Biology: From Protein Complexes to Cellular Networks*. Wiley.

2 Röder, C., Kupreichyk, T., Gremer, L. et al. (2020). Cryo-EM structure of islet amyloid polypeptide fibrils reveals similarities with amyloid-β fibrils. *Nat. Struct. Mol. Biol.* 27 (7): 660–667.

3 Weiel, M., Reinartz, I., and Schug, A. (2019). Rapid interpretation of small-angle X-ray scattering data. *PLoS Comput. Biol.* 15 (3): e1006900.

4 Hermann, M.R. and Hub, J.S. (2019). SAXS-restrained ensemble simulations of intrinsically disordered proteins with commitment to the principle of maximum entropy. *J. Chem. Theory Comput.* 15 (9): 5103–5115.

5 Reinartz, I., Weiel, M., and Schug, A. (2020). FRET dyes significantly affect SAXS intensities of proteins. *Isr. J. Chem.* 60 (7): 725–734.

6 Lau, K.F. and Dill, K.A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22 (10): 3986–3997.

7 Hansmann, U.H.E., Okamoto, Y., and Eisenmenger, F. (1996). Molecular dynamics, Langevin and hybrid Monte Carlo simulations in a multicanonical ensemble. *Chem. Phys. Lett.* 259 (3–4): 321–330.

8 Hardin, C., Eastwood, M.P., Luthey-Schulten, Z., and Wolynes, P.G. (2000). Associative memory Hamiltonians for structure prediction without homology: alpha-helical proteins. *Proc. Natl. Acad. Sci. U.S.A.* 97 (26): 14235–14240.

9 Schug, A., Herges, T., and Wenzel, W. (2003). Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Lett.* 91 (15): 158102.

10 Neri, M., Anselmi, C., Cascella, M. et al. (2005). Coarse-grained model of proteins incorporating atomistic detail of the active site. *Phys. Rev. Lett.* 95 (21): 218102.

11 Bujnicki, J.M. (2006). Protein-structure prediction by recombination of fragments. *ChemBioChem* 7 (1): 19–27.

12 Zheng, W., Schafer, N.P., Davtyan, A. et al. (2012). Predictive energy landscapes for protein–protein association. *Proc. Natl. Acad. Sci. U.S.A.* 109 (47): 19244–19249.

13 Leaver-Fay, A., Tyka, M., Lewis, S.M. et al. (2011). ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules (eds. M.L. Johnson, L. Brand). In: *Methods in Enzymology*, vol. 487, 545–574. Elsevier.

14 Kurowski, M.A. and Bujnicki, J.M. (2003). Genesilico protein structure prediction meta-server. *Nucleic Acids Res.* 31 (13): 3305–3307.

15 Dill, K.A. and MacCallum, J.L. (2012). The protein-folding problem, 50 years on. *Science* 338 (6110): 1042–1046.

16 Moult, J., Fidelis, K., Kryshtafovych, A. et al. (2016). Critical assessment of methods of protein structure prediction: progress and new directions in round XI. *Proteins Struct. Funct. Bioinf.* 84: 4–14.

17 Adhikari, B. (2020). DEEPCON: protein contact prediction using dilated convolutional neural networks with dropout. *Bioinformatics* 36 (2): 470–477.

18 Fukuda, H. and Tomii, K. (2020). DeepECA: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. *BMC Bioinformatics* 21 (1): 1–15.

19 Wu, Q., Peng, Z., Anishchenko, I. et al. (2019). Protein contact prediction using metagenome sequence data and residual neural networks. *Bioinformatics* 36 (1): 41–48. https://doi.org/10.1093/bioinformatics/btz477.

20 Senior, A.W., Evans, R., Jumper, J. et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577 (7792): 706–710.

21 Weigt, M., White, R.A., Szurmant, H. et al. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* 106 (1): 67–72.

22 Schug, A., Weigt, M., Onuchic, J.N. et al. (2009). High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl. Acad. Sci. U.S.A.* 106 (52): 22124–22129.

23 Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins Struct. Funct. Bioinf.* 18 (4): 309–317.

24 Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci. U.S.A.* 91 (1): 98–102.

25 Lockless, S.W. and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286 (5438): 295–299.

26 Schug, A. and Onuchic, J.N. (2010). From protein folding to protein function and biomolecular binding by energy landscape theory. *Curr. Opin. Pharmacol.* 10 (6): 709–714.

27 Morcos, F., Pagnani, A., Lunt, B. et al. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* 108 (49): E1293–E1301.

28 De Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nat. Rev. Genet.* 14 (4): 249.

29 Morcos, F., Schafer, N.P., Cheng, R.R. et al. (2014). Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. U.S.A.* 111 (34): 12408–12413.

30 Figliuzzi, M., Jacquier, H., Schug, A. et al. (2015). Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.* 33 (1): 268–280.

31 Levy, R.M., Haldane, A., and Flynn, W.F. (2017). Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* 43: 55–62.

32 Marks, D.S., Colwell, L.J., Sheridan, R. et al. (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6 (12): e28766.

33 Aurell, E. and Ekeberg, M. (2012). Inverse Ising inference using all the data. *Phys. Rev. Lett.* 108 (9): 090201.

34 Jones, D.T., Buchan, D.W.A., Cozzetto, D., and Pontil, M. (2011). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28 (2): 184–190.

35 Ekeberg, M., Lövkvist, C., Lan, Y. et al. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E* 87 (1): 012707.

36 Michel, M., Hayat, S., Skwark, M.J. et al. (2014). PconsFold: improved contact predictions improve protein models. *Bioinformatics* 30 (17): i482–i488.

37 Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife* 3: e02030.

38 Sułkowska, J.I., Morcos, F., Weigt, M. et al. (2012). Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U.S.A.* 109 (26): 10340–10345.

39 Dago, A.E., Schug, A., Procaccini, A. et al. (2012). Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* 109 (26): E1733–E1742.

40 Hopf, T.A., Colwell, L.J., Sheridan, R. et al. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149 (7): 1607–1621.

41 Nugent, T. and Jones, D.T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl. Acad. Sci. U.S.A.* 109 (24): E1540–E1547.

42 Tian, P., Boomsma, W., Wang, Y. et al. (2014). Structure of a functional amyloid protein subunit computed using sequence variation. *J. Am. Chem. Soc.* 137 (1): 22–25.

43 Ovchinnikov, S., Kinch, L., Park, H. et al. (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* 4: e09248.

44 Uguzzoni, G., Lovis, S.J., Oteri, F. et al. (2017). Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc. Natl. Acad. Sci. U.S.A.* 114 (13): E2662–E2671.

45 De Leonardis, E., Lutz, B., Ratz, S. et al. (2015). Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.* 43 (21): 10444–10455.

46 Weinreb, C., Riesselman, A.J., Ingraham, J.B. et al. (2016). 3D RNA and functional interactions from evolutionary couplings. *Cell* 165 (4): 963–975.

47 Pucci, F. and Schug, A. (2019). Shedding light on the dark matter of the biomolecular structural universe: progress in RNA 3D structure prediction. *Methods* 162–163: 68–73. 10.1016/j.ymeth.2019.04.012.

48 Cheng, R.R., Morcos, F., Levine, H., and Onuchic, J.N. (2014). Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.* 111 (5): E563–E571.

49 Gueudré, T., Baldassi, C., Zamparo, M. et al. (2016). Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc. Natl. Acad. Sci. U.S.A.* 113 (43): 12186–12191.

50 Bitbol, A.-F., Dwyer, R.S., Colwell, L.J., and Wingreen, N.S. (2016). Inferring interaction partners from protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* 113 (43): 12180–12185.

51 Case, D.A. III, Cheatham, T.E., Darden, T. et al. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.* 26 (16): 1668–1688. ISSN 0192-8651 (Print) 0192-8651 (Linking). https://doi.org/10.1002/jcc.20290.

52 Brooks, B.R., Bruccoleri, R.E., Olafson, B.D. et al. (1983). CHARMM - A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4 (2): 187–217. URL <Go to ISI>://A1983QP42300010.

53 Helms, V. (2007). Protein dynamics tightly connected to the dynamics of surrounding and internal water molecules. *ChemPhysChem* 8 (1): 23–33.

54 Eyrisch, S. and Helms, V. (2007). Transient pockets on protein surfaces involved in protein–protein interaction. *J. Med. Chem.* 50 (15): 3457–3464.

55 Ahmad, M., Gu, W., Geyer, T., and Helms, V. (2011). Adhesive water networks facilitate binding of protein interfaces. *Nat. Commun.* 2 (1): 1–7.

56 Wagoner, J. and Baker, N.A. (2004). Solvation forces on biomolecular structures: a comparison of explicit solvent and Poisson–Boltzmann models. *J. Comput. Chem.* 25 (13): 1623–1629. ISSN 0192-8651 (Print) 0192-8651 (Linking). https://doi.org/10.1002/jcc.20089.

57 Donnini, S., Tegeler, F., Groenhof, G., and Grubmuller, H. (2011). Constant pH molecular dynamics in explicit solvent with lambda-dynamics. *J. Chem. Theory Comput.* 7 (6): 1962–1978. ISSN 1549-9626 (Electronic) 1549-9618 (Linking). https://doi.org/10.1021/ct200061r.

58 Zhuravlev, P.I., Wu, S., Potoyan, D.A. et al. (2010). Computing free energies of protein conformations from explicit solvent simulations. *Methods* 52 (1): 115–121. ISSN 1095-9130 (Electronic) 1046-2023 (Linking). https://doi.org/S1046-2023(10)00138-6 [pii] 10.1016/j.ymeth.2010.05.003.

59 Paschek, D., Nymeyer, H., and Garcia, A.E. (2007). Replica exchange simulation of reversible folding/unfolding of the trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. *J. Struct. Biol.* 157 (3): 524–533. ISSN 1047-8477 (Print) 1047-8477 (Linking). https://doi.org/S1047-8477(06)00329-7 [pii] 10.1016/j.jsb.2006.10.031.

60 Cheung, M.S., Garcia, A.E., and Onuchic, J.N. (2002). Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. U.S.A.* 99 (2): 685–690. URL <Go to ISI>://000173450100029.

61 Chen, J. and Brooks, C.L. III (2008). Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys. Chem. Chem. Phys.* 10 (4): 471–481. ISSN 1463-9076 (Print) 1463-9076 (Linking). https://doi.org/10.1039/b714141f.

62 Chen, J., Brooks, C.L. III, and Khandogin, J. (2008). Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin.*

Struct. Biol. 18 (2): 140–148. ISSN 0959-440X (Print) 0959-440X (Linking). https://doi.org/S0959-440X(08)00007-9 [pii] 10.1016/j.sbi.2008.01.003.

63 Gallicchio, E. and Levy, R.M. (2004). AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem.* 25 (4): 479–499. ISSN 0192-8651 (Print) 0192-8651 (Linking). https://doi.org/10.1002/jcc.10400.

64 Brooks, S.P. and Morgan, B.J.T. (1995). Optimization using simulated annealing. *The Statistician* 44 (2): 241–257.

65 Herges, T., Schug, A., Merlitz, H., and Wenzel, W. (2003). Stochastic optimization methods for structure prediction of biomolecular nanoscale systems. *Nanotechnology* 14 (11): 1161–1167. URL <Go to ISI>://WOS:000187038400003.

66 Schug, A. and Wenzel, W. (2004). Predictive in silico all-atom folding of a four-helix protein with a free-energy model. *J. Am. Chem. Soc.* 126 (51): 16736–16737. https://doi.org/10.1021/ja0453681—ISSN 0002-7863. URL <Go to ISI>://WOS:000225910400026.

67 Schug, A., Wenzel, W., and Hansmann, U.H.E. (2005). Energy landscape paving simulations of the trp-cage protein. *J. Chem. Phys.* 122 (19). https://doi.org/194711 Artn 194711. URL <Go to ISI>://000229743500056.

68 Das, R. and Baker, D. (2008). Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* 77: 363–382. ISSN 0066-4154 (Print) 0066-4154 (Linking). https://doi.org/10.1146/annurev.biochem.77.062906.171838.

69 Kass, I. and Horovitz, A. (2002). Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins Struct. Funct. Bioinf.* 48 (4): 611–617. https://doi.org/10.1002/prot.10180.

70 White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2007). [4] - Features of protein–protein interactions in two?component signaling deduced from genomic libraries. In: *Two?Component Signaling Systems, Part A*, *Methods in Enzymology*, vol. 422 (ed. M.I. Simon, B.R. Crane, and A. Crane), 75–101. Academic Press. https://doi.org/https://doi.org/10.1016/S0076-6879(06)22004-4.

71 Besag, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc. Ser. D (Statistician)* 24 (3): 179–195.

72 Byrd, R.H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* 16: 1190–1208. https://doi.org/10.1137/0916069.

73 Seemayer, S., Gruber, M., and Söding, J. (2014). CCMpred-fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* 30 (21): 3128–3130. https://doi.org/10.1093/bioinformatics/btu500.

74 Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proc. Natl. Acad. Sci. U.S.A.* 110 (39): 15674–15679.

75 Ekeberg, M., Hartonen, T., and Aurell, E. (2014). Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.* 276: 341–356. https://doi.org/https://doi.org/10.1016/j.jcp.2014.07.024.

76 Zerihun, M.B., Pucci, F., Peter, E.K., and Schug, A. (2020). pydca v1. 0: a comprehensive software for direct coupling analysis of RNA and protein sequences. *Bioinformatics* 36 (7): 2264–2265.

77 Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.* 4 (1): 1–106. https://doi.org/10.1561/2200000015.

78 Balakrishnan, S., Kamisetty, H., Carbonell, J.G. et al. (2011). Learning generative models for protein fold families. *Proteins Struct. Funct. Bioinf.* 79 (4): 1061–1078. (https://doi.org/10.1002/prot.22934.

79 Dunn, S.D., Wahl, L.M., and Gloor, G.B. (2007). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24 (3): 333–340. https://doi.org/10.1093/bioinformatics/btm604.

80 Pucci, F., Zerihun, M.B., Peter, E.K., and Schug, A. (2020). Evaluating DCA-based method performances for RNA contact prediction by a well-curated data set. *RNA* 26 (7): 794–802.

81 Morcos, F., Jana, B., Hwa, T., and Onuchic, J.N. (2013). Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U.S.A.* 110 (51): 20533–20538.

82 Tang, Y., Huang, Y.J., Hopf, T.A. et al. (2015). Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat. Methods* 12 (8): 751–754.

83 Toth-Petroczy, A., Palmedo, P., Ingraham, J. et al. (2016). Structured states of disordered proteins from genomic sequences. *Cell* 167 (1): 158–170.

84 Dos Santos, R.N., Morcos, F., Jana, B. et al. (2015). Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci. Rep.* 5: 13652.

4

Computational Protein–Protein Docking

Martin Zacharias

Technical University Munich, Physics Department and Center of Functional Protein Assemblies, Ernst-Otto-Fischer-Str. 8, D-85748 Garching, Germany

4.1 Introduction

Proteins are essential for basically all processes in living systems including metabolic and transport processes as well as transcription and translation of genetic information or mediating signal transduction. Although some of the biological tasks are processed by single proteins that can act as enzymes, for example, to catalyze metabolic reactions, a great majority of cellular functions depend on complexes that consist of protein subunits. These complexes are well-ordered systems that require specific assembly of proteins but can also involve the association with other biological molecules such as RNA or DNA [1, 2]. Understanding the function of such assemblies requires the knowledge of the structure and the molecular details of all involved interactions [1, 3, 4].

In recent years, the possibility to modify or design completely new artificial protein–protein interactions and multi-protein complexes of desired function has become increasingly attractive [1, 5, 6]. It may offer the rational creation of new enzymes catalyzing chemical reactions not easily controllable by common chemical approaches or allows one to construct artificial cell-like systems consisting of designed proteins and protein–protein complexes. Both the analysis of natural complexes as well as design of synthetic complexes require structural knowledge and understanding of the associated molecular interactions.

The experimental determination of protein–protein complex structures is one option and indeed the number of solved complex structures is increasing steadily [7]. So far, most protein–protein complex structures have been determined by X-ray crystallography [8]. The approach, however, requires the formation of highly ordered regular crystals. It involves purification of large amounts of partner proteins and the formed complexes need to be relatively stable for crystallization. Often multi-protein complexes and in particular transient complexes are in a dynamic equilibrium possibly interfering with stable crystallization [9]. Nuclear magnetic resonance (NMR) spectroscopy [10, 11] is another experimental method

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

for the structure determination of mostly dimeric complexes. It is limited to the high-resolution structure determination of protein–protein complexes of small size (<20 kDa). However, if the structure of the partner proteins is known it can be used to assist in structural modeling and affinity determination even of large multi-protein complexes [11]. New developments in the area of CryoEM (electron microscopy under cryogenic conditions) often allow rapid structure determination in particular of large biological assemblies [12, 13]. For a structure determination by CryoEM, a sufficient number of object images from different viewpoints but no crystals are required. The object images are combined and averaged to obtain a high-resolution view of the structure [13]. Structures of small particles are not accessible because of the limited contrast relative to a noisy background.

Hence, despite great progress, in the foreseeable future, it will still be impossible to determine experimentally all putative and transient protein–protein complexes of a cell [4]. However, there are very rapid experimental techniques [14, 15] and coevolution bioinformatics methods ([16], see also Chapter 3) to identify protein–protein interactions and to elucidate a network of all interactions in a cell or between a pathogen (e.g. virus) and the host cell. Techniques like yeast-two-hybrid screens [14] or *in vivo* chemical cross-linking [17] can detect association between protein partners without providing structures. Eventually, the knowledge of atomic resolution structures of all protein–protein interactions in a cell is desired. Hence, accurate structure prediction and structural modeling are of importance to provide realistic structural models of complexes [18–22]. Protein–protein docking refers to computational methods for predicting binary protein–protein complexes or prediction of complexes with several partners using the isolated protein structures or structural models as input. Several docking methodologies have been developed in recent years and with the help of the community-wide Critical Assessment of PRedicted Interactions (CAPRI) experiment [23–25] the progress in protein–protein docking prediction methods has been extensively monitored over more than the last 20 years. In this challenge, participating groups test the performance of docking methods in blind predictions of protein–protein complex structures that help to foster progress and ideas in the field. The CAPRI consortium has also defined useful criteria to evaluate the quality of a predicted protein–protein complex (if the experimental structure is known) in terms of structural deviation from the experimental reference (native complex) and number of native contacts between atoms at the predicted interface (Table 4.1).

In general, the binding of biological macromolecules is driven by the associated change in free energy which depends on the structural and physicochemical properties of the binding partners. It is influenced by direct interactions between binding partners but also by the surrounding solvent and ions and the change in average conformation and conformational fluctuations upon association. Optimal complementarity at binding interfaces was proposed by E. Fischer [26] as a decisive factor to achieve high affinity and specificity of binding. In fact, the majority of protein–protein docking methodologies consider the partner proteins as rigid

Table 4.1 Quality criteria for protein–protein docking according to the CAPRI challenge.

Quality/Criteria	Incorrect	Acceptable (*)	Medium (**)	High (***)
% Native contacts	<10	≥10	≥30	≥50
Ligand-RMSD	–	5 Å < RMSD ≤ 10 Å	1 Å < RMSD ≤5 Å	RMSD ≤1 Å
Interface RMSD	–	or 2 Å < RMSD ≤ 4 Å	or 1 Å < RMSD ≤ 2 Å	or RMSD ≤1 Å

The quality criteria are evaluated with respect to the experimental native complex. "% Native contacts" indicates the percentage of predicted contacts (atoms pair with distances <5 Å) that are also found in the native complex. RMSD indicates root-mean-square deviation with respect to the native complex. Here, "Ligand-RMSD" refers to the RMSD of the smaller ligand protein from the native placement after best superposition of the larger protein partner (receptor) on the native complex. "Interface-RMSD" is the RMSD of all interface atoms (all atoms within 10 Å of the partners in the native complex) after best superposition onto the native interface.
The stars indicate the quality of the solution, a one star (*) solution is acceptable, (**) medium, (***) high quality solution.

structures and use optimal complementarity as the main criteria for selecting possible binding geometries [19, 21]. It considerably simplifies the computational search problem, and in the first section of this chapter, the principles of rigid protein–protein docking methods will be introduced. However, binding can involve conformational changes, and efforts to efficiently account for such changes during docking will also be discussed. A schematic outline of the protein–protein docking methodology and associated tasks is given in Figure 4.1. In many cases, experimental data or data from bioinformatics resources are helpful to guide the docking search or re-score docked complexes. In the subsequent sections of the chapter, the possibility to use existing complexes as templates for docking and the refinement and final scoring of predictions will be covered followed by a discussion of future challenges in the field.

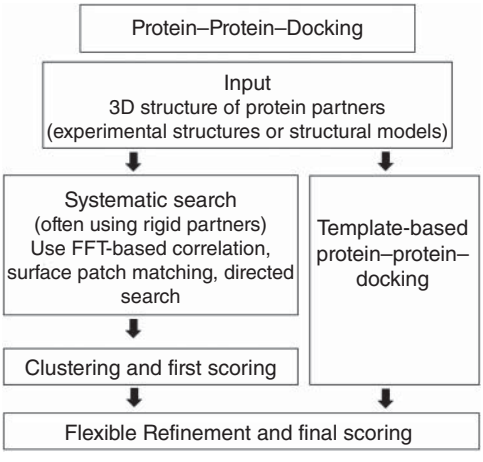


Figure 4.1 Schematic flow chart of the most important steps of computational protein–protein docking approaches.

4.2 Rigid Body Protein–Protein Docking Approaches

In case of treating protein partners as rigid units, the configurational variables are limited to three translational and three rotational degrees of freedom for each protein. The computational task is then to generate possible geometries of protein–protein complexes with complementarity at putative binding interfaces. Some degree of steric overlap between docking partners can be tolerated to implicitly account for conformational adjustments upon association (Figure 4.2). There are several efficient computational methods available for the rapid generation of putative binding geometries. It has been recognized by Katchalski–Katzir and coworkers [27] already in the early 1990s that fast Fourier transform (FFT) correlation techniques are well suited to efficiently locate overlaps between two complementary protein surfaces. In the standard setting of the FFT-docking approach, the partners are mapped onto cubic grids. In the second step, the grid points are given different values to indicate the interior or the surface of a protein and to mark the outside space of the protein (Figure 4.2a). With such representation, a complementarity score can be calculated for the two proteins by calculating the

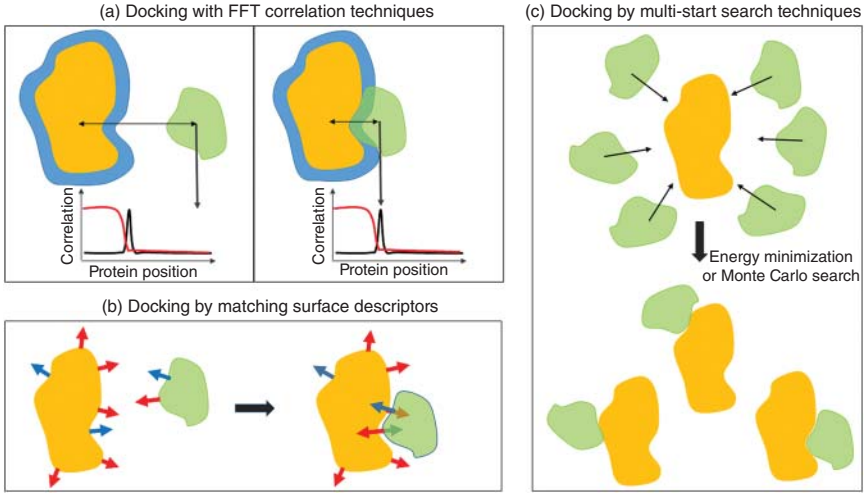


Figure 4.2 Illustration of the most common types of systematic docking methodologies. (a) Docking based on solving a correlation task [27]: The overlap of one partner (green) with a surface (blue) of the second partner protein (yellow) can be calculated by solving a correlation integral. The correlation is favorable as long as the ligand overlaps with the surface region (black line in (a)) but becomes unfavorable with increasing overlap with the interior of the receptor (red line). Using a grid to discretize the protein the correlation can be solved rapidly by fast Fourier transformation (FFT) and the best possible overlap for a given relative protein orientation can be extracted. (b) Geometric hashing can be used to match surface descriptors on both protein partners [28]. In this case, concave (blue arrows) and convex areas (red arrows) on the protein surface are illustrated. Docking is performed by finding best matches of concave and convex regions (other matching characteristics can also be used). (c) Docking by systematic energy minimization or dynamics simulations starts from thousands of different relative placements of the partner proteins (upper part of panel (c)) to obtain locally stable putative docking geometries (lower part of the panel in (c)).

correlation or correlation integral of just the two grids representing each protein or protein surface. The correlation can be approximated by the sum of all the pair products of the grid entries.

Such calculation is rather time-consuming but one can make use of the Fourier correlation theorem. Hence, the corresponding correlation integrals (or discrete sums) can be computed easily in Fourier space. Typically, one protein partner is termed receptor (usually the larger partner) and the other is the ligand protein. The discrete Fourier transform for the receptor grid needs to be calculated only once. Due to the special shifting properties of Fourier transforms, the different translations of the ligand grid with respect to the receptor grid can be computed by a simple multiplication in Fourier space. This process is repeated for various relative orientations of the two proteins. Several available computer programs for protein–protein docking use the Cartesian FFT algorithm and it is also part of most protein–protein docking web-servers (Table 4.2). In case of the standard Cartesian FFT methods, a FFT for each orientation of the ligand-protein relative to the receptor is required. Rotation steps of 10–15° are typically used but they can vary for different programs [35].

Instead of working with Cartesian coordinates, it is also possible to correlate in spherical polar basis functions that represent, for example, the shape of the protein surface [50]. Instead of solving the translation correlation task only for discrete relative orientations of the proteins in Cartesian space, it is then also possible to solve the whole multi-dimensional (orientation and translation) search in Fourier space [51]. The FFT correlation techniques enable conducting rigid protein–protein docking for typically sized protein partners in a few minutes on a workstation computer. It is also possible to rank solutions with methods that allow the correlation of multi-term potentials. In this case, the scoring function needs to be expressed in terms of spherical basis functions characterizing the surface properties of the protein partners [52]. The full partition function of the rigid docking problem can be obtained and thermodynamic and structural properties can be extracted [51]. One drawback of FFT correlation methods is, however, that they can be applied to protein dimer prediction but require sequential application when it comes to molecular assemblies consisting of several protein partners.

Geometric hashing of surface properties is an alternative approach that can also be used to rapidly produce putative protein–protein complex geometries (Figure 4.2b). Typically, a set of triangles approximately representing the protein surface are stored in a hash table. The triangles represent geometrical (concave, convex) or physicochemical (polar, hydrophobic) features of a surface segment. Matching of complementary triangles from the two different molecules produces putative complex geometries. Similar matching triangles on the surface of protein partners can be quickly identified by means of a hash key. PatchDock is an example program employing geometric hashing [28, 49, 53].

Protein–protein docking can also be performed using directed or guided search methods based on Molecular Dynamics (MD), Brownian Dynamics (BD), Monte Carlo (MC) simulations, or multi-start energy minimization. It is possible to limit the simulation degrees of freedom to translation and orientation of rigid partner

Table 4.2 Protein–protein docking program examples and associated websites or web servers.

Correlation FFT	3D Dock [29]	http://www.sbg.bio.ic.ac.uk/docking
Correlation FFT	ClusPro [30]	https://cluspro.bu.edu/login.php
Correlation FFT	DOT [31]	https://www.sdsc.edu/CCMS/DOT
Correlation FFT	GRAMM-X [32]	http://vakser.compbio.ku.edu/resources/gramm/grammx
Correlation polar FFT	Hex [33]	http://hexserver.loria.fr
Correlation FFT	MolFit [34]	www.weizmann.ac.il/Chemical_Research_Support/molfit
Correlation FFT	ZDock [35]	http://zdock.umassmed.edu
Correlation FFT	MEGADOCK [36]	www.bi.cs.titech.ac.jp/megadock
Correlation FFT	F2Dock [37]	http://www.cs.utexas.edu/~bajaj/cvc/software/f2dock.shtml
Correlation FFT	pyDock [38]	https://life.bsc.es/pid/pydockweb
Guided: MC minimization	PROBE [39]	http://pallab.serc.iisc.ernet.in/probe
Guided: Multi-start Energy minimization	ATTRACT [40, 41], PTOOLS [42]	www.attract.ph.tum.de
Guided: Multi-start Energy minimization	HawkDock [43]	http://cadd.zju.edu.cn/hawkdock
Guided: distance restraints	FroDock [44]	http://frodock.chaconlab.org
Monte Carlo on a Swarm	SwarmDock [45]	https://bmm.crick.ac.uk/~svc-bmm-swarmdock
Monte Carlo + minimization	RosettaDock [46]	https://www.rosettacommons.org/software
Guided: data-driven, MD + Energy minimization	HADDOCK [47]	http://haddock.chem.uu.nl
Guided: MC minimization	ICM-Disco [48]	http://www.molsoft.com/icm_pro.html
Geometric hashing	PatchDock [49]	https://bioinfo3d.cs.tau.ac.il/PatchDock
Geometric hashing	SymmDock [28]	https://bioinfo3d.cs.tau.ac.il/SymmDock

proteins [19]. However, the computational efforts of such approaches are higher than FFT correlation methods or geometric hashing, but their great advantage is the possibility to include different types of conformational flexibility already at the initial systematic search step. In addition, it is rather straightforward to extend them to simultaneous docking of several protein partners. Coarse-grained (reduced instead of atomistic) representations of amino acids in proteins are often employed to further reduce computational costs [40, 54, 55]. Examples for this type of approaches are the ATTRACT [40], HawkDock [43], RosettaDock [46, 56], SwarmDock [45], and HADDOCK [57] programs (see also Table 4.2).

4.3 Accounting for Conformational Changes during Docking

The success of a protein–protein docking search is judged by two main measures. First, it is important to obtain docking solutions that are as close as possible relative to the real complex structure. Such solutions are usually termed best docking prediction. Second, the predicted structures that resemble the native complex most closely should also give the best energy or docking score (the best energy or best-scored complex is called the top rank solution). During development of protein–protein docking programs, it was quickly realized that docking of bound partner structures is typically much more successful than using unbound partner structures [19, 22, 58]. The term bound partner structures refers here to protein structures that have been extracted from the known complex structure whereas unbound structures have been determined in the absence of the partner (Figure 4.3). Conformational differences between bound and unbound structures can cause both an increased deviation of the predicted complex structure closest to the native complex and a decrease in ranking or scoring of such near-native predicted complex (Figure 4.3). Depending on the target difficulty indicated as a degree of conformational changes accompanying the binding process, it is possible that none

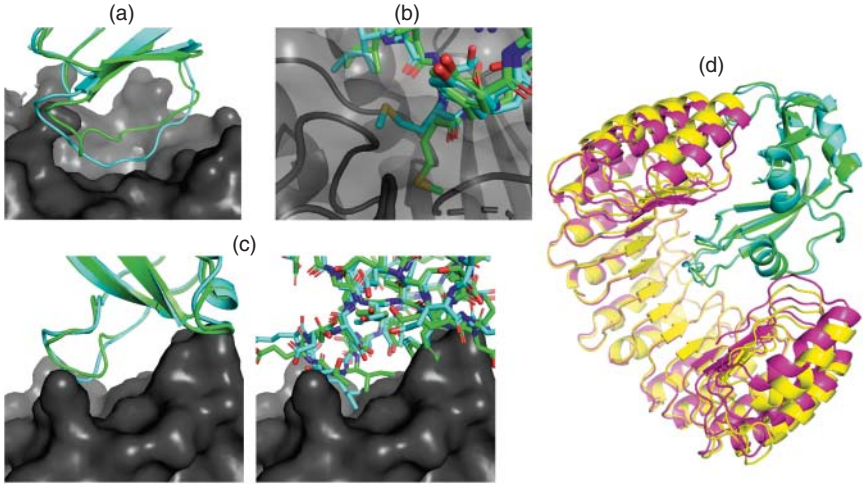


Figure 4.3 Types of conformational changes associated with protein–protein complex formation. (a) Change in backbone loop conformation illustrated for the protein–protein complex PDB id 1ACB with the unbound enzyme inhibitor as green cartoon and the bound structure as blue cartoon (enzyme receptor as gray surface representation). (b) Change in side-chain conformation (stick model) between unbound (green) and bound conformation (blue; receptor in gray) for enzyme-inhibitor complex PDB id 2SNI. (c) Example of backbone and side-chain conformational difference between unbound (green) and bound structure (light blue) of one partner protein in the complex PDB id 1EFN. (d) Global backbone changes (opening/closing motion) are observed by comparing the cartoon representation of the Ribonuclease A inhibitor in the unbound (magenta, PDB id 2bnh) vs. the bound forms (yellow, PDB id 1DFJ). The unbound (green) vs. bound (blue) structures of RNAseA are also shown.

of the rigidly docked complexes comes sufficiently close to the native complex to detect it as most realistic prediction [24, 25]. Hence, especially in these cases, it can be desirable to efficiently account for local side chains and loop transitions at the protein surface but also account for global motions of large protein domains already during the systematic docking search [18, 19, 22, 56]. Often experimental partner structures are not available but only of homologs and structural models of one or both partners need to be generated using comparative modeling. These homology-modeled structures are typically of lower accuracy than experimentally determined structures and can contain side chain or loop misplacements that can also interfere with rigid docking of the partner proteins [59].

Representing the partner proteins by multiple conformations (ensemble of structures) is a simple and straightforward approach to account indirectly for flexibility [60, 61]. To limit the size of the ensemble, protein structures can be generated along sterically allowed deformation directions (e.g. distance or orientation of separate domains) [19, 62, 63]. It is then possible to directly employ rigid docking approaches; however, this results in more docking solutions and possibly additional false-positive geometries.

Several docking approaches include explicit modeling of both side chains as well as backbone changes during a systematic search at the expense of much larger computational efforts compared to rigid docking [40, 45, 46, 64, 65]. In many cases, the partner proteins undergo not only local adjustments (e.g. conformational adaptation of side chains and backbone relaxation at the interface) but also more global conformational changes that involve domain opening–closing motions [22]. One can use MD simulations of protein partners to detect such global motions but alternatively, very rapid elastic network models (ENM) can also be used to detect sterically allowed global changes [66, 67]. ENM calculations are based on modeling interactions between protein atoms by simple distance-dependent springs and despite their simplicity are very successful in describing the global mobility of proteins around a stable state. The predicted soft collective degrees of freedom from such approaches overlap often well with observed global changes in proteins [68, 69]. The soft collective normal modes can then be used as additional variables during docking by energy minimization [68, 70] or Monte Carlo approaches [45]. It can result in improved geometry and ranking of near-native docking solutions and can also lead to an enrichment of solutions close to the native complex structure (illustrated in Figure 4.4).

A significant number of protein–protein interactions involve disordered protein partners or at least the coupled folding and binding of segments of proteins. In such cases, it is possible to employ directly MD or MC simulations for docking. These techniques allow for full flexibility of proteins but are computationally demanding and are therefore more frequently applied in the docking refinement steps (see below) [19]. For prediction of interactions that involve partially disordered segments which become structured only upon binding it can be useful to employ protein–peptide docking methods. Those approaches often require only the sequence of the peptide segment and the protein partner structure as input [71–73] and putative binding geometries and conformations of the docked peptide segment are generated on the fly during the docking process.

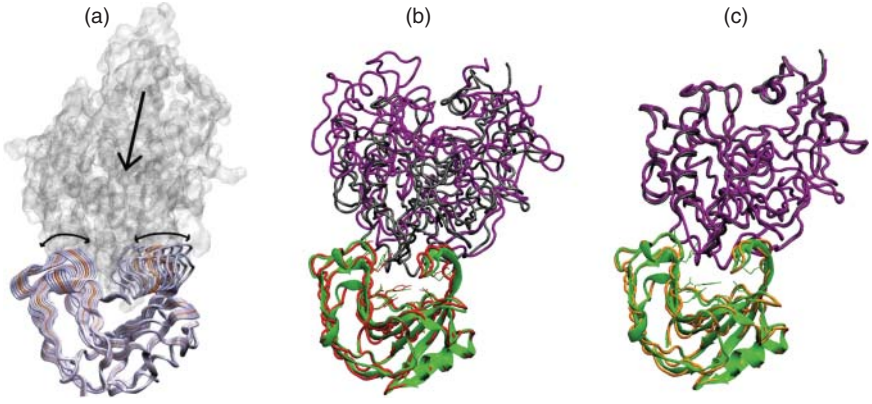


Figure 4.4 (a) Inclusion of global degrees of freedom during docking [68]. Elastic network models (ENM) allow identification of soft global collective degrees of freedom of a protein (e.g. domain opening/closing). It is illustrated by superposition of structures (backbone representation) deformed in the softest normal mode for the xylanase enzyme. The deformation motion allows entry and binding of a xylanase inhibitor protein (gray van der Waals spheres). (b) Best docking result of an inhibitor protein (magenta) to the enzyme using rigid docking compared to the native binding placement (gray backbone representation). The unbound enzyme receptor is indicated in red and the bound structure in green. (c) Inclusion of the soft collective modes during docking results in the best inhibitor placement (magenta) in close agreement with the native placement (gray). The slightly opened enzyme receptor structure in the energy minimized complex is shown in orange (experimental complex structure PDB id 1T6G).

4.4 Integration of Bioinformatics and Experimental Data for Protein–Protein Docking

The prediction accuracy of protein–protein docking is limited due to several approximations (e.g. semi-rigid partner structures, limited accuracy of scoring schemes). Often, a docking search provides several putative solutions with similar scoring. The inclusion of experimental or bioinformatics data on putative binding regions or residues involved in binding can greatly improve prediction accuracy. One option is to integrate low- or high-resolution experimental data directly during the search to steer the docking engine toward binding geometries compatible with the additional information [18]. These approaches are frequently termed "integrative modeling" ([4, 73–75], see also Chapter 7). Alternatively, one can also use the experimental data to filter out false predictions at post-processing stages of an unbiased docking search [76]. In this way, mutagenesis data on residues identified as part of the binding interface, evolutionary conserved surface residues, or coevolutionary data on interface residues can be included in the prediction [16].

In addition, experimental data can be used to guide docking of proteins. The chemical cross-linking of amino acid residues with compounds that contain two chemical reactive groups at a certain distance allows one to identify neighboring residues on protein surfaces or between surfaces of proteins forming a complex [77]. Typically, proteolytic digestion combined with mass spectrometry can be used

to identify cross-link sites on protein partners. A large variety of compounds has been developed in recent years that can also be applied *in vivo* to cross-link protein–protein complexes. Identified cross-links can be included as upper bound distance restraints during docking or for screening a set of solutions. It has been used to assist in modeling large complexes [17, 77] or to improve the rank of the near-native predictions and refine the structure of symmetric complexes using cross-link distances as cutoff [78]. Several docking programs can directly include cross-linking data as spatial restraints to guide and evaluate docking results (e.g. ATTRACT [79], HADDOCK [57], or Rosetta docking programs [46]). Biophysical techniques such as small-angle X-ray scattering (SAXS) in solution can also be used for the rapid characterization of structural and dynamic properties of protein–protein complexes at low resolutions [4]. SAXS data calculated as a form of convolution can be used directly during sampling to guide the search stage [80]. However, most approaches use SAXS data as a filter to refine and rank the final predictions [81–83].

The Cryo-EM method achieves often atomic detail resolution for large biomolecular complexes. In case of sufficient resolution, MD-based methods are extremely powerful and allow nowadays an almost automatic generation of Cryo-EM-based structures. However, frequently only a low-resolution electron-density envelope (15–20 Å resolution) can be obtained which, nevertheless, can be useful for evaluating the shape of the macromolecular complex. In some docking programs, the fitting of substructures into low-resolution Cryo-EM density has been integrated as restraint during docking. For example, the HADDOCK program or the ATTRACT approach can include Cryo-EM data as restraints in addition to other sources of experimental and bioinformatics data for generating putative models of complexes [84]. Using ATTRACT, it was demonstrated that inclusion of low-resolution Cryo-EM data (in the resolution range of 15–20 Å) is highly efficient to guide docking to near-native geometries for the majority of complexes in a large benchmark set [85].

4.5 Template-Based Protein–Protein Docking

Most specific and high-affinity protein–protein interactions in living cells are mediated by reoccurring domains. In case of protein structure prediction, evidence exists that the number of protein domain fold types is limited and even already largely covered by the present structures in the protein database. It has also been recognized that comparative modeling methods used for structure prediction of single proteins could also be used to predict the structure of entire protein–protein complexes [59, 86]. There is evidence that it is also possible to represent all protein–protein binding interfaces by a limited set of interface types mostly already available in the protein database [86, 87]. On the other hand, a large number of antibody–antigen interfaces and possible antigen–antibody complexes indicates that even for one protein type an enormous variety of interface surfaces can in principle be generated.

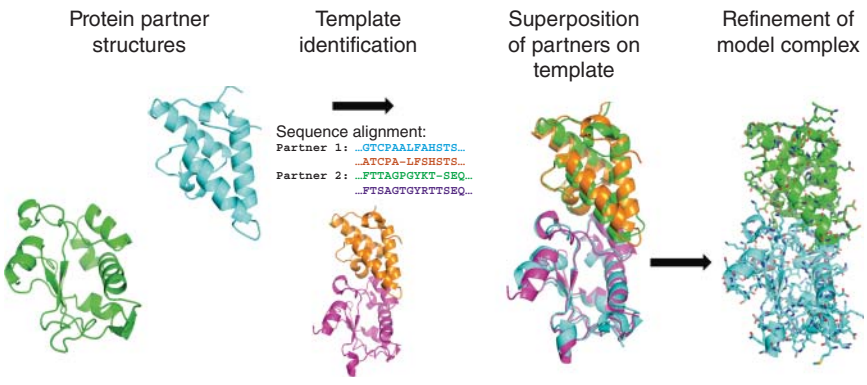


Figure 4.5 Flowchart of template-based docking of protein partners. For a given pair of interacting protein partners, an appropriate template complex is identified in most cases by a sequence similarity search. In the next step, an appropriate superposition on the template using either the full structures or just the interfaces produces an initial complex that can be further refined by energy minimization or related approaches.

A prerequisite for template-based docking is the availability of homologous template structures for the complex or at least part of it (Figure 4.5). Natural complexes that share greater than 35% sequence identity are likely to also share similar structures and interaction modes. Several methods have been published recently to extend available homology modeling methods to allow modeling of protein–protein complexes [59, 88–90] (Table 4.3, see also Chapter 6). To apply the comparative modeling approach to more putative protein–protein interactions, it is desirable to extend it to even remotely related protein complexes. Chen & Skolnick developed an approach of combining a multimeric threading method to detect and align remotely related multimeric proteins and structural refinement (M-Tasser) [94]. The structural refinement allows for conformational change of backbone and side chains similar to methods used to refine homology modeled single-chain proteins. On a large test set of >200 dimers, the method was able to identify correct templates in a large fraction of cases and showed improvement of the final predicted structure compared to the starting template structure by on average 1.5 Å. With a growing number of experimentally determined protein–protein complex structures, template-based docking may have a similar impact on structural biology as comparative modeling of monomeric proteins. With an appropriate template available, it has been demonstrated that template-based docking outperforms nontemplate-based approaches [59, 95].

Recently, great progress has been achieved to accurately model the structure of proteins using artificial intelligence deep neural network methods with the AlphaFold2 [96] and RoseTTAFold [97] approaches. In principle, such approaches can also be counted as template-based methods; however, not only the structural data of a single known structure but of the entire protein database is employed to predict protein structures. It has been demonstrated that AlphaFold2 and RoseTTAFold can also be used very successfully to predict the structure of protein–protein complexes [98].

Table 4.3 Examples of template-based protein–protein docking methods.

Program	Method	Website
PRISM [90]	PP-docking based on interface similarity to template	http://cosbi.ku.edu.tr/prism/index.php
IWRAP [88]	PP-docking by interface threading approach	http://iwrap.csail.mit.edu
PrePPI [89]	PP-docking based on interface similarity	https://bhapp.c2b2.columbia.edu/PrePPI
HDOCK [91]	Docking based on sequence homology	http://hdock.phys.hust.edu.cn
KBDOCK [87]	Template-based docking using contact data	http://kbdock.loria.fr
EVcomplex [92]	Does not use a template structure but extracts PPI information from coevolution data based on many sequences, which can be used to assist docking	https://evcomplex.hms.harvard.edu
InterEvDocks [93]	Complex prediction using evolutionary and coevolution data of interface	http://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2

4.6 Flexible Refinement of Docked Complexes

Frequently, complex structures obtained from an initial rigid or semi-flexible protein–protein docking or comparative template-based modeling programs are of limited accuracy. To generate an accurate realistic structural model, further structural refinement is required. The limitations of rigid docking strategies in combination with a rescoring step have been investigated on large sets of test cases [99]. Good performance was found for proteins that undergo minor conformational changes upon complex formation (<1 Å RMSD between unbound and bound structures) but unsatisfactory results for cases with significant binding induced conformational changes or applications that involved homology modeled proteins.

Hence, most protein–protein docking protocols consist of a preliminary exhaustive systematic docking search followed by a refinement step for a subset of putative complexes from the initial search (illustrated in Figure 4.6). Sometimes, several refinement steps are involved [19]. Often hundreds or thousands of initial solutions are refined and therefore computational efficiency of the refinement step is critical. Most FFT-based approaches such as ClusPro [30] or ZDOCK employ energy minimization and short MD simulations to relax structures and remove sterical overlap [100]. Refinement steps of the HADDOCK program employ several energy minimizations and dynamics steps in dihedral variables followed by a MD simulation in Cartesian coordinates with the option to include explicit solvent [65]. The FireDock approach [53] uses a combination of rigid body moves and side-chain

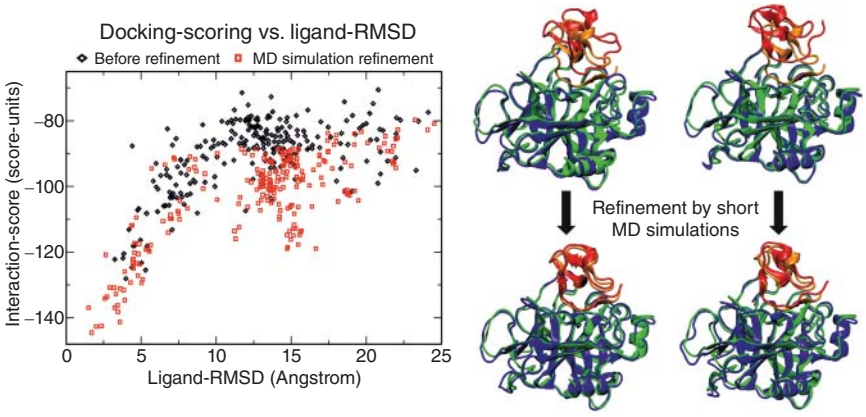


Figure 4.6 Docked complex structures can often be refined by short Molecular Dynamics (MD) simulations [70]. The docking interaction score vs. deviation of the predicted complex from the native experimental complex structure (in terms of the Ligand-RMSD) frequently gives a "funnel-like" plot (left panel). Successful refinement can improve both the score of the complexes and reduces the deviation from the native complex (compare red and black dots). The right panel indicates two examples of successful refinement for the trypsin/trypsin-inhibitor (PDB ID 1PPE) complex with the predicted structures (receptor/ligand protein in blue/red) superimposed on the native complex (receptor/ligand protein in green/orange). The refinement can result in a relative placement of the ligand protein in closer agreement with the native structure than the initially docked complexes.

optimization to improve the surface complementarity of docking solutions obtained by PatchDock [49]. Predicted protein–protein complexes can also be refined with the Rosetta molecular modeling suite [56]. It can be used for flexible refinement of just the side chains at the interface but also in combination with "backrub" motions to modify the backbone geometry [101]. Standard energy minimization or short MD simulation often does not move the partner structures significantly from the starting geometry and results only in local relaxation of steric stress. Especially, when the unbound protein structures differ from the conformation in the bound complex, the initial systematic search does not result in any prediction close to the native complex. Docking refinement strategies like the iATTRACT approach [102] combine energy minimization in the global translational and rotational variables with full atomic flexibility of only the predicted interface region. In this way, small side-chain movements can trigger larger scale whole-body movements of partner proteins.

Progress in computer hardware offers the option to use long MD simulations of putative docked complexes under realistic conditions including full flexibility of partners and explicit solvent molecules for refinement. To improve sampling, advanced methods, such as meta-dynamics or replica-exchange simulations (REMD), with an added biasing potential to the partners in the replicas (H-REMD or BP-REMD approach) can be used to avoid trapping nonspecific geometries and sample a broad range of conformations [103]. This technique has shown promising results for rapidly refining docked solutions including full flexibility of binding partners.

4.7 Scoring of Docked Complexes

Whereas simple criteria such as degree of steric overlap and surface complementarity are used in the initial stages of protein–protein docking, final re-scoring also of refined solutions involves more sophisticated evaluations [25, 104, 105]. It can be based on a molecular–mechanics type force field function that includes van der Waals (VDW) interactions, electrostatic Coulomb interactions, and solvation contributions. For example, the ATTRACT approach [79] (after atomistic refinement), the HADDOCK [57], and Rosetta [46] protein–protein docking programs use such a method, typically, assigning optimal weights to the different force field terms. These weights are optimized on a benchmark set of known complexes and a large set of incorrect decoy complexes.

Knowledge-based or statistical potentials for scoring docking solutions are also available and frequently used to evaluate docking solutions [104–108]. The central idea of a statistical potential is to look at the observed probability of finding residues/atoms (or pairs of atoms/residues) at a protein–protein interface and relate this to the expected probability based on a random distribution of the residues on a protein surface (reviewed in [105]). The obtained probability ratio is often evaluated for different distances between residues or atoms at the interface and can be related to an effective free energy function (by taking the logarithm of the probability ratio). Although the concept can be used to design effective multi-body potentials, in most cases only pairwise contacts or distances between residues are considered [104]. For optimizing such potentials, machine-learning techniques are increasingly being used [109–111]. However, a major limitation is still given by the relatively small number of known complexes and protein–protein interfaces relative to a large number of possible amino acid residue or chemical group combinations at interfaces.

Instead of scoring single predicted complex structures, several techniques can be used to account for the ensembles of solutions. In the molecular mechanics Poisson-Boltzmann/Generalized Born surface area (MMPBSA/MMGBSA) approaches an ensemble of docked conformations in the vicinity of the starting complex is generated using MD-simulations which is analyzed using a molecular mechanics force field (similar to the single structure scoring) combined with the Poisson–Boltzmann or Generalized Born approach to implicitly account for solvation effects. This technique is more demanding than scoring single structures but was used quite successfully for scoring protein–protein complexes [105, 112, 113].

All the above-considered scoring methods just count for the interaction between partners in the complex structure. However, a binding process is not only driven by the interaction between partners but also by other energetic and entropic contributions that together determine the binding free energy. In addition to interactions between the proteins, the binding free energy is influenced by changes in solvation, by the energy of deforming the unbound structures into the bound conformations, by the entropic cost of reducing the rotational and translational freedom of one partner relative to the other and by the change in conformational entropy of the partners (usually a restriction of conformational mobility) [114].

In principle, all these effects can be calculated in free energy simulations of protein–protein binding using appropriate advanced sampling MD simulations [105, 114–116]. For the calculation of the absolute binding free energy, one typically restrains the conformations of the partners to stay close to the starting structure in the bound (or predicted) complex and includes restraints to keep the relative orientation of the partners. The stepwise dissociation can be achieved by addition of an appropriate biasing potential along a distance coordinate between both proteins and the associated free energy (work) along the dissociation path can be extracted. Finally, simulations are performed at the bound and dissociated states to calculate the free energy of releasing the restraints. Although computationally very demanding, such methods can be used to evaluate single complexes [116] or a set of complexes using a coarse-grained model [117]. The methodology was recently also evaluated systematically on 20 test systems and including 50 decoy complexes for each test case in combination with an implicit solvent description [118] and later inclusion of explicit solvent [119]. The performance was found to be better than scoring based just on interaction energies, but further developments are necessary to improve accuracy and convergence. Several studies based on multiple MD simulations or different advanced sampling schemes including REMD approaches or nonequilibrium MD methods indicate promising results for identifying, scoring, and refining putative protein–protein complexes including explicit solvent and full flexibility of protein partners [103, 120, 121].

4.8 Conclusions and Future Developments

In recent years, protein–protein docking methods have evolved to become standard tools for the rapid modeling and structure generation of putative complexes. In the form of available docking programs or as docking web servers, these tools are frequently used also by non-experts working in the field of protein–protein interactions. Complex structures can be generated by *ab initio* docking but in many cases, especially of natural protein–protein interactions, template-based docking is the method of choice. In particular, the application of binary protein–protein interactions is highly evolved with many available approaches and many successful applications. A further significant extension of modeling protein–protein complexes has been achieved by recent deep learning-based structure prediction methods such as AlphaFold2 [96] or RoseTTAFold [97]. These techniques offer great promise to systematically provide structural models of many important proteins and protein–protein interactions in various organisms [98]. For many protein–protein docking cases the inclusion of experimental information or data from bioinformatics restricts the search and can further improve the prediction results. Also, the community-wide docking challenge CAPRI is very helpful in regularly monitoring the progress of the field and identifying and promising new developments.

Nevertheless, there are challenges in the field that are still difficult to tackle. For example, experimental methods are available to map the protein–protein interaction network between a virus and a host cell. Structure prediction for all

these interactions using computational docking is difficult but it may become feasible using deep neural network approaches. However, the lack of sufficiently accurate structures of protein partners, the involvement of large conformational changes during binding (e.g. coupled folding and binding of protein segments), and the simultaneous involvement of several proteins in complexes formed by multiple proteins are factors that complicate the prediction. Another important aspect is possible weak and very transient interactions that are frequently of significant functional importance in cellular processes. In contrast to highly stable interactions, such weakly bound complexes are not well represented in the set of known protein–protein complexes. Hence, current docking and template or deep neural network-based methods, as well as scoring schemes, may not be well suited for such types of interactions.

The integration of experimental data (e.g. from chemical cross-linking or Cryo-EM-tomography) but also bioinformatics data offers great promise to generate structural models of large and transient assemblies in the cell. If the structures of individual stable parts of molecular assemblies are known to atomic resolution, the inclusion even of low-resolution data can provide sufficient constraints to accurately model many multi-protein complexes even in a crowed cell-type environment.

In recent years, the design of entirely new protein–protein interactions to generate synthetic protein–protein complexes with desired function has become a major research focus. Also in this field protein–protein docking methods can contribute. They can be used to check if the desired interaction is stable and the only possible geometry for a set of designed or selected proteins.

Acknowledgments

The author likes to thank S. Fiorucci, S. de Vries, C. Schindler, and T. Siebenmorgen for helpful discussions and the Deutsche Forschungsgemeinschaft for continuous support.

References

1 Wilson, C.J., Bommarius, A.S., Champion, J.A. et al. (2018). Biomolecular assemblies: moving from observation to predictive design. *Chem. Rev.* 118 (24): 11519–11574.
2 Marsh, J.A. and Teichmann, S.A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.* 84 (1): 551–575.
3 Johnson, G.T., Autin, L., Al-Alusi, M. et al. (2015). cellPACK: a virtual mesoscope to model and visualize structural systems biology. *Nat. Methods* 12 (1): 85–91.
4 Soni, N. and Madhusudhan, M.S. (2017). Computational modeling of protein assemblies. *Curr. Opin. Struct. Biol.* 44: 179–189.

5 Huang, P.-S., Boyken, S.E., and Baker, D. (2016). The coming of age of de novo protein design. *Nature* 537 (7620): 320–327.

6 King, N.P., Bale, J.B., Sheffler, W. et al. (2014). Accurate design of co-assembling multi-component protein nanomaterials. *Nature* 510 (7503): 103–108.

7 Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nat. Struct. Mol. Biol.* 10 (12): 980–980.

8 Mosca, R., Céol, A., and Aloy, P. (2013). Interactome 3D: adding structural details to protein networks. *Nat. Methods* 10 (1): 47–53.

9 Shoemaker, B.A. and Panchenko, A.R. (2007). Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.* 3 (4): e43.

10 Güntert, P. (2008). Automated structure determination from NMR spectra. *Eur. Biophys. J.* 38 (2): 129.

11 Clore, G.M. and Gronenborn, A.M. (1998). Determining the structures of large proteins and protein complexes by NMR. *Trends Biotechnol.* 16 (1): 22–34.

12 Bai, X., McMullan, G., and Scheres, S.H.W. (2015). How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci* 40 (1): 49–57.

13 Elmlund, D., Le, S.N., and Elmlund, H. (2017). High-resolution cryo-EM: the nuts and bolts. *Curr. Opin. Struct. Biol.* 46: 1–6.

14 Rajagopala, S.V., Sikorski, P., Kumar, A. et al. (2014). The binary protein–protein interaction landscape of *Escherichia coli*. *Nat. Biotechnol.* 32 (3): 285–290.

15 Babu, M., Bundalovic-Torma, C., Calmettes, C. et al. (2018). Global landscape of cell envelope protein complexes in *Escherichia coli*. *Nat. Biotechnol.* 36 (1): 103–112.

16 Cong, Q., Anishchenko, I., Ovchinnikov, S., and Baker, D. (2019). Protein interaction networks revealed by proteome coevolution. *Science* 365 (6449): 185–189.

17 Chavez, J.D., Mohr, J.P., Mathay, M. et al. (2019). Systems structural biology measurements by in vivo cross-linking with mass spectrometry. *Nat. Protoc.* 14 (8): 2318–2343.

18 Bonvin, A.M. (2006). Flexible protein–protein docking. *Curr. Opin. Struct. Biol.* 16 (2): 194–200.

19 Zacharias, M. (2010). Accounting for conformational changes during protein–protein docking. *Curr. Opin. Struct. Biol.* 20 (2): 180–186.

20 Im, W., Liang, J., Olson, A. et al. (2016). Challenges in structural approaches to cell modeling. *J. Mol. Biol.* 428 (15): 2943–2964.

21 Zhang, Q., Feng, T., Xu, L. et al. (2016). Recent advances in protein–protein docking. *Curr. Drug Targets* 17 (14): 1586–1594.

22 Harmalkar, A. and Gray, J.J. (2021). Advances to tackle backbone flexibility in protein docking. *Curr. Opin. Struct. Biol.* 67: 178–186.

23 Lensink, M.F., Méndez, R., and Wodak, S.J. (2007). Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins Struct. Funct. Bioinf.* 69 (4): 704–718.

24 Lensink, M.F. and Wodak, S.J. (2010). Docking and scoring protein interactions: CAPRI 2009. *Proteins Struct. Funct. Bioinf.* 78 (15): 3073–3084.

25 Lensink, M.F., Velankar, S., and Wodak, S.J. (2017). Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins Struct. Funct. Bioinf.* 85 (3): 359–377.

26 Koshland, D.E. (1994). Das Schüssel-Schloß-Prinzip und die Induced-fit-Theorie. *Angew. Chem.* 106 (23–24): 2468–2472.

27 Katchalski-Katzir, E., Shariv, I., Eisenstein, M. et al. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U.S.A.* 89 (6): 2195–2199.

28 Mashiach, E., Schneidman-Duhovny, D., Peri, A. et al. (2010). An integrated suite of fast docking algorithms. *Proteins* 78 (15): 3197–3204.

29 Carter, P., Lesk, V.I., Islam, S.A., and Sternberg, M.J.E. (2005). Protein–protein docking using 3D-Dock in rounds 3, 4, and 5 of CAPRI. *Proteins Struct. Funct. Bioinf.* 60 (2): 281–288.

30 Kozakov, D., Hall, D.R., Xia, B. et al. (2017). The ClusPro web server for protein–protein docking. *Nat. Protoc.* 12 (2): 255–278.

31 Mandell, J.G., Roberts, V.A., Pique, M.E. et al. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein Eng. Des. Sel.* 14 (2): 105–113.

32 Tovchigrechko, A. and Vakser, I.A. (2006). GRAMM-X public web server for protein–protein docking. *Nucleic Acids Res.* 34 (suppl_2): W310–W314.

33 Macindoe, G., Mavridis, L., Venkatraman, V. et al. (2010). HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Res.* 38 (suppl_2): W445–W449.

34 Heifetz, A., Katchalski-Katzir, E., and Eisenstein, M. (2002). Electrostatics in protein–protein docking. *Protein Sci.* 11 (3): 571–587.

35 Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. *Proteins Struct. Funct. Bioinf.* 52 (1): 80–87.

36 Ohue, M., Shimoda, T., Suzuki, S. et al. (2014). MEGADOCK 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers. *Bioinformatics* 30 (22): 3281–3283.

37 Chowdhury, R., Rasheed, M., Keidel, D. et al. (2013). Protein–protein docking with F2Dock 2.0 and GB-Rerank. *PLoS One* 8 (3): e51307.

38 Jiménez-García, B., Pons, C., and Fernández-Recio, J. (2013). PyDockWEB: a web server for rigid-body protein–protein docking using electrostatics and desolvation scoring. *Bioinformatics* 29 (13): 1698–1699.

39 Mitra, P. and Pal, D. (2011). PRUNE and PROBE--two modular web services for protein–protein docking. *Nucleic Acids Res.* 39 (Web Server issue): W229–W234.

40 Zacharias, M. (2003). Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* 12 (6): 1271–1282.

41 de Vries, S.J., Schindler, C.E.M., Chauvot de Beauchêne, I., and Zacharias, M. (2015). A web interface for easy flexible protein–protein docking with ATTRACT. *Biophys. J.* 108 (3): 462–465.

42 Schneider, S., Saladin, A., Fiorucci, S. et al. (2012). ATTRACT and PTOOLS: open source programs for protein–protein docking. In: *Computational Drug Discovery and Design*, Methods in Molecular Biology (ed. R. Baron), 221–232. New York, NY: Springer New York https://doi.org/10.1007/978-1-61779-465-0_15.

43 Weng, G., Wang, E., Wang, Z. et al. (2019). HawkDock: a web server to predict and analyze the protein–protein complex based on computational docking and MM/GBSA. *Nucleic Acids Res.* 47 (W1): W322–W330.

44 Ramírez-Aportela, E., López-Blanco, J.R., and Chacón, P. (2016). FRODOCK 2.0: fast protein–protein docking server. *Bioinformatics* 32 (15): 2386–2388.

45 Moal, I.H. and Bates, P.A. (2010). SwarmDock and the use of normal modes in protein–protein docking. *Int. J. Mol. Sci.* 11 (10): 3623–3648.

46 Chaudhury, S., Berrondo, M., Weitzner, B.D. et al. (2011). Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One* 6 (8): e22477.

47 Deplazes, E., Davies, J., Bonvin, A.M.J.J. et al. (2016). Combination of ambiguous and unambiguous data in the restraint-driven docking of flexible peptides with HADDOCK: the binding of the spider toxin PcTx1 to the acid sensing ion channel (ASIC) 1a. *J. Chem. Inf. Model.* 56 (1): 127–138.

48 Fernández-Recio, J., Totrov, M., and Abagyan, R. (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins Struct. Funct. Bioinf.* 52 (1): 113–117.

49 Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H.J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 33 (suppl_2): W363–W367.

50 Ritchie, D.W. and Kemp, G.J.L. (2000). Protein docking using spherical polar Fourier correlations. *Proteins Struct. Funct. Bioinf.* 39 (2): 178–194.

51 Padhorny, D., Kazennov, A., Zerbe, B.S. et al. (2016). Protein–protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proc. Natl. Acad. Sci.* 113 (30): E4286–E4293.

52 Venkatraman, V., Sael, L., and Kihara, D. (2009). Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. *Cell Biochem. Biophys.* 54 (1): 23–32.

53 Mashiach, E., Schneidman-Duhovny, D., Andrusier, N. et al. (2008). FireDock: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Res.* 36 (suppl_2): W229–W232.

54 Zacharias, M. (2005). ATTRACT: protein–protein docking in CAPRI using a reduced protein model. *Proteins Struct. Funct. Bioinf.* 60 (2): 252–256.

55 Ruiz Echartea, M.E., Chauvot de Beauchêne, I., and Ritchie, D.W. (2019). EROS-DOCK: protein–protein docking using exhaustive branch-and-bound rotational search. *Bioinformatics* 35 (23): 5003–5010.

56 Kuroda, D. and Gray, J.J. (2016). Pushing the backbone in protein–protein docking. *Structure* 24 (10): 1821–1829.

57 Dominguez, C., Boelens, R., and Bonvin, A.M.J.J. (2003). HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* 125 (7): 1731–1737.

58 Andrusier, N., Mashiach, E., Nussinov, R., and Wolfson, H.J. (2008). Principles of flexible protein–protein docking. *Proteins Struct. Funct. Bioinf.* 73 (2): 271–289.

59 Chakravarty, D., McElfresh, G.W., Kundrotas, P.J., and Vakser, I.A. (2020). How to choose templates for modeling of protein complexes: insights from benchmarking template-based docking. *Proteins Struct. Funct. Bioinf.* 88 (8): 1070–1081.

60 Zhu, G., Liu, W., Bao, C. et al. (2018). Investigating energy-based pool structure selection in the structure ensemble modeling with experimental distance constraints: the example from a multidomain protein Pub1. *Proteins Struct. Funct. Bioinf.* 86 (5): 501–514.

61 Boehr, D.D., Nussinov, R., and Wright, P.E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* 5 (11): 789–796.

62 Bahar, I., Lezon, T.R., Bakan, A., and Shrivastava, I.H. (2010). Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem. Rev.* 110 (3): 1463–1497.

63 Ritchie, D.W. (2008). Recent progress and future directions in protein–protein docking. *Curr. Protein Pept. Sci.* 9 (1): 1–15.

64 Marze, N.A., Roy Burman, S.S., Sheffler, W., and Gray, J.J. (2018). Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics* 34 (20): 3461–3469.

65 de Vries, S.J., van ADJ, D., Krzeminski, M. et al. (2007). HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins Struct. Funct. Bioinf.* 69 (4): 726–733.

66 Bahar, I., Lezon, T.R., Yang, L.-W., and Eyal, E. (2010). Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.* 39 (1): 23–42.

67 Krieger, J.M., Doruker, P., Scott, A.L. et al. (2020). Towards gaining sight of multiscale events: utilizing network models and normal modes in hybrid methods. *Curr. Opin. Struct. Biol.* 64: 34–41.

68 May, A. and Zacharias, M. (2008). Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein–protein docking. *Proteins Struct. Funct. Bioinf.* 70 (3): 794–809.

69 May, A. and Zacharias, M. (2005). Accounting for global protein deformability during protein–protein and protein–ligand docking. *Biochim. Biophys. Acta, Proteins Proteomics* 1754 (1): 225–231.

70 Glashagen, G., de Vries, S., Uciechowska-Kaczmarzyk, U. et al. (2020). Coarse-grained and atomic resolution biomolecular docking with the ATTRACT approach. *Proteins* 88 (8): 1018–1028.

71 Ciemny, M., Kurcinski, M., Kamel, K. et al. (2018). Protein–peptide docking: opportunities and challenges. *Drug Discovery Today* 23 (8): 1530–1537.

72 Schindler, C.E.M., de Vries, S.J., and Zacharias, M. (2015). Fully blind peptide-protein docking with pepATTRACT. *Structure* 23 (8): 1507–1515.

73 Alam, N., Goldstein, O., Xia, B. et al. (2017). High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLoS Comput. Biol.* 13 (12): e1005905.

74 Lasker, K., Phillips, J.L., Russel, D. et al. (2010). Integrative structure modeling of macromolecular assemblies from proteomics data. *Mol. Cell. Proteomics* 9 (8): 1689–1702.

75 Schmidt, C., Macpherson, J.A., Lau, A.M. et al. (2017). Surface accessibility and dynamics of macromolecular assemblies probed by covalent labeling mass spectrometry and integrative modeling. *Anal. Chem.* 89 (3): 1459–1468.

76 Xia, B., Vajda, S., and Kozakov, D. (2016). Accounting for pairwise distance restraints in FFT-based protein–protein docking. *Bioinformatics* 32 (21): 3342–3344.

77 Chavez, J.D. and Bruce, J.E. (2019). Chemical cross-linking with mass spectrometry: a tool for systems structural biology. *Curr. Opin. Chem. Biol.* 48: 8–18.

78 Vreven, T., Schweppe, D.K., Chavez, J.D. et al. (2018). Integrating cross-linking experiments with Ab initio protein–protein docking. *J. Mol. Biol.* 430 (12): 1814–1828.

79 de Vries, S.J., Schindler, C.E.M., Chauvot de Beauchêne, I., and Zacharias, M. (2015). A web Interface for easy flexible protein–protein docking with ATTRACT. *Biophys. J.* 108 (3): 462–465.

80 Schneidman-Duhovny, D., Hammel, M., Tainer, J.A., and Sali, A. (2016). FoXS, FoXSDock and MultiFoXS: single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* 44 (W1): W424–W429.

81 Jiménez-García, B., Pons, C., Svergun, D.I. et al. (2015). pyDockSAXS: protein–protein complex structure by SAXS and computational docking. *Nucleic Acids Res.* 43 (W1): W356–W361.

82 Xia, B., Mamonov, A., Leysen, S. et al. (2015). Accounting for observed small angle X-ray scattering profile in the protein–protein docking server cluspro. *J. Comput. Chem.* 36 (20): 1568–1572.

83 Sønderby, P., Rinnan, Å., Madsen, J.J. et al. (2017). Small-angle X-ray scattering data in combination with RosettaDock improves the docking energy landscape. *J. Chem. Inf. Model.* 57 (10): 2463–2475.

84 van Zundert, G.C.P., Melquiond, A.S.J., and Bonvin, A.M.J.J. (2015). Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data. *Structure* 23 (5): 949–960.

85 de Vries, S.J., Chauvot de Beauchêne, I., Schindler, C.E.M., and Zacharias, M. (2016). Cryo-EM data are superior to contact and Interface information in integrative modeling. *Biophys. J.* 110 (4): 785–797.

86 Kundrotas, P.J., Zhu, Z., Janin, J., and Vakser, I.A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci.* 109 (24): 9438–9441.

87 Ghoorah, A.W., Devignes, M.-D., Smaïl-Tabbone, M., and Ritchie, D.W. (2014). KBDOCK 2013 : a spatial classification of 3D protein domain family interactions. *Nucleic Acids Res.* 42 (D1): D389–D395.

88 Hosur, R., Xu, J., Bienkowska, J., and Berger, B. (2011). iWRAP: an Interface threading approach with application to prediction of cancer-related protein–protein interactions. *J. Mol. Biol.* 405 (5): 1295–1310.

89 Zhang, Q.C., Petrey, D., Garzón, J.I. et al. (2013). PrePPI: a structure-informed database of protein–protein interactions. *Nucleic Acids Res.* 41 (D1): D828–D833.

90 Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011). Predicting protein–protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat. Protoc.* 6 (9): 1341–1354.

91 Yan, Y., Tao, H., He, J., and Huang, S.-Y. (2020). The HDOCK server for integrated protein–protein docking. *Nat. Protoc.* 15 (5): 1829–1852.

92 Hopf, T.A., Schärfe, C.P.I., Rodrigues, J.P.G.L.M. et al. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. Kuriyan, J., ed. *eLife* 3: e03430.

93 Quignot, C., Rey, J., Yu, J. et al. (2018). InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic Acids Res.* 46 (W1): W408–W416.

94 Chen, H. and Skolnick, J. (2008). M-TASSER: an algorithm for protein quaternary structure prediction. *Biophys. J.* 94 (3): 918–928.

95 Szilagyi, A. and Zhang, Y. (2014). Template-based structure modeling of protein–protein interactions. *Curr. Opin. Struct. Biol.* 24: 10–23.

96 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873): 583–589.

97 Baek, M., DiMaio, F., Anishchenko, I. et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373 (6557): 871–876.

98 Humphreys, I.R., Pei, J., Baek, M. et al. Computed structures of core eukaryotic protein complexes. *Science* 374 (6573): eabm 4805.

99 Pons, C., Grosdidier, S., Solernou, A. et al. (2010). Present and future challenges and limitations in protein–protein docking. *Proteins* 78 (1): 95–108.

100 Li, L., Chen, R., and Weng, Z. (2003). RDOCK: refinement of rigid-body protein docking predictions. *Proteins Struct. Funct. Bioinf.* 53 (3): 693–707.

101 Chaudhury, S. and Gray, J.J. (2008). Conformer selection and induced fit in flexible backbone protein–protein docking using computational and NMR ensembles. *J. Mol. Biol.* 381 (4): 1068–1087.

102 Schindler, C.E.M., de Vries, S.J., and Zacharias, M. (2015). iATTRACT: simultaneous global and local interface optimization for protein–protein docking refinement. *Proteins Struct. Funct. Bioinf.* 83 (2): 248–258.

103 Siebenmorgen, T., Engelhard, M., and Zacharias, M. (2020). Prediction of protein–protein complexes using replica exchange with repulsive scaling. *J. Comput. Chem.* 41 (15): 1436–1447.

104 Gromiha, M.M., Yugandhar, K., and Jemimah, S. (2017). Protein–protein interactions: scoring schemes and binding affinity. *Curr. Opin. Struct. Biol.* 44: 31–38.

105 Siebenmorgen, T. and Zacharias, M. (2020). Computational prediction of protein–protein binding affinities. *WIREs Comput. Mol. Sci.* 10 (3): e 1448.

106 Liu, S., Zhang, C., Zhou, H., and Zhou, Y. (2004). A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins Struct. Funct. Bioinf.* 56 (1): 93–101.

107 Chuang, G.-Y., Kozakov, D., Brenke, R. et al. (2008). DARS (decoys as the reference state) potentials for protein–protein docking. *Biophys. J.* 95 (9): 4217–4227.

108 Moreira, I.S., Martins, J.M., Coimbra, J.T.S. et al. (2015). A new scoring function for protein–protein docking that identifies native structures with unprecedented accuracy. *Phys. Chem. Chem. Phys.* 17 (4): 2378–2387.

109 Sasse, A., de Vries, S.J., Schindler, C.E.M. et al. (2017). Rapid design of knowledge-based scoring potentials for enrichment of near-native geometries in protein–protein docking. *PLoS One* 12 (1): e0170625.

110 Ballester, P.J. and Mitchell, J.B.O. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* 26 (9): 1169–1175.

111 Feliu, E., Aloy, P., and Oliva, B. (2011). On the analysis of protein–protein interactions via knowledge-based potentials for the prediction of protein–protein docking. *Protein Sci.* 20 (3): 529–541.

112 Chen, F., Liu, H., Sun, H. et al. (2016). Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein–protein binding free energies and re-rank binding poses generated by protein–protein docking. *Phys. Chem. Chem. Phys.* 18 (32): 22129–22139.

113 Spiliotopoulos, D., Kastritis, P.L., Melquiond, A.S.J. et al. (2016). dMM-PBSA: a new HADDOCK scoring function for protein-peptide docking. *Front. Mol. Biosci.* 3: 46.

114 Mobley, D.L. and Gilson, M.K. (2017). Predicting binding free energies: frontiers and benchmarks. *Annu. Rev. Biophys.* 46 (1): 531–558.

115 Woo, H.-J. and Roux, B. (2005). Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci.* 102 (19): 6825–6830.

116 Gumbart, J.C., Roux, B., and Chipot, C. (2013). Efficient determination of protein–protein standard binding free energies from first principles. *J. Chem. Theory Comput.* 9 (8): 3789–3798.

117 May, A., Pool, R., van Dijk, E. et al. (2014). Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins. *Bioinformatics* 30 (3): 326–334.

118 Siebenmorgen, T. and Zacharias, M. (2019). Evaluation of predicted protein–protein complexes by binding free energy simulations. *J. Chem. Theory Comput.* 15 (3): 2071–2086.

119 Siebenmorgen, T. and Zacharias, M. (2020). Efficient refinement and free energy scoring of predicted protein–protein complexes using replica exchange with repulsive scaling. *J. Chem. Inf. Model.* 60 (11): 5552–5562.

120 Perthold, J.W. and Oostenbrink, C. (2017). Simulation of reversible protein–protein binding and calculation of binding free energies using perturbed distance restraints. *J. Chem. Theory Comput.* 13 (11): 5697–5708.

121 Perthold, J.W. and Oostenbrink, C. (2019). GroScore: accurate scoring of protein–protein binding poses using explicit-solvent free-energy calculations. *J. Chem. Inf. Model.* 59 (12): 5074–5085.

5

Identification of Putative Protein Complexes in Protein–Protein Interaction Networks

Sudharshini Thangamurugan, Markus Hollander, and Volkhard Helms

Saarland University, Center for Bioinformatics, Saarland Informatics Campus, Postfach 15 11 50, 66041 Saarbrücken, Germany

5.1 Protein–Protein Interaction Networks

In biological systems, most cellular and molecular mechanisms involve the activity of proteins. Rarely, only a single protein regulates or executes a complete mechanism. Instead, proteins frequently bind to other biomolecules, often other proteins, to execute cellular functions. Protein–protein interactions (PPIs) are highly specific physical contacts between two or more proteins that are formed due to the conformational and physicochemical properties of the involved proteins. The molecular details of individual PPIs are discussed in more detail in Chapters 2 and 4 and will be omitted here.

While 60–70% of the makeup of biological cells consists of water, 40–55% of the remaining dry weight consists of proteins [1]. Hence, freely diffusing cytosolic proteins frequently collide with other cellular proteins and may occasionally remain bound to each other for a short time as a nonspecific assembly. Only a small portion of these contacts will involve two or more proteins that are actually meant to bind to each other. In this book, we focus on such specific interaction pairs. Based on their lifetime, specific PPIs can be classified as either transient or stable interactions. Transient (specific) interactions between proteins are short-lived interactions formed to perform functions, such as signal transduction, or that lead to further changes (e.g. sodium–potassium pump). Stable interactions between proteins are long-lasting and often serve the purpose of forming macromolecular machinery (e.g. hemoglobin or RNA polymerase).

For a single protein, all its physical interactions with other proteins can be represented as a mathematical graph where the vertices represent the proteins and the undirected edges connecting the vertices represent the physical interactions between the proteins. Such a protein-centered network provides an idea about the protein complexes in which the protein of interest may be involved, and about their biological function. For example, the enzyme aspartate semialdehyde dehydrogenase from *Arabidopsis thaliana* appears to be part of three different protein complexes that are

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

active either in an oxidation–reduction process, in methionine biosynthesis, or in lysine biosynthesis [2]. In contrast, protein–protein networks (PPINs) are global PPI graphs or networks that present an overview of all PPIs existing in an organism. These comprehensive networks are cataloged by several established databases, such as the Biological General Repository for Interaction Datasets (BioGRID), mentha, the Search Tool for Retrieval of Interacting Genes/Proteins (STRING), the Molecular INTeraction Database (MINT), the protein Interaction database (IntAct), and others. Figure 5.1 illustrates the connectivity of a small toy PPIN.

In graphs, the degree of a vertex is the number of edges connected to it, and in PPINs the degree thus measures the number of interactions involving the protein represented by the vertex. One way of examining general connectivity and topology of a PPIN is to compute its degree distribution, which describes the frequency of each vertex degree occurring in the given network. Degree distributions are often visualized in plots that display the vertex degrees on the x-axis and their respective frequency on the y-axis. Upon analyzing multiple PPIN of different species, it was discovered that the networks have a "scale-free" topology irrespective of the species [3]. In scale-free networks, the degree distribution follows a power-law with negative exponent γ, where the probability of a vertex degree k is given by $P(k) = k^{-\gamma}$. As a consequence, the highly connected proteins, called hubs, occur at a much higher frequency than in an exponentially decaying scenario, in which the probability is $P(k) = e^{-\gamma k}$. This scale-free nature implies that the average length of the shortest pathway between any two vertices increases much slower as a function of network size than expected.

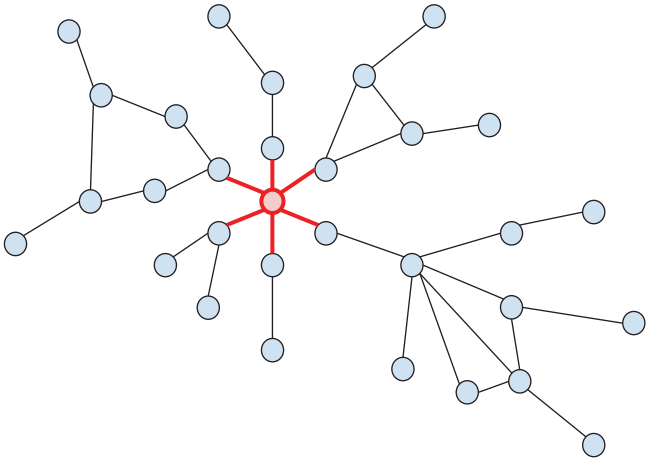


Figure 5.1 Schematic representation of a protein–protein interaction network. The circles are called the vertices of the network. Each one represents all copies of an individual protein type. The lines connecting the vertices are called edges and represent physical contact between two proteins. The degree of a vertex measures the number of edges connected to it. The vertex highlighted in red has six edges connected to it or six binding partners, and hence its degree is six. Note that this representation does not carry information on whether multiple interactions of one protein can occur simultaneously, potentially leading to the formation of a larger protein complex, or not.

5.2 Integration of Various PPI Resources in Public Data Repositories

There generally exist two kinds of protein interaction databases: primary databases and metadatabases. Primary databases directly compile the results of multiple experimental interaction assays. Well-known examples are the Biomolecular Interaction Network Database (BIND) [4], the IntAct molecular interaction database [5], the MINT [6], the Database of Interacting Proteins (DIP) [7], and the BioGRID [8]. In contrast, metadatabases typically integrate data from multiple primary databases. For example, the Integrated Interactions Database (IID) [9] compiles data from BIND, BioGRID, DIP, MINT, IntAct, and a few others, while the Agile Protein Interactomes DataServer [10] holds interactions from BioGRID, DIP, IntAct, MINT, and the Human Protein Reference Database (HPRD) [11]. For model organisms, the metadatabase mentha [12] integrates evidence-based interactions from BioGRID, DIP, IntAct, and MINT. Table 5.1 provides an overview of selected primary and meta-PPI databases.

An international collaboration of major contributors of PPI data, the International Molecular Exchange (IMEx) consortium [13], has established guidelines to maintain a consistent set of uniquely defined molecular identifiers and interactions. IntAct, MINT, DIP, and BIND are a few primary databases that are active members of the IMEx consortium. The metadatabase STRING [14] is notable since it offers interactions from both the IMEx consortium and BioGRID.

Table 5.1 Overview of selected primary and meta protein–protein interaction databases.

Database	Type	PPI Source	Species	Website
BIND	Primary	Evidence	Multiple	
BioGRID	Primary	Evidence	Multiple	http://thebiogrid.org
DIP	Primary	Evidence	Multiple	https://dip.doe-mbi.ucla.edu
IntAct	Primary and meta	Evidence	Multiple	https://www.ebi.ac.uk/intact/
MINT	Primary	Evidence	Multiple	http://mint.bio.uniroma2.it
APID	Meta: BioGRID, DIP, HPRD, IntAct, and MINT	Evidence	Multiple	http://apid.dep.usal.es
IID	Meta: IntAct, MINT, BioGRID, BIND, DIP, and others	Evidence and predicted	Multiple	http://iid20.ophid.utoronto.ca
mentha	Meta: BioGRID, DIP, IntAct, MINT and others	Evidence	Multiple	http://mentha.uniroma2.it
STRING	Meta: IMEx consortium and BioGRID	Evidence and predicted	Multiple	http://string-db.org

5.3 Protein–Protein Interaction Networks of Model Organisms

As mentioned before, most biological processes of an organism are mediated via protein interactions. Having an overview of the interactome of an organism contributes to deriving an understanding of which proteins and genes are associated with a certain process or disease. This then enables a better or deeper identification and mechanistic understanding of disease-related pathways and how they may be controlled. Therefore, one important pillar of computational systems biology is to study and compare the PPINs of one or more organisms to understand the mechanisms and regulations of biological systems.

5.3.1 PPIN of *Saccharomyces cerevisiae*

In order to characterize the protein–protein interaction network of the eukaryotic model organism *S. cerevisiae*, [15] used tandem-affinity purification coupled with mass spectrometry (TAP-MS) and Uetz et al. [16] and Fields and Song [17] used the yeast two-hybrid (Y2H) method. The results obtained by these methods yielded PPINs with 16 000–40 000 interactions involving most of the 6000 yeast proteins. As mentioned before, the network exhibits a power-law connectivity distribution, i.e. only few proteins are highly connected and form hubs whereas most of the proteins interact with only very few proteins. Initially, the coverage of PPIs was quite limited in these pioneering experiments (about 10% only), and concern was raised about whether these networks are really scale-free or simply appear scale-free as a consequence of the low coverage [18]. However, subsequent expansion of the coverage showed that they, in fact, have a scale-free topology [19].

An important question is which vertices are most important in such a PPIN. One way to define importance is to characterize whether a gene product is essential for the cell. If one knocks out an "essential" gene from the genome, this is, by definition, lethal to the cell. In contrast, knock-out cells of nonessential genes are still viable. Experimental studies by Winzeler et al. [20] and Giaever et al. [21], showed that around 1120 (19%) of all protein-coding genes of *S. cerevisiae* are "essential." Gene ontology analysis of these genes showed that about 74% of them are involved in metabolic processes and at least 14% in cell cycle regulation. These appear to be the two essential functions for cell survival [22]. Figure 5.2 displays the connectivity among yeast proteins based on the data in a current version of the Mentha database (http://www.mentha.uniroma2.it), whereby proteins are colored according to their essentiality (red) or non-essentiality (green).

Interestingly, when information about protein connectivity was combined with information on the essentiality of genes, it turned out that highly connected "hub proteins" are much more likely to be encoded by essential genes (ca. 60%) than low-degree proteins (ca. 15%) [3]. This makes intuitive sense. Knocking out a highly connected protein will likely cause a large perturbance to cellular processes. This behavior is illustrated in Figure 5.3 that displays the fractions of

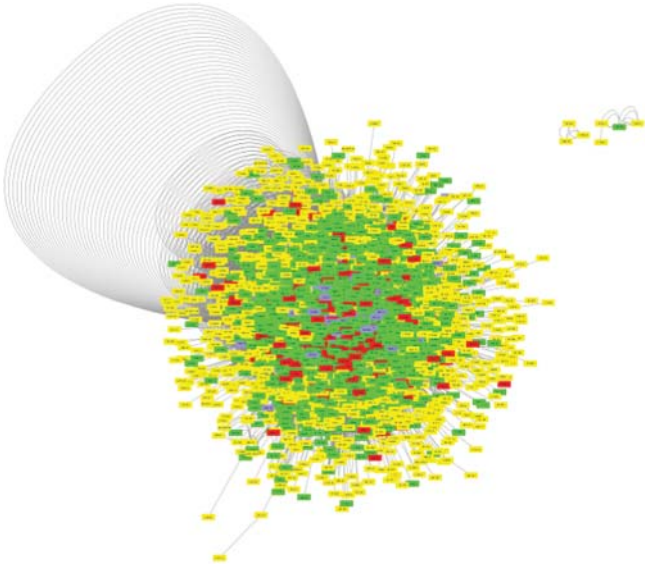


Figure 5.2 Complete interactome of *S. cerevisiae* derived by us from the mentha database and constructed using Cytoscape [23]. The interactome contains 6342 genes and 233 322 interactions. Red vertices represent essential genes (948 essential genes identified), green vertices represent nonessential genes (3583 nonessential genes identified), purple vertices represent conditional genes (270 conditional genes identified), and yellow vertices represent unknown essentiality (1541 genes have unknown essentiality).

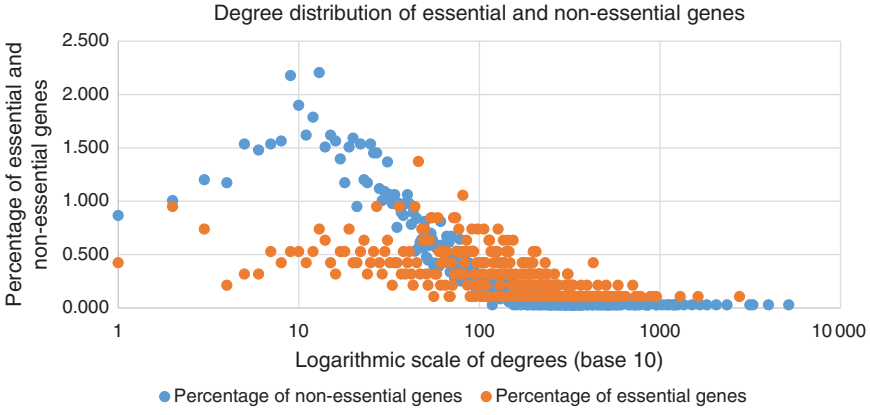


Figure 5.3 For the degree distributions for essential proteins and nonessential proteins (shown on the *x*-axis on a log scale), we color-coded the respective fractions of essential (orange) and nonessential (blue) genes for different degrees. This analysis recovers the previously observed enrichment of essential genes/proteins among the high-degree vertices of the PPIN.

essential vs. nonessential genes as a function of the connectivity of the proteins encoded by them in the PPIN of yeast.

Subsequent to the initial studies based on yeast two-hybrid screens mentioned above, different high-throughput methods were used to determine PPIs in *S. cerevisiae* such as the high-throughput mass-spectrometric protein complex identification (HMS-PCI) technique [24], correlated mRNA expression, and *in silico* predicted interactions. Overall, nowadays there exist confirmed evidence for about 80 000 interactions between proteins of *S. cerevisiae* [25]. When the early data were pooled together, out of the 80 000 interactions, approximately 2400 interactions were common for more than one high-throughput method [26]. This may be due to certain biases in the detection assays. Some methods such as Y2H were reported to have rather high false-positive rates (about 59%) or that they may not be able to detect certain kinds of interactions. For example, it was observed that the yeast two-hybrid method determined comparatively fewer proteins that regulate translation [26]. Hence, Han et al. [27] constructed a "filtered yeast interactome" (FYI) dataset by intersecting the data from different methods. This interactome consists of 2493 high-confidence interactions (observed commonly in at least two methods to rule out false-positive results), 1379 proteins with an average of 3.6 interactions per protein, and 1 large-connected component of 778 proteins. For every hub in the FYI, an average Pearson Correlation Coefficient (avPCC) was calculated correlating the expression levels of the hub and its binding partners under different conditions. The hubs with a degree greater than 5 showed a bimodal probability distribution for a few conditions. The hubs with degrees 5 or less showed a normal distribution centered at 0.1. It was understood that the bimodal distribution suggests two kinds of hub types, static hubs and dynamic hubs, based on their expression profiles, see Figure 5.4. In the 91 identified static hubs, the binding partners interact at the same

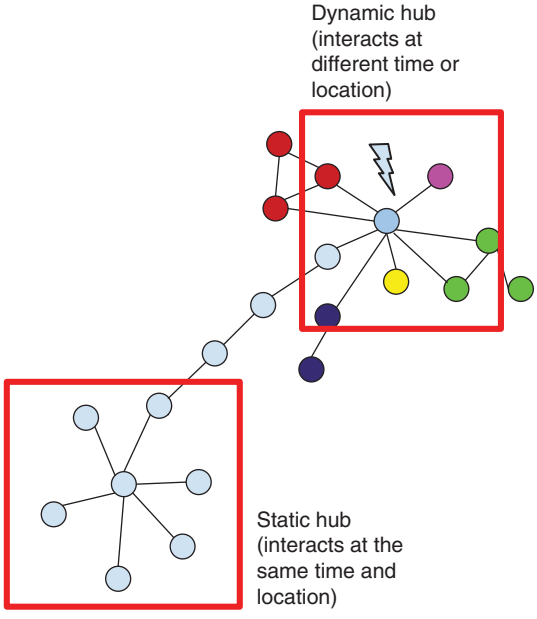


Figure 5.4 Schematic representation of static and dynamic hubs of the network. Proteins of the static hub interact with each other at the same time and location. Proteins of dynamic hubs interact with each other at different times or locations.

time and are involved in the main functional part of the complex. In the 108 identified dynamic hubs, the binding partners interact with each other at different times or in different locations and rather tend to connect separate modules of the PPIN.

To our knowledge, the distribution of essential proteins in either dynamic or static hubs has so far not yet been analyzed. Batada et al. [28], suggested that the existing data for PPIN of *S. cerevisiae* is too little to conclude and differentiate hubs. Out of the 5 conditions from the compendium, a few conditions utilized only 10 data points to differentiate the hubs which may not be enough data. Agarwal et al. [29] reported that if avPCC was calculated for the hubs in all conditions of the compendium, instead of using only five conditions, it yielded only 59 dynamic hubs using the same degree of threshold as 5. This shows that the differentiation of the hubs is mainly based on the expression profile and can vary with different experimental conditions. Hence, it is questionable whether avPCC is a good parameter to differentiate hubs. Also, based on functions, the hubs exhibited a spectrum of structural roles, which makes it difficult to differentiate them as static and dynamic hubs.

5.3.2 PPIN of Human

Based on data from the GTEx consortium, 20 532 potential protein-coding human genes have been annotated [30]. It is an enormous task to completely map the interaction network among all these human proteins. Initially, Stelzl et al. [31] constructed a partial human protein interaction network based on yeast two-hybrid screening of 4456 bait and 5632 prey proteins. This yielded 3186 mostly novel interactions of 1705 proteins. Recently, Agrawal et al. [32] collected network data from studies by Menche et al. [33] and Chatr-Aryamontri et al. [34] and fifteen databases which resulted in a large network of 342 353 interactions of 21 557 proteins.

Shin et al. [35] reviewed the identification of drug targets in the PPIN of humans. The development of any drug begins with the identification of a drug target, i.e a receptor protein having a druggable-binding pocket. As mentioned before, PPIs play an essential role in regulating biological pathways, including disease processes. It has been argued that considering the PPI network of humans is beneficial for determining novel drug targets [36]. In the past years, approximately 40 PPIs from the human interactome were identified as potential drug targets for drug development [37]. New computational structure-based approaches have been presented to determine inhibitors of PPIs that are termed Small Molecule Protein–Protein Interaction Inhibitors (SMPPIs). For example, the Small Ubiquitin-like Modifier (SUMO) protein forms a covalent interaction with proteins that possess a SUMO interaction motif (SIM) by the process called sumoylation. This process regulates general cellular processes, such as cell proliferation, chromosome winding, DNA replication, and DNA repair, and processes that cause neurodegenerative diseases and cancer. Considering the electrostatic similarity with the native-binding partner protein using software called Elekit [38] led to the discovery of an inhibitor that binds to the SUMO protein with low micromolar activity [39] and interferes with the SUMO-SIM interaction. Figure 5.5 illustrates schematically the idea behind the design of mimicking SMPPIs.

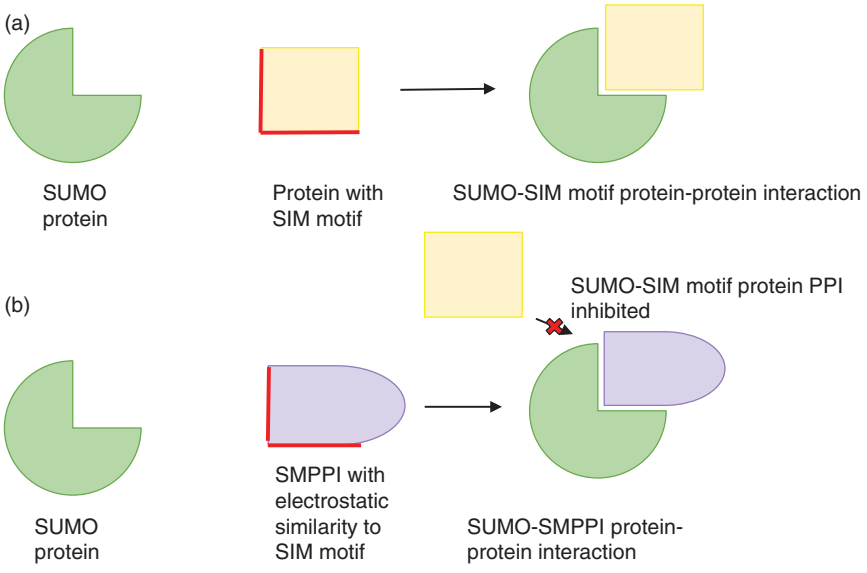


Figure 5.5 Overview of strategy behind blocking SUMO protein interactions. (a) Proteins with SIM motifs (highlighted in red) interact with SUMO proteins by forming covalent bonds through the process called sumoylation. (b) The SUMO proteins are targeted by SMPPIs (region similar to SIM motif highlighted in red) by binding to them and hence inhibiting the binding of proteins carrying SIM motifs.

5.4 Algorithms to Identify Protein Complexes in PPI Networks

Protein complexes often constitute macromolecular machines that play crucial roles in many cellular processes. For example, RNA polymerase is a protein complex formed from 10 individual protein units. It functions as the key enzyme in gene transcription and synthesizing a copy of mRNA from a DNA template. To the aim of better understanding cellular mechanisms, many mathematical algorithms were developed to identify putative protein complexes based on interactomics data. An intuitive idea was put out whereby putative protein complexes can be identified from a PPIN by detecting dense regions in a weighted PPIN containing many connections or ones with large weights [40]. Below, we review several algorithms that have been shown to be able to detect protein complexes in a PPIN.

5.4.1 Molecular Complex Detection (MCODE)

In 2003, Bader and Hogue [41] published one of the first graph-theoretic clustering algorithms called Molecular Complex Detection (MCODE) that estimates densely connected regions in the PPIN of an organism as putative molecular complexes. The algorithm proceeds in three steps: (i) vertex weighting, (ii) complex prediction, and

(iii) post-processing, where vertices are added or filtered out from the complex based on certain connectivity criteria.

5.4.1.1 Definitions

Let us assume that the PPIN used as basis for finding protein complexes is given as a graph G. In this graph, each protein is represented by a vertex v, while each edge e represents an interaction between two proteins. V is then the set of all proteins and E the set of all protein–protein interactions in the PPIN, giving the following definition of G:

$$G = (V, E)$$

In G, the total number of proteins in the network is given by $|V|$, and the total number of interactions by $|E|$. The maximum number of possible edges $|E|_{\max}$ occurs in fully connected networks, where each protein is connected to all other proteins. If the PPIN is given as a graph without self-edges, i.e. cases where proteins interact with themselves are not considered, each protein can be connected to at most $|V-1|$ other proteins. Since the direction of the interaction is generally not a factor, PPINs are typically undirected graphs, and the resulting number of edges thus has to be divided by 2, to not count the same edge twice:

$$|E|_{\max} = \frac{|V| \cdot (|V|-1)}{2}$$

If the PPIN includes self-edges, each protein can form $|V|$ edges, of which the self-edge is already unique. In other words, PPINs with self-edges have $|V|$ additional edges compared to PPINs without self-edges:

$$|E|_{\max} = \frac{|V| \cdot (|V|-1)}{2} + |V| = \frac{|V| \cdot (|V|+1)}{2}$$

The overall density d_G of a graph G is commonly defined as the fraction of the number of edges $|E|$ over the maximum number of edges: $d_G = \frac{|E|}{|E|_{\max}}$. However, to identify putative protein complexes, methods like MCODE consider local, rather than network-wide interaction density. A subgraph g of G consists of a subset of vertices V_g and the subset of edges E_g that connect the vertices in V_g. The local density d_g of subgraph g is defined analogously to the overall density d_G:

$$d_g = \frac{|E_g|}{|E_g|_{\max}}$$

An example of such a subgraph is the neighborhood N_v of vertex v, which in addition to v itself contains all vertices directly connected to it. The density of N_v measures the connectivity among the direct neighbors of v.

A k-core is another type of subgraph in which all included vertices have a degree of at least k, see Figure 5.6. The most densely connected subgraph is the k-core with the highest possible k, called $k_{\max}$. A $k_{\max}$-core can also be constructed from the neighborhood of a vertex v, see Figure 5.7, and its density is referred to as the core-clustering coefficient d_v.

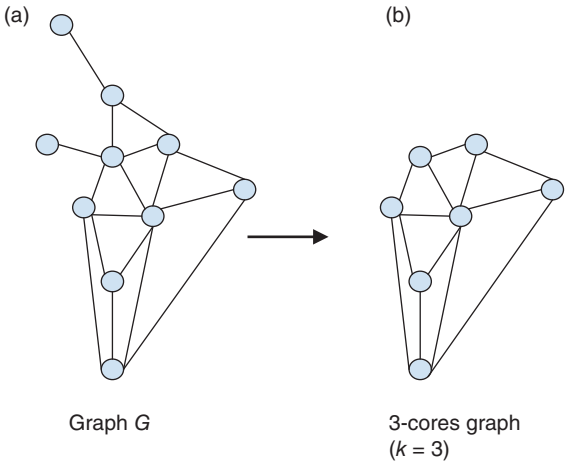


Figure 5.6 (a) Represents a graph *G*. When the *k*-value is set to 3, then (b) represents the subgraph *g* or the 3-core graph, in which all vertices of *g* have at least a degree of 3.

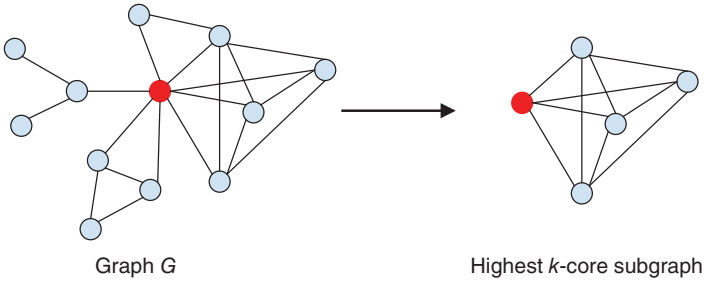


Figure 5.7 For the highlighted red vertex v in graph G (left), the k-core with maximal degree $k_{max} = 4$ is identified in the neighborhood N_v of v (right). The core-clustering coefficient or density of v is calculated as $d_v = (2 \cdot 10)/(5 \cdot 4) = 1$, showing that it is a fully connected subgraph. The weight of v is $w_v = k_{max} \cdot d_v = 4$.

5.4.1.2 Algorithm

Step 1 – Vertex weighting: Given the input PPIN as graph G, all vertices are weighted based on their local density. For each vertex v of G, its neighborhood N_v of directly connected vertices is constructed. In that neighborhood, the k-core with the highest k, i.e. the k_{max}-core, is identified. Figure 5.7 gives an example in which vertex v has degree 8 but none of its neighbors has a degree higher than 4. Thus $k_{max} = 4$ and the identified k_{max}-core consists of v and the 4 other vertices with degree 4.

Subsequently, the core-clustering coefficient d_v of v is calculated as the density of the identified k_{max}-core. The core-clustering coefficient focuses entirely on densely connected neighbors, which are assumed to be more indicative of putative protein complexes, and is thus not negatively affected by the presence of less connected neighbors, unlike the clustering coefficient that considers all original neighbors of v. Finally, the weight w_v of v is computed from k_{max} and the core-clustering coefficient, further emphasizing local network density:

$$w_v = k_{max} \cdot d_v$$

Pseudocode

Input: graph $G = (V, E)$

For each vertex v in V, calculate:

N_v: set of direct neighbors
K: $k_{\max}$-core of N_v with vertices V_K and edges E_K
$k_{\max}$: highest k-value of K
d_v: $\frac{|E_K|}{|E_K|_{\max}}$, density of K
w_v: $k_{\max} \cdot d_v$, weight of v

Step 2 – Complex prediction: In the second step, the graph with weight-annotated vertices is considered as input. The algorithm first selects the vertex with the highest weight as the start vertex for constructing the first complex. Based on the vertex weight percentage (VWP) parameter, the inclusion threshold for this complex is set as the specified fraction of the start vertex weight. The algorithm traverses through the graph and adds the neighbors of the start vertex to the current complex if their individual weights exceed the inclusion threshold. In subsequent recursions, the neighbors of newly added complex members are checked for the threshold as well and added to the complex if the condition is fulfilled. Since proteins cannot be assigned to more than one complex in this step, a vertex is not checked more than once. Finally, this process halts when no more vertices can be added to the current complex which is now considered complete. This process is repeated by using the unvisited vertex with the highest weight as the start for the construction of the next complex until no more complexes can be constructed.

Pseudocode

Input: graph $G = (V, E)$ with vertex weights w, weight percentage p, start vertex s, and current complex C.
If s was previously visited: return
Else:
t: $w_s \cdot p$, the inclusion threshold
For each neighbor vertex v of vertex s:
If $w_v > t$: add v to C
Recursion with v instead of s

Step 3 – Post-processing: The last step removes those constructed complexes that do not contain a single 2-core subgraph. It is optionally possible to increase the size of complexes and allow potential overlap between them, by adding neighbors of complex vertices to the complex, if the neighborhood density exceeds a "fluff" parameter without marking these vertices as visited. The remaining complexes are assigned a score and ranked accordingly. To favor larger and denser complexes, the score S_C of complex C is derived from the density d_C of the complex subgraph and its number of vertices $|V_C|$:

$$S_C = d_C \cdot |V_C|$$

The time complexity of the entire algorithm is polynomial with $O(|V| \cdot |E| \cdot h^3)$, where $|V|$ is the number of vertices, $|E|$ is the number of edges, and h is the vertex size of the average vertex neighborhood in the input graph, G.

5.4.1.3 Examples

MCODE can be run conveniently with the MCODE plugin of the Cytoscape software (https://apps.cytoscape.org/apps/mcode). In the PPIN of *S. cerevisiae* based on PPI data from the mentha database [12] that is shown in Figure 5.2, this results in 75 putative protein complexes when default parameters of MCODE are used (degree cutoff set to 2, vertex score cutoff set to 0.2, k-core set to 2, loops included, and maximum depth value set to 100). When compared to the gold standard CYC2008 dataset for *S. cerevisiae* compiled by Pu et al. [42], a majority of the complexes identified by MCODE showed partial overlap with multiple CYC2008 complexes. Figure 5.8 presents two examples from these 75 complexes. The large putative complex shown in panel (a) contains parts of several known protein complexes. It appears unlikely that it would assemble as a physical unit at one particular time in the yeast cell. In contrast, all vertices of the small complex shown in panel (b) belong to the known ribonuclease MRP complex.

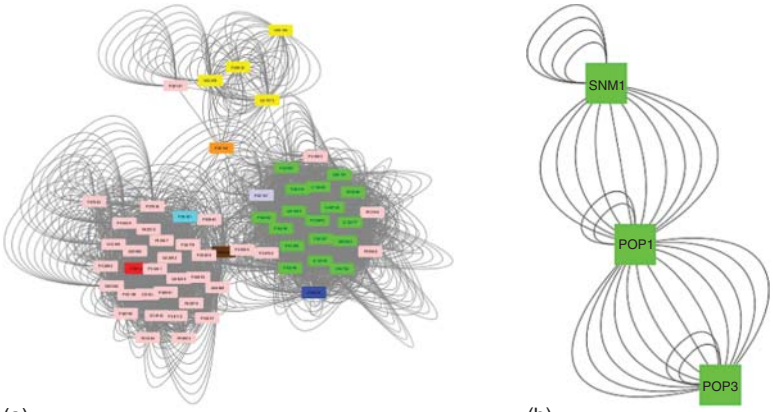


Figure 5.8 Two putative protein complexes constructed by the MCODE algorithm based on the complete interactome of *S. cerevisiae* according to PPI data from mentha. These complexes were then compared to the CYC2008 gold standard set. (a) Protein complex identified by MCODE with a score of 20.984, involving 60 vertices (proteins) and 3119 interactions. In this complex, 17 vertices (green) belong to the 19/22s regulator complex (CYC2008 lists 22 proteins for this complex), 1 vertex (purple) belongs to the 20S proteasome complex (14 proteins in CYC2008), 1 node (orange) belongs to the ISW1b complex (3 proteins in CYC2008), 1 vertex (red) belongs to the Noc1p/Noc2p complex (2 proteins in CYC2008), 1 vertex (blue) belongs to the Png1p/Rad23p complex (2 proteins in CYC2008), 4 vertices (yellow) belong to the RSC complex (17 proteins in CYC2008), 1 vertex (brown) belongs to the UTP B complex (6 proteins in CYC2008) and 1 node (cyan) belongs to the cytoplasmic ribosomal large subunit complex (81 proteins in CYC2008). The remaining vertices (Pink) are not annotated to any complexes in CYC2008. (b) Protein complex identified by MCODE with score of 2.5, 3 vertices, and 14 interactions. In this complex, all 3 vertices (green) belong to the ribonuclease MRP complex (CYC2008 lists 10 proteins for this complex).

5.4.2 Clustering with Overlapping Neighborhood Expansion (ClusterONE)

Subsequent to MCODE, Nepusz et al. [43] introduced the Clustering with Overlapping Neighborhood Expansion (ClusterONE) graph clustering algorithm that takes a weighted PPIN as input and constructs overlapping protein complexes. The algorithm introduced in Section 5.4.1, MCODE, detects clusters by identifying densely connected regions of a network and assigning them to non-overlapping complexes. However, in the case of a PPIN, proteins tend to have multiple functions and may, depending on the situation, belong to more than one complex. ClusterONE addresses the combinatorial nature of overlapping complexes and thus accounts for one protein potentially participating in multiple complexes. As mentioned in Maruyama and Kuwahara [44], the CYC2008 gold-standard data set for *S. cerevisiae* contains 408 protein complexes, out of which 216 pairs of two complexes overlap with each other. Most pairs (151) share only one protein, but 10 share 7 proteins and 1 pair even shares 17 proteins. As an example, the same authors discovered that the commitment complex and the U4/U6·U5 tri-snRNP complex that both bind to different types of RNA share the four proteins Smb1p, Smd1p, Smd2p, and Smd3p.

The algorithm ClusterONE detects overlapping complexes in mainly three steps: (i) The proteins are grouped based on high cohesiveness by a greedy algorithm that is run repeatedly from different starting proteins to identify multiple and overlapping complexes. (ii) The extent of overlap between complexes is quantified between each pair of groups, and those groups with an overlap score above a preset threshold are merged. (iii) Finally, all those complexes are discarded that are formed from fewer than three proteins and those with density below a preset threshold.

5.4.2.1 Definitions

For this algorithm, the input PPIN is given as a graph G that in addition to the set of protein vertices V and the set of protein interaction edges E also contains a set of edge weights W:

$$G = (V, E, W)$$

For a group of selected proteins V, one can distinguish internal edges, which represent interactions between members of V, and outgoing edges, which represent interactions between members of V and proteins in the rest of the PPIN, see Figure 5.9. The cohesiveness $f(V)$ of the selected proteins relative to the rest of the network can be assessed by comparing the summed weight of the internal edges $w^{\text{in}}(V)$ to the summed weight of the outgoing edges $w^{\text{out}}(V)$:

$$f(V) = \frac{w^{\text{in}}(\text{V})}{w^{\text{in}}(V) + w^{\text{out}}(V) + p \cdot |V|}$$

Since not all existing protein interactions are known, the penalty term $p \cdot \lceil V \rceil$ is added, which accounts for every member of V having p unknown outgoing interactions. The cohesiveness weighs the density of physical interactions among a group of proteins against the average density in their environment. High cohesiveness can mean two scenarios: (i) The group of proteins V forms dense and reliable edges

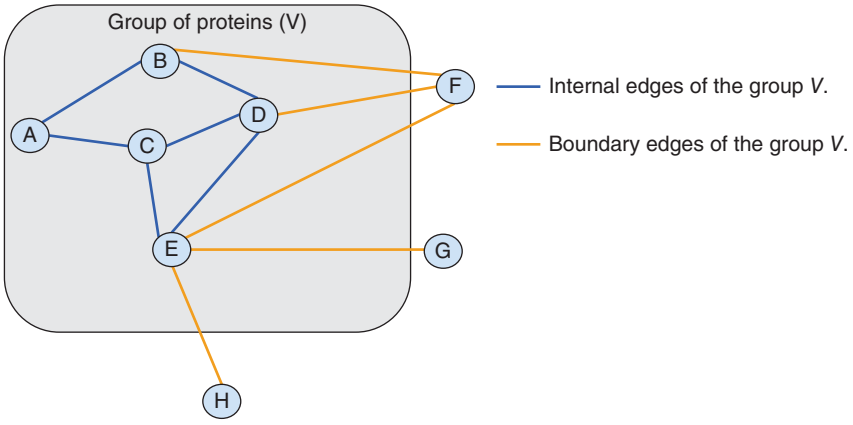


Figure 5.9 Schematic representation of a group of proteins (vertices in the grey box) selected within a PPIN. The blue lines represent the internal edges within the group. The orange lines represent the edges that connect vertices inside the group to the rest of the network. For example, with all edge weights set to 1 and $p = 0$, this group would have $w^{\text{in}}(V) = 6$ and $w^{\text{out}}(V) = 5$, resulting in cohesiveness of $f(V) = 6/11$ (with the penalty term set to zero).

among themselves (high $w^{\text{in}}(V)$), or (ii) the group of proteins is more or less separated from the rest of the network (low $w^{\text{out}}(V)$). Protein groups with cohesiveness values above $1/3$ can be considered as good candidates for putative complexes, since, above this threshold, the internal weights start to outweigh the external weights.

5.4.2.2 Algorithm

Step 1 – Group assembly: ClusterONE employs a greedy algorithm to assemble cohesive groups of proteins. Each group V initially consists of the unvisited protein with the highest degree in the PPIN up to that point. In each step, all proteins participating in outgoing interactions are evaluated: An external protein v is added to the group if doing so increases the cohesiveness $f(V)$ of the group, i.e. when $f(V+v) > f(V)$, whereas an internal protein is removed from the group if its removal improves group cohesiveness, i.e. when $f(V-v) > f(V)$. Once no further improvements to $f(V)$ can be made, the current group is considered to be a local optimum and the algorithm begins assembling the next group, until all proteins have been examined. This process is illustrated in an example in Figure 5.10.

Step 2 – Assembly of candidate complexes: Since ClusterONE allows proteins to participate in more than one group, this step examines the overlap between the locally optimal cohesive groups identified in the previous step. The overlap score $\omega(A, B)$ of two groups A and B is computed by considering the number of proteins both groups have in common, $|A \cap B|$, and the total number of proteins in each group:

$$\omega(A, B) = \frac{|A \cap B|^2}{|A||B|}$$

All pairs of cohesive groups with overlap score $\omega(A, B) > 0.8$ are labeled as connected, and all groups that are directly or indirectly connected to each other are

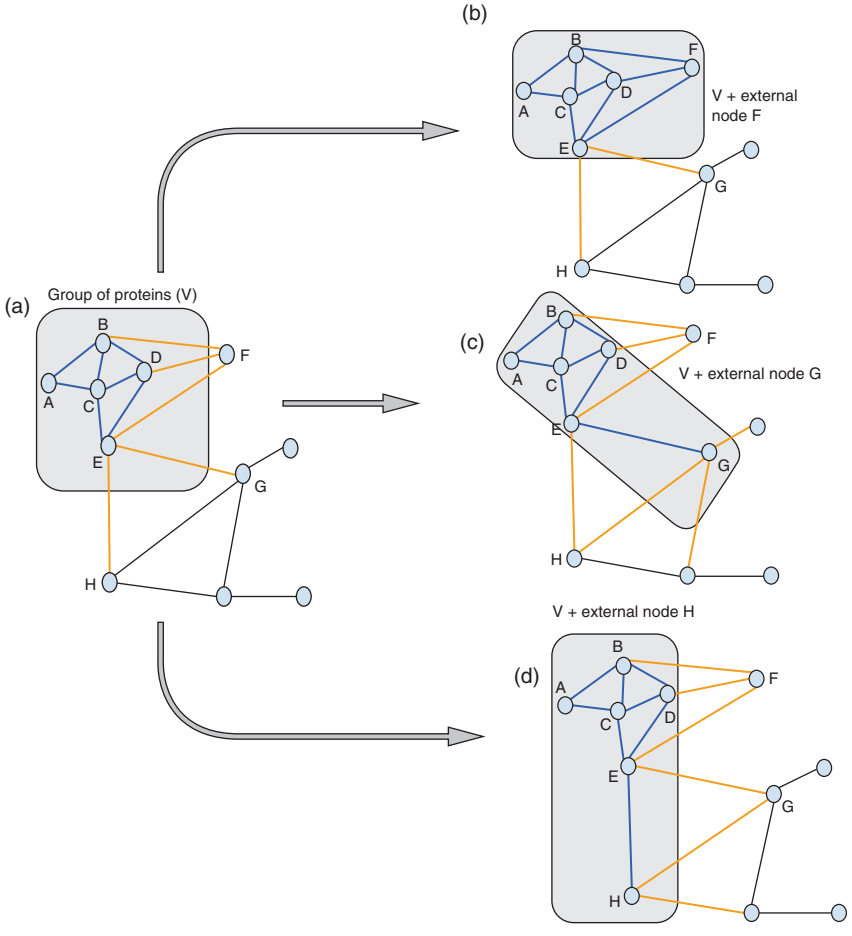


Figure 5.10 Workflow of the ClusterONE algorithm showing the cohesive growth of protein group V. (a) In this example, the group V consists of the five vertices A, B, C, D, and E. Assuming that all edge weights are set to 1 and the penalty $p = 0$, the cohesiveness of this group is $f(V) = 7/12$ with $w^{in}(V) = 7$ and $w^{out} = 5$. The group starts to grow by adding external vertices to or removing internal vertices from V based on the resulting changes in cohesiveness. The greedy algorithm adds an external vertex v only if $f(V + v) > f(V)$. Panels (b), (c), and (d) show different options for expanding V. (b) Adding the external vertex F would increase $f(V)$ to $f(V + F) = 10/12$. In contrast, (c) adding vertex G would lower $f(V)$ to $f(V + G) = 8/15$, and (d) adding H would similarly lower it to $f(V + H) = 8/14$. The greedy algorithm thus only adds vertex F to group V and the new group cohesiveness is $f(V) = 10/12$. In the next iteration, the expansion process terminates with $V = \{A, B, C, D, E, F\}$ as a locally optimal cohesive group, since adding G or H in addition to F would not increase $f(V)$ any further, with $f(V + G) = 11/15$ and $f(V + H) = 11/14$ both <10/12. The algorithm then restarts the expansion process by selecting the yet unvisited protein with the highest degree.

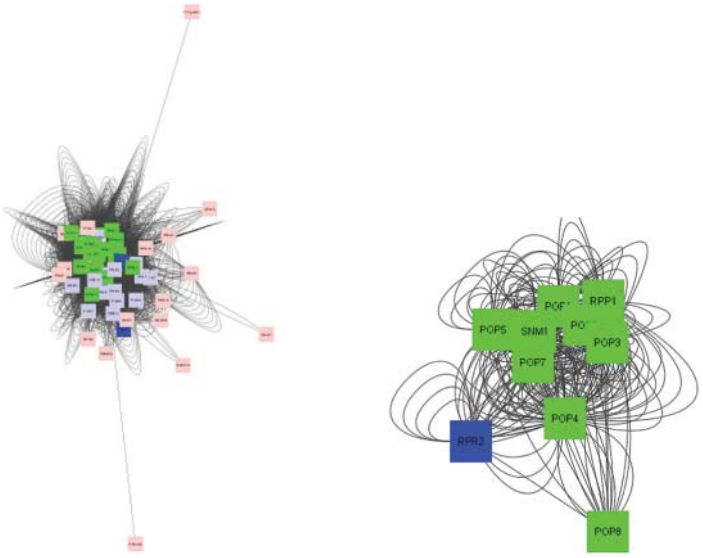
merged to form candidate complexes. Cohesive groups that do not overlap with and are not connected to other groups are classified as candidate complexes without merging.

Step 3 – Candidate filtering: The final step assesses the size and density d_C of each candidate complex C according to the provided threshold δ, with d_C as defined in Section 5.4.1.1. Only candidate complexes consisting of more than three proteins and $d_C > \delta$ are retained, whereas the rest are removed.

For the same PPIN from yeast (see Figure 5.2), the ClusterONE plugin of Cytoscape identified 842 clusters as putative protein complexes when using default parameters (minimum size of cluster set to 3, minimum density set to auto-tuned (0.3 for weighted graphs and 0.5 for unweighted graphs) and vertex penalty set to 2). Figure 5.11 shows three examples of these putative protein complexes (two are found in (a), one in (b)).

5.4.3 Domain-Aware Cohesiveness Optimization (DACO)

As mentioned earlier, ClusterONE accounts for one important characteristic of protein complexes, namely that they may overlap. However, it does not consider the



(a) (b)

Figure 5.11 Protein complex prediction based on the *S. cerevisiae* PPIN according to data from mentha using the ClusterONE algorithm. (a) Cluster with quality of 0.461, *p*-value of 0.000 22, 57 vertices, and 4346 interactions. Green vertices belong to the 19/22s regulator complex (22 out of 22 in CYC2008 detected by ClusterONE, MCODE only 17, see Figure 5.8a), purple vertices belong to the 20S proteasome complex (14 detected out of 14, MCODE only 1), and blue vertices belong to the Png1p/Rad23p complex (2 detected out of 2, MCODE only 1). As for MCODE, the cluster identified by ClusterONE contains further proteins that are not known to be part of this complex. (b) Protein complex identified by ClusterONE with quality of 0.613, *p*-value of 0.000 16, 10 vertices, and 245 interactions. The green vertices belong to the ribonuclease MRP complex, the blue vertex belongs to the nucleolar ribonuclease P complex. CYC2008 lists 10 proteins for the ribonuclease MRP complex. Out of them, ClusterONE identified 9 whereas MCODE identified only 3 (see Figure 5.8b).

spatial topology or physical aspects underlying the interactions. For instance, for multiple proteins to interact, all binding partners must be expressed at the same time, and they must not sterically overlap with other partners of the same protein complex. This second aspect was the main motivation behind the development of the Domain-Aware Cohesiveness Optimization (DACO) algorithm [45]. It is a combinatorial approach that adds a domain-level structural consideration for binding partners to the local cohesiveness optimization of ClusterONE and is similarly based on weighted protein interactions. The algorithm predicts protein complexes in a given PPIN on the basis of the following inputs: (i) A probability-weighted PPIN, (ii) a list of seed proteins, (iii) a threshold for the expansions, and (iv) a maximum search depth to control performance. An overview is shown in Figure 5.12.

Step 1 – DDIN construction: First, the algorithm translates the given weighted PPIN to the domain level by annotating proteins with known domains automatically retrieved from the databases Pfam [46] and InterPro [47], as well as domain-domain interactions queried from the databases IDDI [48] and DOMINE [49]. Once the domain-domain interaction network (DDIN) is constructed from the given PPIN, the algorithm executes DACO to identify putative protein complexes.

Step 2 – Greedy expansion: Similar to ClusterONE, DACO utilizes a greedy approach to form locally optimal cohesive groups of proteins. However, the groups are initialized with the seed proteins provided as input, not the unvisited proteins with the highest degree in the network. Due to the inclusion of domain-level information, the definition of incident and boundary proteins of a group differs slightly as well (Figure 5.13).

Using the weighted protein interactions, DACO calculates in each step the cohesiveness $f(V)$ for the current protein group V with the formula introduced in the previous Section 5.4.2.1 about ClusterONE. For each incident vertex v it considers whether the inclusion of v increases the group cohesiveness, i.e. $f(V+v) > f(V)$, and for each boundary vertex b whether its removal leads to an improvement, i.e. $f(V-b) > f(V)$. In cases where removing a boundary vertex b yields the largest increase, b is removed from V and the domains previously

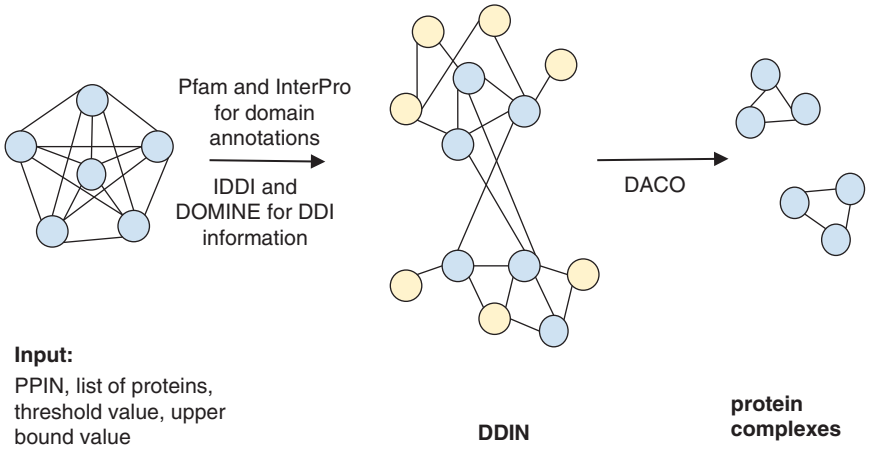


Figure 5.12 Workflow of the DACO algorithm.

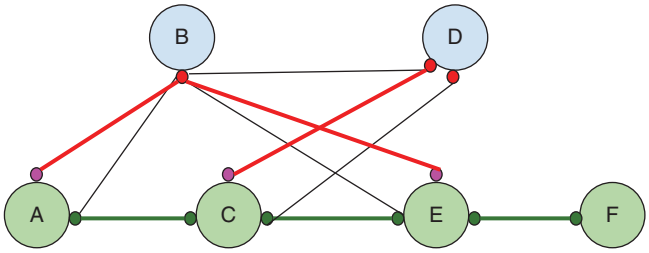


Figure 5.13 Incident and boundary proteins of a protein group *V*. The proteins (large circles) are annotated with domains (small circles) and connected by domain-domain interactions (edges). Members of group *V* are shown in green (A, C, E, F), as are internal edges and domains occupied by them. The blue incident vertices (B and D) are vertices outside of *V* that have an edge (red) to an unoccupied domain (pink) in *V*. Boundary vertices (A and F) are those members of *V* that have only one occupied domain.

occupied by the connections with b are marked as available. Should addition of incident vertex v lead to the largest improvement, DACO examines the supporting domain-domain interactions and, in the case of multiple options, selects the one with the highest probability derived from the weights of the PPIN. The corresponding domains are marked as occupied. If neither addition nor removal can improve group cohesiveness any further, the current group is considered locally optimal, and the algorithm continues with the next seed protein.

5.5 Summary

Protein–protein interaction networks provide an overview of the connectivity around specific proteins of a biological cell, the protein complexes in which these proteins are potentially involved, and the overall interactome of the organism. For a specific organism, a complete overview of all the existing PPIs is represented as a global PPI network. This PPIN typically has a "scale-free" topology where the distribution of the vertex degrees follows a power law with a negative exponent, leading to a higher frequency of highly connected "hub" proteins than is expected in a random graph. Protein–protein interaction data to construct these networks are accessible from two general kinds of public repositories, primary and metadatabases.

Protein complexes are crucial entities for regulating cellular activity and determining the behavior of a cell. Hence, to better understand the inner workings of cell, it is beneficial to identify protein complexes and examine how they interact with each other. Various algorithms were developed to recognize the protein complexes from PPINs. Here, we discussed three of them, MCODE, ClusterONE, and DACO. The MCODE algorithm constructs putative protein complexes by identifying densely connected areas of a PPIN. Promising seed proteins are identified by a vertex weighting strategy based on each protein's local neighborhood density. The ClusterONE algorithm improves on MCODE by also detecting partially overlapping protein complexes. Finally, the DACO algorithm considers whether multiple

interactions between one protein and other proteins can be realized simultaneously. Obviously, this is not the case when multiple interactions would involve the same binding interface. To resolve such cases, DACO also considers the domain nature of the individual proteins and knowledge about all domain-domain interactions characterized so far. In a formed complex, each protein domain is then only allowed to mediate one interaction each. Similar to ClusterONE, DACO identifies putative protein complexes having high cohesiveness in the PPIN and that may partially overlap, but also considers whether these can be physically realized when considering the domain makeup of the involved proteins.

Initially, each experimental assay could only detect a portion of the formed PPIs, the error rates were quite high for some methods, and the overlap between different technologies was rather low. In the meantime, these methods have become more mature, the datasets have expanded in size, and the coverage of the full PPINs has increased. It is difficult to estimate how much is still missing and how much is still to come. Over the past few years, interest has shifted to developing technologies, both of experimental and computational nature, which enable the characterization of condition-specific PPINs, e.g. for particular human tissues or for a particular disease condition [50]. These specific PPINs facilitate the study of interaction rewiring events, for example between healthy and disease conditions, which may affect protein complex formation and pathways of interest.

References

1 Milo, R. (2013). What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* 35: 1050–1055. https://doi.org/10.1002/bies.201300066.

2 Gorka, M., Swart, C., Siemiatkowska, B. et al. (2019). Protein complex identification and quantitative complexome by CN-PAGE. *Sci. Rep.* 11523: https://doi.org/10.1038/s41598-019-47829-7.

3 Barabási, A.-L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5: 101–113. https://doi.org/10.1038/nrg1272.

4 Bader, G.D., Betel, D., and Hogue, C.W.V. (2003). BIND: the biomolecular interaction network database. *Nucleic Acids Res.* 31: 248–250. https://doi.org/10.1093/nar/gkg056.

5 Orchard, S., Ammari, M., Aranda, B. et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42: D358–D363. https://doi.org/10.1093/nar/gkt1115.

6 Licata, L., Briganti, L., Peluso, D. et al. (2021). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 40: D857–D861. https://doi.org/10.1093/nar/gkr930.

7 Salwinski, L., Miller, C.S., Smith, A.J. et al. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32: D449–D451. https://doi.org/10.1093/nar/gkh086.

8 Oughtred, R., Rust, J., Chang, C. et al. (2021). The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* 30: 187–200. https://doi.org/10.1002/pro.3978.

9 Kotlyar, M., Pastrello, C., Malik, Z., and Jurisica, I. (2019). IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* 47: D581–D589. https://doi.org/10.1093/nar/gky1037.

10 Alonso-López, D., Campos-Laborie, F.J., Gutiérrez, M.A. et al. (2019). APID database: redefining protein–protein interaction experimental evidence and binary interactomes. *Database* baz005: https://doi.org/10.1093/database/baz005.

11 Prasad, T.S.K., Goel, R., Kandasamy, K. et al. (2009). Human protein reference database–2009 update. *Nucleic Acids Res.* 37: D767–D772. https://doi.org/10.1093/nar/gkn892.

12 Calderone, A., Castagnoli, L., and Cesareni, G. (2013). Mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods* 10: 690–691. https://doi.org/10.1038/nmeth.2561.

13 Porras, P., Barrera, E., Bridge, A. et al. (2020). Towards a unified open access dataset of molecular interactions. *Nat. Commun.* 11: 6144. https://doi.org/10.1038/s41467-020-19942-z.

14 Szklarczyk, D., Nastou, K., Lyon, D. et al. (2020). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49: D605–D612. https://doi.org/10.1093/nar/gkaa1074.

15 Gavin, A.C., Bösche, M., Krause, R. et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141–147. https://doi.org/10.1038/415141a.

16 Uetz, P., Giot, L., Cagney, G. et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403: 613–627. https://doi.org/10.1038/35001009.

17 Fields, S. and Song, O.-k. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340: 245–246. https://doi.org/10.1038/340245a0.

18 Han, J.-D., Dupuy, D., Bertin, N. et al. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat. Biotechnol.* 23: 839–844. https://doi.org/10.1038/nbt1116.

19 Wang, H., Kakaradov, B., Collins, S.R. et al. (2009). A complex-based reconstruction of the *Saccharomyces cerevisiae* interactome. *Mol. Cell. Proteomics* 8: 1361–1381. https://doi.org/10.1074/mcp.M800490-MCP200.

20 Winzeler, E.A., Shoemaker, D.D., Astromoff, A. et al. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285: 901–906. https://doi.org/10.1126/science.285.5429.901.

21 Giaever, G., Chu, A.M., Ni, L. et al. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418: 387–391. https://doi.org/10.1038/nature00935.

22 Zhang, Z. and Ren, Q. (2015). Why are essential genes essential? - the essentiality of *Saccharomyces genes*. *Microb. Cell* 2: 280–287. https://doi.org/10.15698/mic2015.08.218.

23 Shannon, P., Markiel, A., Ozier, O. et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13: 2498–2504. https://doi.org/10.1101/gr.1239303.

24 Ho, Y., Gruhler, A., Heilbut, A. et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415: 180–183. https://doi.org/10.1038/415180a.

25 Zhang, Q.C., Petrey, D., Garzón, J.I. et al. (2013). PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.* 41: D828–D833. https://doi.org/10.1093/nar/gks1231.

26 von Mering, C., Krause, R., Snel, B. et al. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417: 399–403. https://doi.org/10.1038/nature750.

27 Han, J.D., Bertin, N., Hao, T. et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430: 88–93. https://doi.org/10.1038/nature02555.

28 Batada, N.N., Hurst, L.D., and Tyers, M. (2006). Evolutionary and physiological importance of hub proteins. *PLoS Comput. Biol.* 2: –e88. https://doi.org/10.1371/journal.pcbi.0020088.

29 Agarwal, S., Deane, C.M., Porter, M.A., and Jones, N.S. (2010). Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS Comput. Biol.* 6: e1000817. https://doi.org/10.1371/journal.pcbi.1000817.

30 Pertea, M., Shumate, A., Pertea, G. et al. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* 19 (2018): https://doi.org/10.1186/s13059-018-1590-2.

31 Stelzl, U., Worm, U., Lalowski, M. et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122: 957–968. https://doi.org/10.1016/j.cell.2005.08.029.

32 Agrawal, M., Zitnik, M., and Leskovec, J. (2018). Large-scale analysis of disease pathways in the human interactome. *Pac. Symp. Biocomputing* 23: 111–122.

33 Menche, J., Sharma, A., Kitsak, M. et al. (2015, 2015). Uncovering disease-disease relationships through the incomplete interactome. *Science* 347 (6224): 1257601. PMID: 25700523; PMCID: PMC4435741. https://doi.org/10.1126/science.1257601.

34 Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R. et al. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 43: D470–D478. https://doi.org/10.1093/nar/gku1204.

35 Shin, W.-H., Christoffer, C.W., and Kihara, D. (2017). In silico structure-based approaches to discover protein-protein interaction-targeting drugs. *Methods* 131: 22–32. https://doi.org/10.1016/j.ymeth.2017.08.006.

36 Han, Y., Wang, C., Klinger, K. et al. (2021). An integrative network-based approach for drug target indication expansion. *PLoS One* 16: e0253614. https://doi.org/10.1371/journal.pone.0253614.

37 Arkin, M.R., Tang, Y., Wells, J.A. (2014). Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. *Chem. Biol.*, 21, 1102–1114, https://doi.org/10.1016/j.chembiol.2014.09.001.

38 Voet, A., Berenger, F., and Zhang, K.Y. (2013). Electrostatic similarities between protein and small molecule ligands facilitate the design of protein-protein interaction inhibitors. *PLoS One* 8: e75762. https://doi.org/10.1371/journal.pone.0075762.

39 Voet, A., Ito, A., Hirohama, M. et al. (2014). Discovery of small molecule inhibitors targeting the SUMO–SIM interaction using a protein interface consensus approach. *MedChemComm* 5: 783–786. https://doi.org/10.1039/C3MD00391D.

40 Guimei, L., Limsoon, W., and Hon, N.C. (2009). Complex discovery from weighted PPI networks. *Bioinformatics* 25: 1891–1897. https://doi.org/10.1093/bioinformatics/btp311.

41 Bader, G.D. and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf.* 4: 2. https://doi.org/10.1186/1471-2105-4-2.

42 Pu, S., Wong, J., Turner, B. et al. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* 37: 825–831. https://doi.org/10.1093/nar/gkn1005.

43 Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* 9: 471–472. https://doi.org/10.1038/nmeth.1938.

44 Maruyama, O. and Kuwahara, Y. (2017). RocSampler: regularizing overlapping protein complexes in protein-protein interaction networks. *BMC Bioinf.* 18: 491. https://doi.org/10.1186/s12859-017-1920-5.

45 Will, T. and Helms, V. (2014, 2014). Identifying transcription factor complexes and their roles. *Bioinformatics* 30: i415–i421. https://doi.org/10.1039/bioinformatics/btu448.

46 Mistry, J., Chuguransky, J., Williams, L. et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49: D412–D419. https://doi.org/10.1093/nar/gkaa913.

47 Blum, M., Chang, H., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson. L., Salazar, G.A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D.H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N, Thomas. P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A., Finn, R.D. (2020). The InterPro protein families and domains database: 20 years on*Nucleic Acids Res.*, 49, p. D344-D354, https://doi.org/10.1093/nar/gkaa977.

48 Kim, Y., Min, B., and Yi, G.-S. (2012). IDDI: integrated domain-domain interaction and protein interaction analysis system. *Proteome Sci.* 10: S9. https://doi.org/10.1186/1477-5956-10-S1-S9.

49 Yellaboina, S., Tasneem, A., Zaykin, D.V. et al. (2011). DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res.* 39 (Database Issue): D730–D735. https://doi.org/10.1093/nar/gkq1229.

50 Will, T. and Helms, V. (2019). Differential analysis of combinatorial protein complexes with CompleXChange. *BMC Bioinf.* 20: 300. https://doi.org/10.1186/s12859-019-2852-z.

6

Structure, Composition, and Modeling of Protein Complexes

Olga V. Kalinina

Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), 66123 Saarbrücken, Germany
Saarland University, Drug Bioinformatics, Medical Faculty, 66421 Homburg, Germany
Saarland University, Center for Bioinformatics, 66123 Saarbrücken, Germany

6.1 Protein Complex Structure

6.1.1 Protein Quaternary Structure

The sequence of amino acids in a protein forms its primary structure, local 3D arrangement of amino acids defines its secondary structure, and structural contacts between secondary structure elements lead to the formation of a tertiary structure that fully describes the 3D shape of a single protein chain. However, many proteins form assemblies of higher order, where several (identical or nonidentical) chains come together to form a noncovalently bound protein complex. Such an assembly is called a quaternary protein structure and may include, along with proteins, other molecules such as metal ions or low molecular weight cofactors. Individual protein chains in such an assembly are called *protomers* or *subunits*; if all subunits are identical proteins, such an assembly is called a *homomer* (*homooligomer*), and *heteromer* (*heterooligomer*) otherwise. Depending on the number of subunits in a complex, they may be called *dimers* (2 subunits), *trimers* (3), *tetramers* (4), etc.

The higher-order protein complexes described above exist for a reason – their functionality often extends beyond the sum of functionalities of their individual units. Functions of multi-protein complexes may include co-localization of multiple active sites that act as a conveyor belt for the synthesis of a certain molecule (e.g. the tryptophan synthase complex, [1]), formation of a shared active site between subunits (e.g. NADP-dependent isocitrate dehydrogenase family, [2]), combinatorial swapping of subunits (e.g. in antibodies), and formation of large cellular structures (e.g. actin filaments, microtubules) or of key molecular machines (ribosome, proteasome, DNA and RNA polymerase complexes, etc.). Another example of functional change related to complex formation is cooperative ligand binding – a phenomenon when the number of binding sites occupied by a ligand depends non-linearly on the ligand concentration [3]. The binding of some ligands may induce a large-scale change in a protein structure (e.g. binding of calcium atoms switches

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

the conformation of calmodulin from closed to open), a phenomenon known as *induced fit* (see also Chapter 4). Induced fit, however, may be a consequence not only of ligand binding but also of protein complex formation (e.g. actin/myosin interaction, for further examples see [4]).

While some proteins exist almost solely in their complex form – such complexes are commonly known as stable – others are only formed for short periods of time. Such complexes are labeled as transient complexes. In this chapter, we will mainly focus on the prediction of 3D structure of stable protein complexes that consist of more than two subunits.

The experimental methods for protein structure elucidation include X-ray crystallography, NMR spectroscopy, electron microscopy (EM), cryo-EM tomography, and approaches that present a combination of the aforementioned main ones. While NMR spectroscopy is better suited to the resolution of small structures, other approaches are theoretically capable of resolving large multi-protein complexes. However, for the most commonly used approach, X-ray crystallography, the difficulty of solving the structure often correlates strongly with the size of the protein, thus limiting its applicability in the field of protein complex structure prediction.

Like all other experimentally resolved protein 3D structures, experimentally resolved multi-protein complexes are stored in the Protein Data Bank (PDB). Naturally, the interest of structural biologists has been directed to multi-protein structures that are involved in the key biochemical pathways in the cell, and a great many of them have been resolved (see a virtual tour of the PDB[1] and Chapter 5). However, for a large (and probably unknown) number of proteins that participate in multimeric assemblies, only structures of individual monomers have been experimentally resolved. The methods and resources reviewed in this chapter can be split into seven major categories: classification of protein-protein interaction interfaces, classification of protein complexes, assignment of quaternary structure from X-ray crystallography data, combinatorial docking, homology-based complex reconstruction, de novo prediction of complexes from sequence, and assisted docking (Table 6.1).

6.1.2 Classification of Protein–Protein Interaction Interfaces

At the time of this writing, the PDB contains 173 090 protein-containing structures, of which 97 040 (56%) contain more than one protein chain in their assumed biologically relevant stoichiometry (biological assembly). A nonredundant set of 19 855 structures from the PDB constructed by Marsh and Teichmann [52] contains 7972 biological monomers, 9206 homooligomers, and 2677 heterooligomers. Numerous methods have been developed to classify this wealth of structural information. One of the approaches is to cluster protein–protein interactions based on the evolutionary history and/or functional family of the interacting partners, and not on geometry and biophysical properties of the interacting interfaces (although in practice the former often implies the latter). In this context, one considers the problem of

1 https://cdn.rcsb.org/pdb101/molecular-machinery/

Table 6.1 Overview of methods and databases.

Category	Example tools	Comments
Classification of protein–protein interaction interfaces	iPfam [5]	Pfam annotations
	DOMMINO [6, 7]	SCOP and SUPERFAMILY annotations
	PIBASE [8]	SCOP annotations, clustering based on geometry
	SCOPPI [9]	SCOP annotations, geometric characterization of interfaces
	SNAPPI-DB [10]	SWISS-PROT, SCOP, CATH, Pfam, InterPro, GO terms annotations; structural alignment of similar chains
	DOMINE [11]	Meta-method including many of above mentioned and other tools
	INstruct [12]	Includes additional information from low-resolution methods
	ProtCID [13–15]	Clustering based on Pfam
Classification of protein complexes	PiQSi [16], 3Dcomplex [17]	Classification of large multimeric complexes based on their topology
Assignment of quaternary structure from X-ray crystallography data	PQS [18]	Change of the solvent-accessible surface area (Δasa) upon complex formation
	PITA [19]	Δasa; pairwise residue interaction potentials
	PISA [20]	Change of the Gibbs free energy (solvation, contact-dependent and electrostatic interactions, entropy)
	Valdar and Thornton [21]	Δasa, residue conservation on the interface; one-layered artificial neural network
	NOXclass [22]	Δasa, interface complementarity, amino acid composition, and conservation; support vector machine (SVM)
	DiMoVo [23]	Geometric complementarity based on Voronoi tessellations; SVM
	IPAC [24]	Geometric and physicochemical features; Bayes classifier
	IChemPIC [25]	Physicochemical properties of the interface contacts; random forest
	Luo et al. [26]	Geometric and physicochemical properties of the interface contacts, amino acid, and secondary structure composition, amino acid propensities; random forest
	EPPIC [27; 28]	Per-residue K_a/K_s ratio
	QSalign [29]	Modified structural similarity score (TM-score) for complexes

(*continued*)

Table 6.1 (Continued)

Category	Example tools	Comments
Combinatorial docking	Berchanski and Eisenstein [30]	Symmetric homooligomeric complexes: dimers of dimers
	Berchanski et al. [31]	Symmetric homooligomeric complexes
	SymmDock [32]	Symmetric complexes
	M-ZDOCK [33]	Symmetric complexes with cyclic symmetry
	ClusPro [34]	Symmetric complexes with cyclic, D_2 and D_3 symmetry
	DockTrina [35]	Nonsymmetric trimers
	CombDock [36]	Nonsymmetric complexes of any stochiometry
	Multi-LZerD [37]	Nonsymmetric complexes of any stochiometry
	DockStar [38]	Nonsymmetric star-like complexes of any stochiometry
	3D-MOSAIC [39]	Nonsymmetric complexes of any stochiometry
Homology-based complex reconstruction	PRISM [40]	
De novo prediction from sequence	M-TASSER [41]	Based on prediction of 3D structure of monomers with TASSER
Assisted docking	ATTRACT-EM [42]	Leverages low-resolution electron microscopy (EM) data
	MDFF [43]	Leverages low-resolution EM data followed by molecular dynamics simulations
	PRISM-EM [44]	Leverages low-resolution EM data
	PROXIMO [45]	Restraints from radical probe mass spectrometry (RP-MS)
	Kiselar et al. [46]	Restraints from hydroxyl radical footprinting
	iSPOT [47]	Restraints from small-angle X-ray scattering (SAXS) and hydroxyl radical footprinting
	Berchanski et al. [48]	Restraints from large-scale protein–protein interaction screens
	HADDOCK [49–51]	Up to six protomers, various experimental and bioinformatics restraints

protein–protein interaction as the problem of domain–domain interactions, since domains are the structural, functional, and, most importantly in this context, evolutionary units in the protein space.

iPfam [5] does this by providing Pfam [53] annotations for every pair of interacting proteins in the PDB. DOMMINO [6, 7] solves the same problem by using SCOP [54] and SUPERFAMILY [55] annotation to this end. SCOP classification is also used for construction of PIBASE [8], where interfaces are additionally clustered based on

their geometry, and complexes are clustered based on their topology. SCOPPI [9] adds GO term annotations. In 3did [56], interfaces are clustered based on geometric similarity of the interacting monomers and their mutual arrangement. SNAPPI-DB [10] provides SWISS-PROT, SCOP, CATH, Pfam, InterPro, and GO terms annotations, as well as multiple structural alignments for pairs of interacting proteins from the same structural or functional family. DOMINE [11] integrates a large collection of classification and prediction tools that enable the authors to build the largest at their time database of protein–protein and domain–domain interfaces. INstruct [12] leverages information from low-resolution experimental protein–protein interaction databases in combination with structure reconstruction methods similar to homology modeling to create a high-resolution collection of protein–protein interaction interfaces for humans and six common model organisms.

Another way to classify 3D multi-protein assemblies is to classify interaction interfaces between subunits using the sequence homology of the interacting partners. This has been implemented in the ProtCID[2] database [13–15]. This resource clusters protein interfaces with respect to Pfam [53] families of the interacting subunits and operates on the level of protein domains. The authors show how these clusters can generate unexpected functional hypotheses for some oligomers.

6.1.3 Classification and Evolution of Protein Complexes

PiQSi [16] was developed as a database that stores and links the data on the quaternary structure of homologous proteins from different organisms. As in other databases centered around protein interactions, each subunit of a complex is viewed as a node of a graph, and edges between the nodes represent interaction interfaces.

The first version of PiQSi was curated manually, but afterward, it became a part of 3Dcomplex[3] [17], a hierarchical classification scheme of protein complexes, similar in spirit to SCOP [54], and CATH [57]. In 3Dcomplex, protein assemblies are also represented as graphs, which are aligned by exhaustive mapping while considering three cost functions. The cost functions aim to represent (i) structural similarity of the monomers (same SCOP domain composition), (ii) missing nodes in one of the graphs, and (iii) inconsistency between edges. Symmetric complexes are also identified and classified into their respective groups. Combining the similarity of graph topologies with different levels of sequence identity between monomers from different complexes produces a hierarchical classification scheme used by 3Dcomplex.

A striking observation in any collection of multi-protein assemblies is that the majority of them are symmetric. In the context of protein structures, symmetry usually refers to rotational symmetry. In the above-mentioned nonredundant collection of protein structures [52], 77% of all complexes are homomeric (comprise identical proteins), and the vast majority of them are symmetric. This can be explained by the fact that it is easier to resolve such assemblies experimentally, and/or that symmetric homomeric complexes are energetically more stable. Simulations confirm the latter option

2 http://dunbrack2.fccc.edu/ProtCiD/Default.aspx
3 http://shmoo.weizmann.ac.il/elevy/3dcomplexV6/Home.cgi

and suggest a thermodynamic route for the emergence of symmetry, at least for homodimers [58].

Analysis of the evolution of protein–protein interaction networks and multiprotein complexes from different organisms further suggests that duplication of genes that encode proteins forming homomeric complexes is the driving force of protein complex evolution. This leads to overrepresentation of paralogous (evolutionary related and similar in sequence and structure) subunits in protein complexes [59, 60]. Examples of such complexes are hemoglobin or TRiC/CCT chaperonin. Interestingly, the overall shape of the ancestral homomeric and descendant heteromeric complexes is very similar. For example, for the TRiC/CCT chaperonin, the ancestral complex of the archaeal thermosome from *Thermococcus* sp. with 16 identical subunits was used as a template for homology modeling of the eukaryotic TRiC complex [61].

Using electrospray mass spectrometry, it has been confirmed that the actual dynamic assembly pathways mimic evolutionary paths for complex formation [62–65]: for example, tetramers assemble as dimers of dimers. It was shown that assembly and disassembly of protein complexes proceed stepwise, and often more than one oligomeric state is possible in equilibrium. Dimerization of subcomplexes and cyclization are the two most common mechanisms in complex assembly irrespective of whether homomers or heteromers are considered. Analysis of protein complexes whose quaternary structures can be described as a subset of one another further supports this evolutionary scenario. This creates an implicit periodic table of protein complexes [62].

These insights provide an evolutionary footing for endeavors to classify protein quaternary assemblies, such as 3Dcomplex [17]. 3Dcomplex is a hierarchical classification that considers all biological assemblies from the PDB as a starting point. Each complex is turned into a graph, in which the nodes represent individual protomers, and edges represent interaction interfaces between them. On the highest classification level, all graphs of the same topology are grouped together. Further, the complexes are split into smaller groups by considering the content of SCOP domains for each chain, the number of nonidentical chains in each complex, symmetry of the complex, and different levels of sequence identity between chains from different complexes that can be mapped to each other after graph superimposition. Altogether, 12 levels of classification are defined. At the time of construction in 2006, the classification included 21 037 groups of complexes with 192 symmetric and 265 nonsymmetric graph topologies. More than 96% of them constituted complexes with less than 10 subunits, and, as mentioned above, a vast majority of them are symmetric.

6.2 Methods for Automated Assignment of Biological Assemblies

A total of 87.6% of all structures in the PDB have been resolved by X-ray crystallography. This method is based on creating crystals of the purified target protein,

i.e. repeating identical units spaced equally in three dimensions. A unit of such crystals is called the *asymmetric unit*, and diffraction patterns obtained by illuminating the crystal with an X-ray beam allow for reconstructing 3D coordinates of the non-hydrogen atoms in the asymmetric unit. However, proteins often function as multi-polypeptide chain assemblies, which in the PDB are called *biological assemblies* (see Section 6.1.2). An asymmetric unit may contain one, several, or a part of a biological assembly, and hence a problem presents itself: How to deduce the correct biological assembly from the atom coordinates in an asymmetric unit? In this section, we overview bioinformatics tools developed to this end. First, we consider purely geometric tools and further present tools that use machine-learning methods or additional evolutionary data.

6.2.1 Assignment from Crystallographic Data

One of the earliest tools developed for identifying biological assemblies from crystallographic data is PQS, the protein quaternary structure file server [18]. The key here is to calculate Δasa, the change of the surface area available to the solvent upon complex formation. This idea has been proposed earlier [66]. First, all potential quaternary assemblies are built by grouping chains in the asymmetric unit and applying symmetry operations to it. For each candidate assembly, Δasa per chain is calculated and a cutoff of 400 Å^2 is used. This cutoff is based on the empirical observation that in native complexes Δasa is ∼370–4750 Å^2 for homo-dimers and ∼640–3230 Å^2 for heterodimers [67]. In addition, the automated classifier takes into account the number of buried residues, relative solvent accessible area in the complete assembly, and the change of the solvation energy of folding [68] between the isolated monomer and the assembly, calculated for each monomer separately. Virus capsids are treated differently to account for the icosahedral structure of many capsids. This method was further improved by including pairwise residue interaction potentials across candidate interaction interfaces (PITA, [19]). Other geometrical considerations, such as Voronoi tessellations, can also be applied [23].

A long-standing *de facto* standard for predicting biological assemblies is PISA [20], which is used by the PDB for the assignment of biological assemblies in cases when this information is not provided by the authors. PISA postulates that the change of the Gibbs free energy defines assembly stability and the interaction energy of the interfaces and entropy upon assembly are the major factors that drive complex formation, and provides a computational method that allows to assess them. Change of the interaction energy depends on the change of solvation, spatial contacts, and electrostatic interactions, and change of entropy can be estimated from the changes of translational, rotational, and vibrational entropies, as well as the entropy of the surface atoms. Candidate assemblies are identified by enumerating all combinations of interfaces involved in interactions between two subunits in the asymmetric unit, and the change of the Gibbs free energy upon complex dissociation ΔG_{diss} is calculated. The correct assembly was predicted as the one, for which ΔG_{diss} is greater than 0, whereas ΔG_{diss} is less or equal to 0 for all other assemblies of the same or larger size. In the original publication, PISA was tested on 218 protein–protein

(ranging from monomers to hexamers) and 212 protein-DNA (ranging from dimers to decamers) biological assemblies assigned manually, and produced correct predictions in 84–100% of the cases.

6.2.2 Employing Machine-Learning Methods

With the advent of machine learning, new methods have been introduced that learn certain properties of native assemblies from expert assignments in an automated way and then can be applied to predicting novel assemblies from crystallographic data. An example of a relatively simple problem that can be formulated in this setting is to distinguish between native dimers and monomers [21]. As described in Section 6.2.1, an asymmetric unit may not coincide with the biologically relevant multimeric state, and thus a set of native dimers and monomers has to be considered, and symmetry operations have to be applied to create further multimeric complexes. Two features are considered: contact area (measured as the change of the solvent accessible area upon dimerization) and residue conservation of the interface. A simple one-layered artificial neural network was trained to this end and achieved an accuracy of 98.3%. More methods that leverage evolutionary information will be considered in Section 6.2.3.

A more refined machine learning-based classification into obligate and nonobligate interactions (to differentiate between stable and transient complexes), as well as crystal contacts, was implemented in NOXclass [22]. Several features, including buried surface area, interface complementarity, amino acid composition, and conservation on the interface, were used to create an input, and a support vector machine (SVM) model was trained. The method achieved an accuracy of 91.8% for the three-class (crystal and biological complexes, monomers) classification problem. Another tool for such three-class classification represents protein complexes in terms of properties of their Voronoi tessellations (DiMoVo, [23]). In this approach, each residue is represented by a point in the 3D space, and a plane is placed mid-way and orthogonal to a line between every two residues. This induces a tessellation of the 3D space around the protein structures, whose geometric properties proved to be useful in different problem settings. Indeed, DiMoVo reaches an accuracy of 97% on an extended dataset, whereas NOXclass showed an accuracy of only 76%.

Due to the composition of the training sets, NOXclass and DiMoVo were predominantly trained to recognize native homodimers. Heterooligomers, on the other hand, typically present a larger problem for their correct assignment. IPAC [24] devised a training set focused specifically on heteromeric complexes larger than a dimer and trained a Bayes classifier to detect monomers and native assemblies using a large panel of geometric and physicochemical features. The method demonstrated superior performance compared to PISA and PQS. The authors specifically analyzed the cases where these tools fail. When compared to NOXclass and DiMoVo, IPAC demonstrated a consistent coverage (percent of correctly recovered quaternary structures) of >90% irrespective of whether the validation set contained homo- or heteromeric complexes, whereas DiMoVo failed on the heteromers. Interestingly, NOXclass performed very well for heteromers but was inferior to IPAC for homomeric dimers.

A random forest with 45 input features for discrimination between biological and crystallographic interaction interfaces was implemented in IChemPIC [25]. The authors constructed a very large training dataset of 400 homomeric and heteromeric interfaces. The descriptors used for these interfaces are based on intermolecular interactions (hydrophobic, aromatic, hydrogen bond, ionic bond) between the two selected chains. Compared to PISA, NOXclass, DiMoVo, and EPPIC ([27], see next section for an extended discussion of this tool), IChemPIC demonstrates a superior accuracy (75.0%) and specificity (76.0%) compared to all tools except EPPIC (which is not surprising, since EPPIC uses additional evolutionary information). However, NOXclass is more sensitive (87.8%) and DiMoVo more precise (85.7%) than IChemPIC (74.0% and 75.5%, respectively).

Another random forest-based method [26] uses features derived from geometric and physicochemical properties of the interfaces, amino acid and secondary structure composition, and amino acid propensities to different protein regions. Random forests allow for an easy feature importance analysis, and here features related to the tight packing of residues on the interface and their hydrophobicity proved to be characteristic for biological interfaces. A dataset created for evolutionary analysis of protein–protein interaction interfaces was used for training ([28], also see Section 6.2.3), and the method showed a superior performance across a wide range of statistical measures compared to DiMoVo, PITA, PISA, and EPPIC.

6.2.3 Leveraging Evolutionary Information

It has been observed that homologous proteins tend to form similar assemblies. Indeed, the geometry of interactions between homologs bears significant similarity [69]. Hence information on quaternary structure of some assemblies can be transferred to homologous complexes from other organisms. Another evolutionary consideration that can be taken into account is that amino acids on protein–protein interaction interfaces tend to be more conserved [21]. The first consideration lays the foundation for protein assembly classification schemes reviewed in Section 6.2.2, and the latter is the idea for the classification of biological and crystal interfaces by the tool EPPIC [27, 28]. The authors first showed that the per-residue K_a/K_s ratio (the ratio between nonsynonymous and synonymous substitutions) has characteristic values in the "core" of an interaction interface (the innermost tightly interacting part that gets completely buried upon complex formation) and at its "rim" (outer part of the interface that gets only partially buried) for native biological interfaces, which allows to distinguish them from crystal contacts. In EPPIC, they combined this with a geometric size of the core interface region using a set of *ad hoc* criteria to optimize performance. Further, these predictions were used to identify correct assemblies among all possible assemblies present in a symmetric unit or that can be obtained by symmetry operations upon it [27].

Another method that uses the overall structural similarity of complexes to detect correct assemblies is QSalign [29], which employs a modified TM-score (a widely used measure of structural similarity) [70] developed specifically for comparing 3D structures of complexes, as opposed to 3D structures of individual chains in

the original implementation. This method shows a uniform low error rate of 4% both for dimers and higher-order complexes, whereas the error rate of other tools (PISA and EPPIC) is larger for the latter. Separately, a version for detection of monomers (anti-QSalign) was developed. Integrating all three tools (PISA, EPPIC, and QSalign/anti-QSalign) produced excellent predictions for all three categories of assemblies (monomers, dimers, and higher-order oligomers), demonstrating complementary nature of the methods. The integrated method was implemented in a database called QSbio, which has since also become a part of 3Dcomplex.

Further, it was shown [71] that a combination of residue conservation on the interface, structural clustering of interfaces, and other interface composition descriptors allows to improve the selection of homologous templates for complex reconstruction. Specifically, the suggested method can correctly resolve cases when several quaternary structures are possible within one family of homologous proteins (e.g. for fructose biphosphate aldolases, monomers, dimers, tetramers, and hexamers have been reported). Additionally, the authors have compiled a large dataset of 807 nonredundant proteins with experimentally validated quaternary structures that are balanced to contain both homo- and heterooligomers with varying stoichiometry. At the time of construction of this dataset, for all homooligomeric complexes and for 64% of heterooligomeric complexes a homologous modeling template complex with sequence identity <95% could be identified in the SWISS-MODEL template library [72].

6.3 Computational Approaches to Predicting 3D Structure of Protein Complexes

In this section, we focus on tools that predict 3D structures of multi-protein complexes where the number of monomers is greater than two. These tools often utilize pre-computed structures of dimeric assemblies that can be predicted using one of the protein–protein docking methods reviewed in Chapter 4.

6.3.1 Combinatorial Docking

Docking methods traditionally focus on predicting 3D structure of binary protein complexes (homo- and heterodimers, see Chapter 4). These tools involve geometric and physicochemical compatibility of the two subunits in the 3D space that may or may not include a certain degree of flexibility of individual protein chains. Computationally, this problem is hard enough, due to the size of the search space that needs to be evaluated, and different optimization techniques can be applied. The energy potentials that are used to evaluate candidate poses are statistically derived and known to be imprecise, such that near-native conformations may not be scored best. With this in mind, the problem of *multi*-protein docking appears to be even harder, probably intractable. However, it turns out that this complexity can be leveraged by computational tools in the form of additional constraints that allow for predicting complex assemblies.

Symmetric complexes. One type of multi-protein complexes that allow for a relatively easy reconstruction are homooligomers characterized by a dihedral (D_n) or cyclic (C_n) symmetry. In cyclic symmetry, the monomers are related by rotations by $360°/n$ degrees: every such rotation in 3D produces a complex identical to the original one. Dihedral symmetry in addition to these rotational symmetries includes a perpendicular twofold rotation.

Protein complexes of $2n$ subunits are often n-mers of dimers, with D_2-symmetric homotetramers being particularly abundant [17]. This property has been leveraged in a geometric-based tool [30], where first dimer complexes are reconstructed using an in-house docking method MolFit [73], and then a dimer of dimers is formed while respecting the symmetry restraints. Later, the same group of authors extended their method to model homooligomeric complexes with C_n or D_n ($n > 2$) symmetries [31]. The authors demonstrate the applicability of their method on several protein complexes with cyclic and dihedral symmetries, including recent CAPRI targets (see Chapter 4), achieving RMSD in low-Ångström range.

Another method for reconstruction of cyclically symmetric complexes, SymmDock [32], uses a different docking algorithm to produce initial dimers, PatchDock [74], that is based on the rigid-body representation of monomers and surface matching to identify the interaction interface between them. Each matching pair defines a symmetry axis, and the axis that allows for the best closed circle of monomers is chosen. If a monomer can be split into a set of rigid substructures connected by flexible hinges, these substructures can be considered separately and then efficiently combined into a plausible conformation while docking. In this process, symmetric complexes are constructed for each rigid substructure and then their axes are checked for consistency. In almost all test cases, near-native conformations of the symmetric complexes were successfully recovered. However, it must be noted that the tests for all tools mentioned in this section were run on bound conformations (monomers were derived from experimental structures of the bound complexes, cf. Chapter 4); hence the performance may be overestimated.

A number of other tools use similar ideas to reconstruct cyclically or dihedrally symmetrical complexes: M-ZDOCK [33] for cyclically symmetric complexes, ClusPro [34] for cyclically, D_2 and D_3 symmetric complexes use their respective rigid-body docking algorithms to predict conformations of dimers.

Nonsymmetric complexes. Dealing with nonsymmetric and possibly heteromeric protein complexes presents a larger computational difficulty. The simplest kind of nonsymmetric complexes with more than two monomers are hetero-trimers, and a tool DockTrina [35] aims to predict their structures. The development of a specialized software for trimers is justified by the large number of trimers among naturally occurring complexes (see Section 6.1.3).

For a complex of three monomers A, B, and C, the algorithm starts with pairwise docking poses of all combinations A–B, B–C, and A–C that are generated with yet another rigid-body docking tool Hex [75]. For each set of pairwise poses, they are superimposed in all combinations, such that the second monomer of the last pair is the first monomer of the first pair: For example, a pose of A–B and a pose of B–C are superimposed on top of each other using the monomer B as an anchor.

Then a pose of C–A can be superimposed on top of this complex using the monomer C as an anchor, which will produce a (virtual) second copy of the monomer A, denoted A′, in the complex. If we denote transformations associated with these individual docking poses T^{AB}, T^{BC}, and T^{CA}, the coordinates of A′ can be expressed as $A' = T^{AB} \cdot T^{BC} \cdot T^{CA} \cdot A$. Ideally, if all used docking poses correspond to the actual conformations of the dimers in the sought trimer, A′ should coincide with A. In reality, one can measure the deviation, for example as RMSD between the coordinates of A and A′, which provides a measure of internal consistency of the candidate complex. All candidate trimers generated in this way are then scored with a score $\text{Score} = \text{Score}^{AB} + \text{Score}^{BC} + \text{Score}^{CA} + 0.25\frac{\text{Score}^{max}}{\text{RMSD}}$, where Score^{AB}, Score^{BC}, Score^{CA} are the scores of the corresponding docking poses, Score^{max} is the sum of maximum scores of the three dimers, and RMSD is the RMSD between the coordinates of A and A′. The last term penalizes large RMSDs and favors complexes where A and A′ overlap nicely.

DockTrina was tested on a large set of over 200 symmetric and asymmetric complexes including seven unbound ones (i.e. in these cases the monomers came from experimental structures other than of the sought trimer). In more than a half cases, DockTrina was able to reconstruct a complex with all selected pairwise docking poses having RMSD < 3 Å from the native conformation. Importantly, for the unbound cases, the tool produced near-native complexes (all pairwise docking poses with RMSD < 3 Å) in two cases and acceptable complexes (all pairwise docking poses with RMSD < 10 Å) in four cases. For two more cases, the underlying pairwise docking tool was not able to produce near-native poses. The overall quality of trimer reconstruction of DockTrina is thus remarkably good.

A similar idea can be extended to complexes with the number of monomers greater than three: pairwise docking poses of monomers can be predicted and combined in an optimal way. Unfortunately, the number of combinations of docking poses hereby increases exponentially with the number of monomers. Indeed, for a complex of N subunits and K docking poses per pair, in a naïve approach one needs to assess $K^{\frac{N(N-1)}{2}}$ candidate complexes. Hence, heuristic and optimal algorithms are necessary.

The overall workflow for all combinatorial docking algorithms of this class is similar and consists of three steps (Figure 6.1): (a) generation of pairwise docking poses; (b) combinatorial complex assembly and (c) final scoring of the candidates.

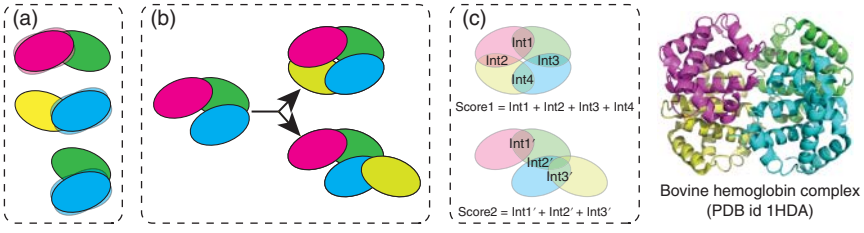


Figure 6.1 Workflow of combinatorial docking algorithms. (a) Pairwise docking poses are generated and scored. (b) Full complexes are assembled in a step-wise fashion. (c) Assembled complexes are scored based on the pairwise docking scores of their subunits.

Examples of the methods that implement this idea are CombDock [36], Multi-LZerD [37], DockStar [38], and 3D-MOSAIC [39].

In CombDock and Multi-LZerD, the sought complex is represented as a spanning tree in a graph where vertices represent subunits and edges possible docking poses between them (for K docking poses, there are K edges between a pair of vertices). Edges are weighted proportionally to the docking scores of individual poses, and the optimal assembly corresponds to the heaviest spanning tree. In an exhaustive search, $N^{N-2}K^{N-1}$ spanning trees have to be assessed (the number of spanning trees in a complete graph with no parallel edges is N^{N-2} [76]). CombDock reduced this number by hierarchical construction of the spanning tree with greedy selection of subtrees. Multi-LZerD employs a genetic algorithm for the search. CombDock scores the resulting complexes based on the shape compemetarity of the produced inter-subunit interfaces and the buried nonpolar surface area. Multi-LZerD implements a physics-based score that is a linear combination of Van der Waals, electrostatics, hydrogen and disulfide bond, solvation, and knowledge-based atom contact terms.

DockStar uses another graph-based representation: a single monomer is chosen as an anchor subunit, and all other monomers are assigned transformations into the anchor frame of reference using the docking poses. The graph is N-partite (N is the number of subunits in the complex), where each set of vertices corresponds to a subunit, and for each transformation, there is a vertex in the corresponding vertex set, and each pair of vertices is joined by an edge with a weight proportional to the corresponding docking score. Then a candidate complex corresponds to choosing one vertex per set, and the optimization problem can be formulated as an integer linear programming (ILP) problem. DockStar efficiently reconstructs star-shaped complexes, and the resulting score is the sum of the pairwise docking scores for the individual docked pairs along the edges in the star-shaped graph. Arbitrary complexes are split into a set of overlapping star-shaped sub-complexes.

3D-MOSAIC implements a greedy assembly strategy, where monomers are added sequentially to an arbitrarily chosen start subunit, and the score is collected on the go as the sum of the scores of the docking poses already included in the complex. In addition to that (and in contrast to DockStar), if adding a subunit induces additional interfaces with other previously added subunits that are similar to a pre-computed docking pose, the score of that docking pose is added to the score of the complex. At each iteration step, high-scoring candidates are clustered, and a user-defined number of clusters is retained for further iterations. Symmetric complexes are further optimized using symmetry operations. It must be emphasized that in order to constrain the solution space, certain knowledge of interaction interfaces between the monomers was assumed in the main application scenario, in that RosettaDock's [77] local docking mode was used that produced candidate poses close to the experimentally observed dimers.

3D-MOSAIC has been extensively tested using a dataset of 308 complexes, both bound and unbound. The predictions are evaluated using tRMSD ("topology RMSD") – a measure introduced by the authors to assess the correctness of complex topology rather than deviations of individual residues. It is similar to iRMSD

("interface RMSD", [69]), a measure introduced earlier to assess the similarity between protein–protein interaction interfaces. In this measure, each subunit is represented by six dots equally distant from its center of mass plus the center of mass itself, and RMSD is measured between the coordinates of these points after an optimal superimposition of each pair of interacting monomers. For 267 of 308 complexes, there was a combination of parameters for 3D-MOSAIC that yielded a tRMSD <2.5 Å, which represents a very close to reality complex conformation.

3D-MOSAIC was compared to CombDock and ClusPro for the complexes where the respective assemblies were available (for ClusPro, 17 cases, in 12 cases also reconstructed with 3D-MOSAIC) or could be reconstructed by running the tool (CombDock, 190 cases). To make the comparison with CombDock fair, the authors performed global docking refraining from using additional information on the location of interaction interfaces. Under a relaxed threshold of tRMSD < 5 Å, 2 complexes were reconstructed correctly with CombDock vs. 19 with 3D-MOSAIC, with a general trend of 3D-MOSAIC to produce lower tRMSD values. Still, 3D-MOSAIC could reconstruct none of the complexes with tRMSD under the strict threshold of 2.5 Å, while CombDock could reconstruct one.

6.3.2 Homology-Based Complex Reconstruction

As we have seen in the previous section, multi-protein assembly reconstruction tools are severely limited by the quality of the structure of dimer complexes, on which they operate. Any additional data on interaction interfaces can thus alleviate this problem and increase the quality of the final assembly. In this section, we will consider a group of methods that, similarly to template-based modeling of protein 3D structure, leverage the information from homologs to this end. The idea behind them is that proteins with similar sequences and structures tend to interact similarly and build similar dimers [69], and hence it is possible to use pairwise complexes of homologs as a starting point for reconstruction of multi-protein assemblies.

A large number of tools employ sequence and structural homology for reconstruction of pairwise complexes, that is regular protein–protein docking ([78–80], see also Chapter 4), but only few were developed further for reconstruction of multi-protein assemblies [40]. One such tool is based on one of these pairwise homology-based methods, PRISM (Protein Interactions by Structural Matching, [79, 81, 82]), which detects experimentally resolved dimers that contain regions near the interaction interface that are similar to certain regions in the target pair of proteins. Similarity, in this case, is represented as a combined sequence- and structure-based similarity score, computed individually for each candidate interaction partner. If some monomers have homologs in different conformations, they are all retained.

For predicting multi-protein assemblies using PRISM [40], all pairwise complexes are first detected, and thus are homologous to some experimentally resolved dimers. Interfaces for all interacting pairs are extracted and clustered according to their structural similarity, their interaction energy is measured with FiberDock [83], and only low-scoring (biologically favorable) pairs are retained. Importantly, unlike in case of the tool described in Section 6.3.1, not all pairs of monomers here

are considered as potentially interacting, but only those that have experimental support from homologous complexes. In the second step of the algorithm, dimers are assembled together into higher-order complexes based on the superimposition of the shared subunits. One protein is added at each iteration, and nonredundant biologically favorable sub-assemblies are filtered. All combinations are explored, and partial assemblies are discarded if clashes occur or no protein can be added with a sufficiently low interaction energy. Finally, candidate assemblies are clustered based on RMSD of the backbone non-hydrogen atoms. The method was tested on a number of protein complexes, including eight that can be assembled from unbound components, with the RMSD of the predicted complex < 6 Å from the experimentally resolved structure.

6.3.3 Prediction from Sequence

Similar to protein tertiary structure, protein quaternary structure can also be predicted directly from sequence. As we have seen in the previous section, homology modeling can be helpful to this end but is naturally limited by the availability of relevant structural templates. Just as in methods for protein structure prediction, one can attempt to predict 3D structure of complexes even in the absence of such information, but the algorithms for this task become increasingly complex, and the solutions decreasingly reliable. The problem considered here is fundamentally different from the one described in the previous sections, since here neither structures of pairwise complexes nor of individual monomers are available.

In predicting 3D structure of individual proteins, when no homologous template is available, one resorts to template-free methods, and TASSER [84] is one of the very successful methods in this field. It employs threading, assembly, and clustering techniques to reconstruct protein 3D structures in the absence of templates with significant sequence similarity. M-TASSER [41] leverages this tool, along with another threading method PROSPECTOR_3 [85] to reconstruct 3D structures of multi-protein assemblies.

In M-TASSER, first 3D models of all chains in the assembly are built using the two threading methods. In this process template monomers (probably without significant sequence similarity) are identified, and multimeric complexes are identified, in case these monomers are a part of a multi-protein assembly. From each such complex, dimers that include the template monomers are extracted, and models of dimers from the sought complex are built and refined. This method does not attempt to build complexes of higher order than dimers but potentially can be combined with combinatorial methods discussed in Section 6.3.1 or with assisted docking methods from Section 6.3.4.

Another way to approach this problem is to try to predict the multimeric state of a protein without predicting its actual 3D structure. An early attempt in this direction involved classifying protein sequences into ones that form homodimers and ones that do not based on amino acid properties using a decision tree [86]. The rationale here is that protein–protein interaction interfaces have different physicochemical properties than the rest of the protein surface, and this can be detected by a

machine-learning classifier trained on an appropriate dataset (cf. Section 6.2.2 for machine-learning tools involving known protein three-dimensional structure). Another study that employed SVMs [87] showed superior performance compared to the decision trees. Homodimers considered in these methods can be of course a part of higher-order structures, but the algorithms make no attempt to predict the actual stoichiometry.

6.3.4 Assisted Docking

As we have seen above (Section 6.3.1), information about approximate location of interaction interfaces can dramatically improve the quality of complex reconstruction. For 3D-MOSAIC, for example, including information on one single interacting residue pair (along with up to ten nonnative contacts to mimic experimental errors) between every two subunits allowed to reconstruct 7 out of 10 test complexes [39]. This indicates that even noisy experimental data could be critical for prediction success. Such experimental data could come in form of low-resolution EM and cryoelectron tomography maps, cross-linking, small-angle X-ray scattering (SAXS), or Förster resonance energy transfer (FRET) experiments. Computationally, incorporation of such restraints can be implemented in form of a weighted docking algorithm ([88], cf. Chapter 4), where certain surface residues receive an additional weight to favor docking poses that involve them, or subsequent filtering of the docking poses can be performed. Alternatively, one can fit high-resolution structures of subunits directly into low-resolution EM maps [89–94] by performing more or less exhaustive searches. Other methods first segment the EM volumes [95], and fit individual subunits simultaneously [96].

ATTRACT-EM (Vries and [42]) has been specifically established to assemble large multi-protein complexes with the help of EM maps based on an earlier docking tool ATTRACT [97], an efficient algorithm with simplified side-chain representation, and aims to perform flexible multi-protein docking. Starting structures are generated by random placement of subunits or are an output of another rigid assembly method. The initial assembly is performed by placing the starting structures into the EM maps using a Gaussian overlap model. Then refinement takes place that favors both favorable interactions between the subunits (from the ATTRACT force field) and the agreement with the EM map.

A similar approach has been adopted in the MDFF method [43], where the fit to the EM maps is further improved with all-atom molecular dynamics (MDs) simulations. This tool has been successfully applied for reconstruction of the HIV-1 capsid at atomic resolution [98].

A version of the described above method PRISM, PRISM-EM [44], leverages EM maps combined with information from complexes of homologs for multi-protein assembly prediction. As in PRISM, subunits are added to a growing complex one by one, taking the pairwise poses from homologous complexes and fitting them into the EM map with the Situs software [99] that was previously developed for fitting atomically resolved 3D structures into EM maps. Candidate subcomplexes are filtered based on the interaction energies between the subunits, clashes, and their fit

in the EM map. Situs provides a score of how well the eventual assembly fits into the given EM map that is adopted by PRISM-EM.

PROXIMO [45] leverages a different type of experimental constraints – the ones coming from the radical probe mass spectrometry (RP-MS) experiments. In this method, oxidizing agent is sprayed on the protein surface [100], and thus on a protein complex surface residues shielded from the solvent by other complex subunits can be identified, which can be used as a source of experimental constraints on pairwise interactions between protein residues. PROXIMO uses a widely-used geometric fit method [73] in combination with this information to constrain the complex conformation space. A number of other methods use different labeling techniques coupled with mass spectrometry as a source of similar geometric restraints [46, 47].

Experimental data on pairwise protein–protein interactions from high-throughput screens, such as yeast two-hybrid (Y2H) or tandem-affinity purification coupled with mass spectrometry (TAP-MS) assays, can also be used as additional restraints for reconstruction of multi-protein assemblies [48]. Another source of such restraints are crosslinking experiments [101, 102].

Finally, one of the most powerful and modern docking tools, HADDOCK [103] also has a mode that allows for assembly of multi-protein complexes [49–51]. A feature that ensures HADDOCK's success over the years is an ability to include various constraints in the form of Ambiguous Interaction Restraints (AIRs) that can be seamlessly incorporated into the docking process by differentiating between the active residues (those believed to participate in the interface) and the passive residues (those believed not to, cf. Chapter 4). The multi-protein mode of HADDOCK allows to dock up to six subunits simultaneously using experimental or bioinformatic (e.g. derived from a consensus interface prediction server CPROT [104]) restraints. It can detect and optimize complexes exhibiting cyclic and dihedral symmetries. These efforts can be further extended to create integrated platforms for reconstruction of large macromolecular assemblies leveraging multiple experimental restraints (e.g. [105], see also Chapter 7).

6.4 Conclusion and Outlook

Detailed resolution of multimeric protein complexes presents a challenge to both experimental and computational approaches. Numerous studies have been performed over the years to classify and predict such assemblies. With X-ray crystallography still being the prime method of structural biology, a separate branch of methods emerged to detect biologically relevant assemblies in its experimental results. With the advent of cryo-EM methods, hybrid methods arose and gained importance. It is to be expected that resolution of cryo-EM will get better with time, eventually reaching atomic in the near future, and hence methods for interpreting cryo-EM results, rather than fitting of structures resolved with other methods or modeled computationally will get more attention.

On the computational side, with computing resources getting more easily available, it gets easier to run large docking or molecular dynamics experiments. Deep

learning-based approaches also enter the field. Computational tools leveraging the immensely successful AlphaFold pipeline [106] are also emerging for protein complex reconstruction, but so far, in case of complexes with more than two proteomers, are largely limited to small homomeric assemblies, such as trimers and tetramers [107].

Acknowledgments

I am grateful to Dr. Vasily Ramensky, Ilya Senatorov, and Liubov Shilova for their critical reading of the draft and stimulating discussions and to the Klaus Faber Foundation for financial support.

References

1 Dunn, M.F., Niks, D., Ngo, H. et al. (2008). Tryptophan synthase: the workings of a channeling nanomachine. *Trends Biochem. Sci.* 33: 254–264. https://doi.org/10.1016/j.tibs.2008.04.008.
2 Hurley, J.H., Dean, A.M., Koshland, D.E., and Stroud, R.M. (1991). Catalytic mechanism of NADP(+)-dependent isocitrate dehydrogenase: implications from the structures of magnesium-isocitrate and $NADP^+$ complexes. *Biochemistry* 30: 8671–8678. https://doi.org/10.1021/bi00099a026.
3 Stefan, M.I. and Novère, N.L. (2013). Cooperative binding. *PLoS Comput. Biol.* 9: e1003106. https://doi.org/10.1371/journal.pcbi.1003106.
4 Goh, C.-S., Milburn, D., and Gerstein, M. (2004). Conformational changes associated with protein–protein interactions. *Curr. Opin. Struct. Biol.* 14: 104–109. https://doi.org/10.1016/j.sbi.2004.01.005.
5 Finn, R.D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 21: 410–412. https://doi.org/10.1093/bioinformatics/bti011.
6 Kuang, X., Dhroso, A., Han, J.G. et al. (2016, 2016). DOMMINO 2.0: integrating structurally resolved protein-, RNA-, and DNA-mediated macromolecular interactions. *Database* bav 114. https://doi.org/10.1093/database/bav114.
7 Kuang, X., Han, J.G., Zhao, N. et al. (2012). DOMMINO: a database of macromolecular interactions. *Nucleic Acids Res.* 40: D501–D506. https://doi.org/10.1093/nar/gkr1128.
8 Davis, F.P. and Sali, A. (2005). PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* 21: 1901–1907. https://doi.org/10.1093/bioinformatics/bti277.
9 Winter, C., Henschel, A., Kim, W.K., and Schroeder, M. (2006). SCOPPI: a structural classification of protein–protein interfaces. *Nucleic Acids Res.* 34: D310–D314. https://doi.org/10.1093/nar/gkj099.
10 Jefferson, E.R., Walsh, T.P., Roberts, T.J., and Barton, G.J. (2007). SNAPPI-DB: a database and API of structures, iNterfaces and alignments for protein–protein

interactions. *Nucleic Acids Res.* 35: D580–D589. https://doi.org/10.1093/nar/gkl836.

11 Yellaboina, S., Tasneem, A., Zaykin, D.V. et al. (2011). DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res.* 39: D730–D735. https://doi.org/10.1093/nar/gkq1229.

12 Meyer, M.J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 29: 1577–1579. https://doi.org/10.1093/bioinformatics/btt181.

13 Xu, Q., Canutescu, A.A., Wang, G. et al. (2008). Statistical analysis of interface similarity in crystals of homologous proteins. *J. Mol. Biol.* 381: 487–507. https://doi.org/10.1016/j.jmb.2008.06.002.

14 Xu, Q. and Dunbrack, R.L. (2020). ProtCID: a data resource for structural information on protein interactions. *Nat. Commun.* 11: 711. https://doi.org/10.1038/s41467-020-14301-4.

15 Xu, Q. and Dunbrack, R.L. Jr., (2011). The protein common interface database (ProtCID)—a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic Acids Res.* 39: D761–D770. https://doi.org/10.1093/nar/gkq1059.

16 Levy, E.D. (2007). PiQSi: protein quaternary structure investigation. *Structure* 15: 1364–1367. https://doi.org/10.1016/j.str.2007.09.019.

17 Levy, E.D., Pereira-Leal, J.B., Chothia, C., and Teichmann, S.A. (2006). 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* 2: e155. https://doi.org/10.1371/journal.pcbi.0020155.

18 Henrick, K. and Thornton, J.M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* 23: 358–361. https://doi.org/10.1016/S0968-0004(98)01253-5.

19 Ponstingl, H., Kabir, T., and Thornton, J.M. (2003). Automatic inference of protein quaternary structure from crystals. *J. Appl. Crystallogr.* 36: 1116–1122. https://doi.org/10.1107/S0021889803012421.

20 Krissinel, E. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* 372: 774–797. https://doi.org/10.1016/j.jmb.2007.05.022.

21 Valdar, W.S.J. and Thornton, J.M. (2001). Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.* 313: 399–416. https://doi.org/10.1006/jmbi.2001.5034.

22 Zhu, H., Domingues, F.S., Sommer, I., and Lengauer, T. (2006). NOXclass: prediction of protein-protein interaction types. *BMC Bioinf.* 7: 27. https://doi.org/10.1186/1471-2105-7-27.

23 Bernauer, J., Bahadur, R.P., Rodier, F. et al. (2008). DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein–protein interactions. *Bioinformatics* 24: 652–658. https://doi.org/10.1093/bioinformatics/btn022.

24 Mitra, P. and Pal, D. (2011). Combining bayes classification and point group symmetry under Boolean framework for enhanced protein quaternary structure inference. *Structure* 19: 304–312. https://doi.org/10.1016/j.str.2011.01.009.

25 Da Silva, F., Desaphy, J., Bret, G., and Rognan, D. (2015). IChemPIC: a random forest classifier of biological and crystallographic protein–protein interfaces. *J. Chem. Inf. Model.* 55: 2005–2014. https://doi.org/10.1021/acs.jcim.5b00190.

26 Luo, J., Guo, Y., Fu, Y. et al. (2014). Effective discrimination between biologically relevant contacts and crystal packing contacts using new determinants. *Proteins Struct. Funct. Bioinf.* 82: 3090–3100. https://doi.org/10.1002/prot.24670.

27 Bliven, S., Lafita, A., Parker, A. et al. (2018). Automated evaluation of quaternary structures from protein crystals. *PLoS Comput. Biol.* 14: e1006104. https://doi.org/10.1371/journal.pcbi.1006104.

28 Duarte, J.M., Srebniak, A., Schärer, M.A., and Capitani, G. (2012). Protein interface classification by evolutionary analysis. *BMC Bioinf.* 13: 334. https://doi.org/10.1186/1471-2105-13-334.

29 Dey, S., Ritchie, D.W., and Levy, E.D. (2018). PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nat. Methods* 15: 67–72. https://doi.org/10.1038/nmeth.4510.

30 Berchanski, A. and Eisenstein, M. (2003). Construction of molecular assemblies via docking: modeling of tetramers with D2 symmetry. *Proteins Struct. Funct. Bioinf.* 53: 817–829. https://doi.org/10.1002/prot.10480.

31 Berchanski, A., Segal, D., and Eisenstein, M. (2005). Modeling oligomers with Cn or Dn symmetry: application to CAPRI target 10. *Proteins Struct. Funct. Bioinf.* 60: 202–206. https://doi.org/10.1002/prot.20558.

32 Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H.J. (2005). Geometry-based flexible and symmetric protein docking. *Proteins Struct. Funct. Bioinf.* 60: 224–231. https://doi.org/10.1002/prot.20562.

33 Pierce, B., Tong, W., and Weng, Z. (2005). M-ZDOCK: a grid-based approach for C_n symmetric multimer docking. *Bioinformatics* 21: 1472–1478. https://doi.org/10.1093/bioinformatics/bti229.

34 Comeau, S.R. and Camacho, C.J. (2005). Predicting oligomeric assemblies: *N*-mers a primer. *J. Struct. Biol.* 150: 233–244. https://doi.org/10.1016/j.jsb.2005.03.006.

35 Popov, P., Ritchie, D.W., and Grudinin, S. (2014). Dock trina: docking triangular protein trimers. *Proteins Struct. Funct. Bioinf.* 82: 34–44. https://doi.org/10.1002/prot.24344.

36 Inbar, Y., Benyamini, H., Nussinov, R., and Wolfson, H.J. (2005). Prediction of multimolecular assemblies by multiple docking. *J. Mol. Biol.* 349: 435–447. https://doi.org/10.1016/j.jmb.2005.03.039.

37 Esquivel-Rodríguez, J., Yang, Y.D., and Kihara, D. (2012). Multi-LZerD: multiple protein docking for asymmetric complexes. *Proteins Struct. Funct. Bioinf.* 80: 1818–1833. https://doi.org/10.1002/prot.24079.

38 Amir, N., Cohen, D., and Wolfson, H.J. (2015). DockStar: a novel ILP-based integrative method for structural modeling of multimolecular protein complexes. *Bioinformatics* 31: 2801–2807. https://doi.org/10.1093/bioinformatics/btv270.

39 Dietzen, M., Kalinina, O.V., Taškova, K. et al. (2015). Large oligomeric complex structures can be computationally assembled by efficiently combining docked

interfaces. *Proteins Struct. Funct. Bioinf.* 83: 1887–1899. https://doi.org/10.1002/prot.24873.

40 Kuzu, G., Keskin, O., Nussinov, R., and Gursoy, A. (2014). Modeling protein assemblies in the proteome. *Mol. Cell. Proteomics* 13: 887–896. https://doi.org/10.1074/mcp.M113.031294.

41 Chen, H. and Skolnick, J. (2008). M-TASSER: an algorithm for protein quaternary structure prediction. *Biophys. J.* 94: 918–928. https://doi.org/10.1529/biophysj.107.114280.

42 de Vries, S.J. and Zacharias, M. (2012). ATTRACT-EM: a new method for the computational assembly of large molecular machines using cryo-EM maps. *PLoS One* 7: e49733. https://doi.org/10.1371/journal.pone.0049733.

43 Trabuco, L.G., Villa, E., Mitra, K. et al. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16: 673–683. https://doi.org/10.1016/j.str.2008.03.005.

44 Kuzu, G., Keskin, O., Nussinov, R., and Gursoy, A. (2016). *PRISM-EM*: template interface-based modelling of multi-protein complexes guided by cryo-electron microscopy density maps. *Acta Crystallogr., Sect. D: Struct. Biol.* 72: 1137–1148. https://doi.org/10.1107/S2059798316013541.

45 Gerega, S.K. and Downard, K.M. (2006). PROXIMO—a new docking algorithm to model protein complexes using data from radical probe mass spectrometry (RP-MS). *Bioinformatics* 22: 1702–1709. https://doi.org/10.1093/bioinformatics/btl178.

46 Kiselar, J.G., Datt, M., Chance, M.R., and Weiss, M.A. (2011). Structural analysis of proinsulin hexamer assembly by hydroxyl radical footprinting and computational modeling. *J. Biol. Chem.* 286: 43710–43716. https://doi.org/10.1074/jbc.M111.297853.

47 Huang, W., Ravikumar, K.M., Parisien, M., and Yang, S. (2016). Theoretical modeling of multiprotein complexes by iSPOT: Integration of small-angle X-ray scattering, hydroxyl radical footprinting, and computational docking. *J. Struct. Biol.* 196: 340–349. https://doi.org/10.1016/j.jsb.2016.08.001.

48 Lasker, K., Phillips, J.L., Russel, D. et al. (2010). Integrative structure modeling of macromolecular assemblies from proteomics data. *Mol. Cell. Proteomics* 9: 1689–1702. https://doi.org/10.1074/mcp.R110.000067.

49 Karaca, E., Melquiond, A.S.J., de Vries, S.J. et al. (2010). Building macromolecular assemblies by information-driven docking. *Mol. Cell. Proteomics* 9: 1784–1794. https://doi.org/10.1074/mcp.M000051-MCP201.

50 de Vries, S.J., van Dijk, A.D.J., Krzeminski, M. et al. (2007). HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins Struct. Funct. Bioinf.* 69: 726–733. https://doi.org/10.1002/prot.21723.

51 de Vries, S.J., van Dijk, M., and Bonvin, A.M.J.J. (2010). The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* 5: 883–897. https://doi.org/10.1038/nprot.2010.32.

52 Marsh, J.A. and Teichmann, S.A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.* 84: 551–575. https://doi.org/10.1146/annurev-biochem-060614-034142.

53 Mistry, J., Chuguransky, S., Williams, L. et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49: D412–D419. https://doi.org/10.1093/nar/gkaa913.

54 Andreeva, A., Howorth, D., Chandonia, J.-M. et al. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 36: D419–D425. https://doi.org/10.1093/nar/gkm993.

55 Wilson, D., Madera, M., Vogel, C. et al. (2007). The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.* 35: D308–D313. https://doi.org/10.1093/nar/gkl910.

56 Stein, A., Russell, R.B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.* 33: D413–D417. https://doi.org/10.1093/nar/gki037.

57 Orengo, C., Michie, A., Jones, S. et al. (1997). CATH – a hierarchic classification of protein domain structures. *Structure* 5: 1093–1109. https://doi.org/10.1016/S0969-2126(97)00260-8.

58 André, I., Strauss, C.E.M., Kaplan, D.B. et al. (2008). Emergence of symmetry in homooligomeric biological assemblies. *Proc. Natl. Acad. Sci.* 105: 16148–16152. https://doi.org/10.1073/pnas.0807576105.

59 Pereira-Leal, J.B., Levy, E.D., Kamp, C., and Teichmann, S.A. (2007). Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* 8: R51. https://doi.org/10.1186/gb-2007-8-4-r51.

60 Plaxco, K.W. and Gross, M. (2009). Protein complexes: the evolution of symmetry. *Curr. Biol.* 19: R25–R26. https://doi.org/10.1016/j.cub.2008.11.004.

61 Kalisman, N., Adams, C.M., and Levitt, M. (2012). Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. *Proc. Natl. Acad. Sci.* 109: 2884–2889. https://doi.org/10.1073/pnas.1119472109.

62 Ahnert, S.E., Marsh, J.A., Hernández, H. et al. (2015). Principles of assembly reveal a periodic table of protein complexes. *Science* 350 (6266), aaa2245.

63 Levy, E.D., Erba, E.B., Robinson, C.V., and Teichmann, S.A. (2008). Assembly reflects evolution of protein complexes. *Nature* 453: 1262–1265. https://doi.org/10.1038/nature06942.

64 Marsh, J.A., Hernández, H., Hall, Z. et al. (2013). Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* 153: 461–470. https://doi.org/10.1016/j.cell.2013.02.044.

65 Marsh, J.A. and Teichmann, S.A. (2014). Parallel dynamics and evolution: protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays* 36: 209–218. https://doi.org/10.1002/bies.201300134.

66 Janin, J. (1997). Specific versus non-specific contacts in protein crystals. *Nat. Struct. Biol.* 4: 973–974. https://doi.org/10.1038/nsb1297-973.

67 Jones, S. and Thornton, J.M. (1995). Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* 63: 31–65. https://doi.org/10.1016/0079-6107(94)00008-W.

68 Eisenberg, D. and McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature* 319: 199–203. https://doi.org/10.1038/319199a0.

69 Aloy, P., Ceulemans, H., Stark, A., and Russell, R.B. (2003). The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.* 332: 989–998.

70 Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinf.* 57: 702–710. https://doi.org/10.1002/prot.20264.

71 Bertoni, M., Kiefer, F., Biasini, M. et al. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.* 7: 10480. https://doi.org/10.1038/s41598-017-09654-8.

72 Biasini, M., Bienert, S., Waterhouse, A. et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42: W252–W258. https://doi.org/10.1093/nar/gku340.

73 Katchalski-Katzir, E., Shariv, I., Eisenstein, M. et al. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci.* 89: 2195–2199. https://doi.org/10.1073/pnas.89.6.2195.

74 Duhovny, D., Nussinov, R., and Wolfson, H.J. (2002). Efficient unbound docking of rigid molecules. In: *Algorithms in Bioinformatics*, Lecture Notes in Computer Science (ed. R. Guigó and D. Gusfield), 185–200. Berlin, Heidelberg: Springer https://doi.org/10.1007/3-540-45784-4_14.

75 Ritchie, D.W., Kozakov, D., and Vajda, S. (2008). Accelerating and focusing protein–protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics* 24: 1865–1873. https://doi.org/10.1093/bioinformatics/btn334.

76 Cayley, A. (1889). A theorem on trees. *Q. J. Math.* 23: 376–378.

77 Chaudhury, S., Berrondo, M., Weitzner, B.D. et al. (2011). Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One* 6: e22477. https://doi.org/10.1371/journal.pone.0022477.

78 Launay, G. and Simonson, T. (2008). Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. *BMC Bioinf.* 9: 427. https://doi.org/10.1186/1471-2105-9-427.

79 Ogmen, U., Keskin, O., Aytuna, A.S. et al. (2005). PRISM: protein interactions by structural matching. *Nucleic Acids Res.* 33: W331–W336. https://doi.org/10.1093/nar/gki585.

80 Zhang, Q.C., Petrey, D., Garzón, J.I. et al. (2013). PrePPI: a structure-informed database of protein–protein interactions. *Nucleic Acids Res.* 41: D828–D833. https://doi.org/10.1093/nar/gks1231.

81 Aytuna, A.S., Gursoy, A., and Keskin, O. (2005). Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* 21: 2850–2855. https://doi.org/10.1093/bioinformatics/bti443.

82 Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat. Protoc.* 6: 1341–1354. https://doi.org/10.1038/nprot.2011.367.

83 Mashiach, E., Nussinov, R., and Wolfson, H.J. (2010). FiberDock: flexible induced-fit backbone refinement in molecular docking. *Proteins Struct. Funct. Bioinf.* 78: 1503–1519. https://doi.org/10.1002/prot.22668.

84 Zhang, Y., Arakaki, A.K., and Skolnick, J. (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins Struct. Funct. Bioinf.* 61: 91–98. https://doi.org/10.1002/prot.20724.

85 Skolnick, J., Kihara, D., and Zhang, Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins Struct. Funct. Bioinf.* 56: 502–518. https://doi.org/10.1002/prot.20106.

86 Garian, R. (2001). Prediction of quaternary structure from primary structure. *Bioinformatics* 17: 551–556. https://doi.org/10.1093/bioinformatics/17.6.551.

87 Zhang, S.-W., Pan, Q., Zhang, H.-C. et al. (2003). Classification of protein quaternary structure with support vector machine. *Bioinformatics* 19: 2390–2396. https://doi.org/10.1093/bioinformatics/btg331.

88 Ben-Zeev, E. and Eisenstein, M. (2003). Weighted geometric docking: Incorporating external information in the rotation-translation scan. *Proteins Struct. Funct. Bioinf.* 52: 24–27. https://doi.org/10.1002/prot.10391.

89 Baker, M.L., Abeysinghe, S.S., Schuh, S. et al. (2011). Modeling protein structure at near atomic resolutions with Gorgon. *J. Struct. Biol.* 174: 360–373. https://doi.org/10.1016/j.jsb.2011.01.015.

90 Ceulemans, H. and Russell, R.B. (2004). Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization. *J. Mol. Biol.* 338: 783–793. https://doi.org/10.1016/j.jmb.2004.02.066.

91 Garzón, J.I., Kovacs, J., Abagyan, R., and Chacón, P. (2007). ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics* 23: 427–433. https://doi.org/10.1093/bioinformatics/btl625.

92 Roseman, A.M. (2000). Docking structures of domains into maps from cryo-electron microscopy using local correlation. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* 56: 1332–1340. https://doi.org/10.1107/S0907444900010908.

93 Rossmann, M.G., Bernal, R., and Pletnev, S.V. (2001). Combining electron microscopic with X-ray crystallographic structures. *J. Struct. Biol.* 136: 190–200. https://doi.org/10.1006/jsbi.2002.4435.

94 Siebert, X. and Navaza, J. (2009). *UROX* 2.0: an interactive tool for fitting atomic models into electron-microscopy reconstructions. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* 65: 651–658. https://doi.org/10.1107/S0907444909008671.

95 Tjioe, E., Lasker, K., Webb, B. et al. (2011). MultiFit: a web server for fitting multiple protein structures into their electron microscopy density map. *Nucleic Acids Res.* 39: W167–W170. https://doi.org/10.1093/nar/gkr490.

96 Kawabata, T. (2008). Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a Gaussian mixture model. *Biophys. J.* 95: 4643–4658. https://doi.org/10.1529/biophysj.108.137125.

97 Zacharias, M. (2003). Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* 12: 1271–1282. https://doi.org/10.1110/ps.0239303.

98 Zhao, G., Perilla, J.R., Yufenyuy, E.L. et al. (2013). Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 497: 643–646. https://doi.org/10.1038/nature12162.

99 Wriggers, W., Milligan, R.A., and McCammon, J.A. (1999). Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* 125: 185–195. https://doi.org/10.1006/jsbi.1998.4080.

100 McKenzie-Coe, A., Montes, N.S., and Jones, L.M. (2021). Hydroxyl radical protein footprinting: a mass spectrometry-based structural method for studying the higher order structure of proteins. *Chem. Rev.* https://doi.org/10.1021/acs.chemrev.1c00432.

101 Förster, F., Lasker, K., Beck, F. et al. (2009). An atomic model AAA-ATPase/20S core particle sub-complex of the 26S proteasome. *Biochem. Biophys. Res. Commun.* 388: 228–233. https://doi.org/10.1016/j.bbrc.2009.07.145.

102 Maiolica, A., Cittaro, D., Borsotti, D. et al. (2007). Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol. Cell. Proteomics* 6: 2200–2211. https://doi.org/10.1074/mcp.M700274-MCP200.

103 Dominguez, C., Boelens, R., and Bonvin, A.M.J.J. (2003). HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* 125: 1731–1737. https://doi.org/10.1021/ja026939x.

104 de Vries, S.J. and Bonvin, A.M.J.J. (2011). CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One* 6: e17695. https://doi.org/10.1371/journal.pone.0017695.

105 Russel, D., Lasker, K., Webb, B. et al. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* 10: e1001244. https://doi.org/10.1371/journal.pbio.1001244.

106 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589. https://doi.org/10.1038/s41586-021-03819-2.

107 Akdel, M., Pires, D.E.V., Porta Pardo, E. et al. (2021). A structural biology community assessment of AlphaFold 2 applications. *bioRxiv* https://doi.org/10.1101/2021.09.26.461876.

7

Live-Cell Structural Biology to Solve Molecular Mechanisms: Structural Dynamics in the Exocyst Function

Altair C. Hernandez[1], Baldo Oliva[1], Damien P. Devos[2], and Oriol Gallego[1]

[1] *Universitat Pompeu Fabra (UPF), Department of Medicine and Life Sciences (MELIS), Carrer del Dr. Aiguader, 88, Barcelona 08003, Spain*
[2] *Universidad Pablo de Olavide-CSIC, Centro Andaluz de Biología del Desarrollo (CABD), Ctra. de Utrera, Km 1, Sevilla 41013, Spain*

7.1 Introduction

The cell machinery is organized in interconnected proteins that assemble into protein complexes and functional networks. Structural information about these protein assemblies in their native environment is crucial to understand the mechanisms that regulate cellular processes. Mapping molecular assemblies into highly detailed models (at atomistic scale) will provide a meaningful description of the functional mechanisms that orchestrate the cell biology. However, this challenge cannot be addressed by any of the available techniques due to their intrinsic limitations (Figure 7.1). In the last decades, *in vitro* approaches have dominated most of structural biology. X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy have long been the standard methods used to obtain numerous protein structures at atomic level, shedding light for the first time on unresolved questions in molecular and cell biology. More recently, cryo-electron microscopy (cryo-EM) has shown its potential to provide protein structures at high resolution without the need for obtaining crystals, which has sped up the process considerably. Despite great advances in resolving protein structures, these methods require the isolation of the sample from its native environment (*in vitro*), and cannot assess functionality under physiological conditions. Within the cell, proteins adopt various conformations to establish transient interactions with other biological molecules that are necessary for their function. Although *in vitro* techniques can resolve structures at atomic level, they are biased toward conformations that can be homogeneously isolated, thus underestimating the subpopulations of transient dynamic states and limiting the mechanistic insight that can be derived.

Establishing a relation between structural features determined in isolation and the functional activity of proteins in a native context is not trivial. To complement the gap between resolution and functionality, *in situ* approaches are on the rise to provide high-resolution measurements in a near-to-physiological context.

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

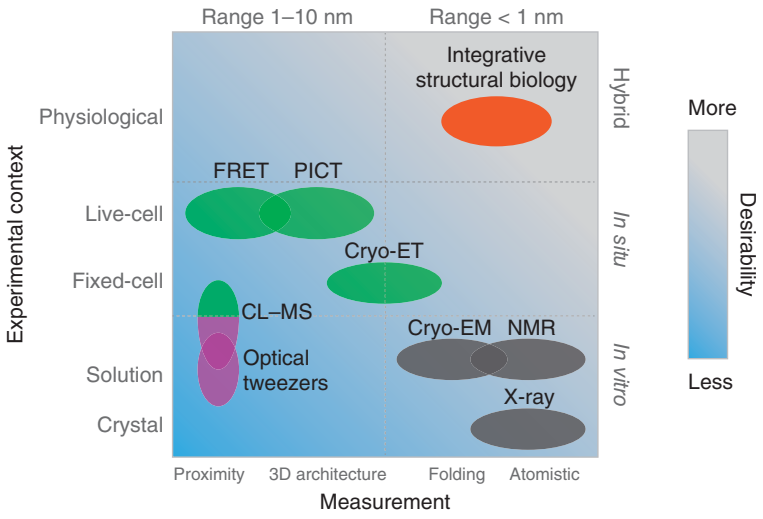


Figure 7.1 Techniques used in structural biology. Overview of the main methods used to gather structural information of biomolecules. Only *in vitro* methods (gray) can achieve atomistic resolution. *In situ* approaches (green) and observations in solution (purple) can provide data in a functional context. Integrative structural biology (orange) can combine information from both *in vitro* and *in situ* methods to achieve physiological data with a better resolution.

In situ structural biology provides structural data of proteins or macromolecular complexes without extracting them from their cellular environment (the term *in cellulo* structural biology is also used).

In situ approaches can resolve heterogeneous populations of conformations and they can explore the interplay with other components of the cell, such as the subcellular localization, interaction with other biomolecules, and dynamics of assembly. However, resolving protein complexes in their native environment is challenging.

Different *in situ* approaches have been developed to study the cell proteome such as crosslinking mass spectrometry (CL–MS) and cryo-electron tomography (cryo-ET). CL–MS uses chemical reagents to label peptides in close proximity. This method can define the interface of short-lived protein–protein interactions and conformational dynamics occurring in the cell. Cryo-ET applies EM to obtain 3D reconstructions (tomograms) of cryopreserved biological samples. These reconstructions are calculated from a set of 2D projection images of the same object, separated by defined tilt angles. Multiple cryo-ET subtomograms capturing several copies of the studied protein complex can be aligned and averaged to increase the resolution of resolved structures. Unfortunately, inherent technical limitations (e.g. chemical reactions and cryo-preservation) continue preventing CL–MS and cryo-ET to visualize the dynamics of the cell machinery *in vivo*.

7.2 Structural Biology Using Light Microscopy Methods

Light-based microscopy involves a broad family of techniques that have been successfully used to study the structure of molecular assemblies. Although the diffraction limit prevents light microscopy from spatially resolving two molecules separated by less than ~300 nm, different strategies have pushed the limits of these methods to a resolutive power that is relevant for structural biology. For instance, optical tweezers offer the possibility of measuring conformational changes at the single-molecule level by the application of forces in the order of piconewton scale [1]. Fluorescence microscopy techniques take advantage of fluorescent tags to identify and track proteins directly in the cellular milieu. Interactions between molecules can be studied through Förster resonance energy transfer (FRET), which is able to measure distances between spectrally different fluorophores in the range between 2 and 10 nm and a temporal resolution of a few milliseconds [2, 3]. Therefore, FRET is able to resolve structural dynamics between predefined labeled elements of the assembly but it is not suitable to generally resolve 3D architectures *de novo*. Localization microscopy identifies the centroid position of spectrally different fluorophores in the image and estimates their separation. This method can measure distances with a precision of 1 nm and essentially no upper limit in the separation between fluorophores [4]. Localization microscopy allowed to map the organization of the kinetochore subunits during metaphase [5]. The so-called super-resolution microscopy methods have pushed the resolving power of light microscopy to resolve fluorophores as close as 2 nm [6]. Localization microscopy and super-resolution microscopy in general have succeeded to resolve large cellular complexes with a 10–20 nm resolution [7, 8]. However, the relatively long time needed for the acquisition of images and the inherent dynamics of the cellular machinery prevented these approaches from being generally used to study protein structures in living cells.

Live-cell fluorescent microscopy presents a unique ability to analyze the structure of biomolecules *in vivo*. PICT (Protein interactions from Imaging of Complexes after Translocation) is a live-cell imaging method that was originally developed to detect and quantitatively characterize protein interactions by combining both live-cell imaging and cell engineering. PICT is based on chemically induced translocation of a protein complex of interest to a static intracellular anchor site [9]. It uses the heterodimerization of FK506-binding protein (FKBP) and FKBP-rapamycin binding domain (FRB) induced by the drug rapamycin for the recruitment of a FRB-tagged bait protein (bait-FRB) to the platform defined by the RFP-FKBP-tagged anchoring protein (anchor-RFP-FKBP) [10]. If the bait-FRB interacts with the GFP-tagged prey (prey-GFP), dual-color fluorescent microscopy will detect an increase in the colocalization of GFP and RFP signals upon rapamycin addition. PICT can be integrated in sophisticated workflows to infer additional structural information of large protein complexes.

In a recent work, a method based on PICT and localization microscopy was developed to determine *de novo* the 3D architecture of protein assemblies [11]. The dynamics in the cellular milieu challenge the structural characterization of cellular complexes *in vivo*. To circumvent this, the PICT method employs engineered yeast cells that harbor immobile anchoring platforms where the studied protein complex

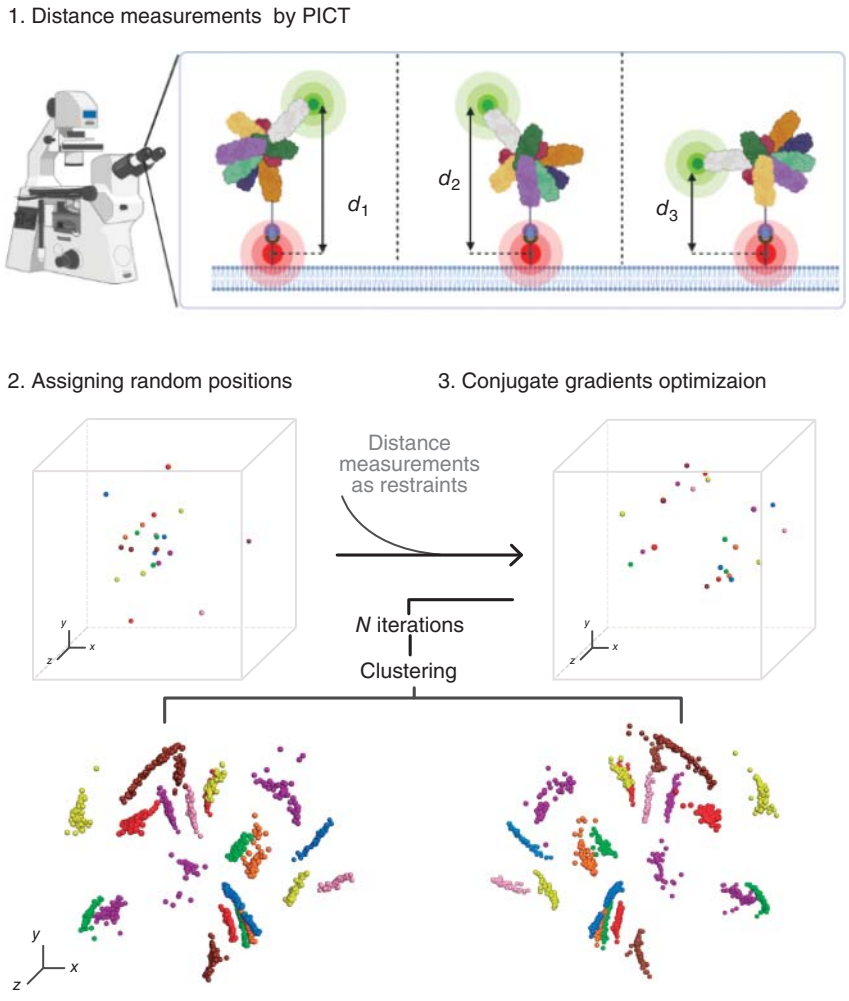


Figure 7.2 Positioning fluorescent tags using PICT. (1) The PICT method allows measuring distances between fluorophores flagging the anchor site (RFP, red) and the termini of the prey-GFP (green) from different orientations using fluorescence microscopy (created with BioRender.com). (2) Tags are represented as spheres and randomly positioned in space. (3) The conjugate gradients algorithm is used to trilaterate tag positions using the set of distance measurements as restraints. This sampling step is iterated to explore the whole space of solutions, each time starting from a random initial configuration. (4) The best scoring solutions that fulfilled all restraints are filtered and superimposed to analyze the ensemble of refined models. Solutions cluster in two populations of solutions that are the mirror image of each other (bottom).

is recruited. Upon recruitment, PICT is capable of measuring the distance between the RFP-labelled anchoring platforms and a subunit fused to GFP (prey-GFP) with a precision of up to 2 nm by localization microscopy (Figure 7.2, Step 1) [9].

Depending on its position within the assembly, each of the bait-FRB used fixes the complex in a specific orientation with respect to the anchoring platform (Figure 7.2). The anchor-RFP-FKBP, which forms reproducible platforms inside the cell, provides a spatial reference across samples. This reference and the controlled anchoring orientation enables integrating the distance measurements to estimate the 3D organization of the complex subunits (tagged to GFP) by trilateration. Trilateration uses distances to a set of reference points to determine the location of an object. The conjugate gradient optimization allows scouting all the possible locations of the fluorophores (prey-GFPs) that are compatible with the distances to the anchor-RFP-FKBP measured upon induced recruitment (Figure 7.2, Step 2).

While sampling, each tested model is scored according to the degree of match with the input data. This allows filtering out the representative subpopulation of best-scoring solutions compatible with all imposed restraints for a posterior analysis and validation [11]. In comparison to X-ray crystallography, NMR, or cryo-EM, the technical requirements and cost to perform PICT are few and widely available. However, the method is restricted to those protein assemblies that can be efficiently tagged and immobilized to the anchoring platform. In addition, although it is able to study the spatial organization of complexes and their relationship with other structures in living cells, light microscopy can only detect and characterize fluorescent labels. Because fluorescent tags are normally attached to the amino-termini (N-) and carboxy-termini (C-) of proteins, light microscopy cannot directly inform about other structural features such as secondary structure or folding of the studied proteins. For instance, PICT can reconstruct an accurate 3D map of the fluorescent tags that label the N- and C- termini of the anchored protein complex (Figure 7.2, Step 2), but it cannot reconstruct the 3D structure of the assembly. In addition, trilateration of the fluorescent tags results in two possible spatial configurations that are mirror images of each other and equally in agreement with the distance measurements. Measurement of distances cannot resolve this ambiguity (Figure 7.2).

Since each method has its own limitations, hybrid approaches that combine *in vitro* and *in situ* structural information are becoming necessary to cross current technical frontiers and to push further knowledge in cell biology.

7.3 Hybrid Methods: Integrative Structural Biology

Computational methods open up the opportunity to further analyze and integrate larger amounts of data generated by life sciences. Integrative structural biology combines data from different experiments to build a more informative depiction of molecular assemblies. By using different sources of information one can represent and describe the assembly with more accuracy and precision than with models based on the individual experiments. The integrative structure determination is a powerful approach for exploring and modeling molecular structures based on

theoretical methods and experimental data, with implications for our understanding of cell biology [12].

To comprehend the integrative structure determination workflow, it is necessary to understand what "model" and "modeling" means. Model is a representation of a real-world assembly that is more informative than the input information on which it is based, which allows extracting detailed features and making testable predictions about future experimental observations. Consequently, modeling is a process of converting some input information into a depiction (model) and its uncertainty [12]. Hence, modeling can be understood as an optimization problem that tries to maximize the accuracy and precision of the model while maintaining the complexity cost of calculation. However, no modeling process can resolve an entire assembly without some degree of uncertainty. Therefore, when modeling it is important to include in the model the propagation of the uncertainty of the input data to measure accuracy. This has been achieved robustly by exploring the conformational space of solutions (models must be consistent with the input information) and not only the model that fits the best with the input data.

The Integrative Modeling Platform (IMP) [13] is an open-source software developed to model the structure of molecular assemblies, by integrating data from diverse biochemical and biophysical experiments. The IMP software package facilitates the scripting of integrative modeling applications, offering a set of tools to develop new model representations, scoring functions, sampling schemes, and analysis methods. Briefly, IMP divides the modeling workflow in four steps: (i) gathering of data; (ii) model representation; (iii) finding models that satisfy the constraints derived from the input data (sampling); (iv) analyses of resulting models and validation. More information about how to model using IMP can be found at the IMP website (http://integrativemodeling.org/).

Several methods based on docking have been also developed over the past few years to study protein complexes. By performing computational docking experiments it is possible to predict or to model the 3D architecture of a biomolecular complex, starting from the structures of the individual molecules in their free, unbound form [14, 15] (see also Chapters 4 and 6). For example, a recent work using molecular docking modeled antibody–antigen complexes by using information from complementary-determining regions and binding epitopes [16]. This workflow was applied to a dataset of 16 complexes and benchmarked its performance by comparing four different docking software suites (ClusPro, LightDock, ZDOCK, and HADDOCK). Other docking-based works have shown its potential to model membrane-associated protein assemblies [17] or to combine several bioinformatic approaches of homology modeling and docking to model the structures of protein complexes identified in protein interaction networks [18, 19].

Different integrative modeling strategies have been implemented when single experiments were not enough to explain the assembly of study. For instance, the molecular architecture of the 26S proteasome was determined by combining data from cryo-EM, X-ray crystallography, and proteomics data about its subunit composition and comparative protein structure models of the component proteins [20]. Another example where the value of integrative modeling is illustrated was the

reconstruction of the topology and structure of the yeast nuclear pore complex (NPC), a large assembly that includes more than 30 types of different proteins [21–23] and more than 450 individual proteins in total. To determine the NPC structure, data from multiple experiments were combined, including stoichiometry from protein quantification, protein proximities from subcomplex purification and CL–MS, protein positions from immuno-EM, and the overall NPC shape from cryo-ET. These models provided fundamental new insights into the function of the NPC controlling the entry and exit from the nucleus of macromolecules, and also shed light on its evolution [22, 24].

Section 7.4 is dedicated to a practical case in which the integrative modeling protocol was applied to reconstruct *de novo* the 3D architecture of the exocyst complex in living cells [11].

7.4 Integrative Modeling: The Case of the Exocyst Complex

Exocytosis is a vesicle trafficking pathway that delivers cargo to the plasma membrane and the extracellular space (Figure 7.3). This cellular pathway is highly conserved across all eukaryotes and it is essential for cell survival, mediating fundamental processes such as cell growth, cell migration, neural development, and tumor invasion [25, 26]. The exocyst is a hetero-octameric protein complex that mediates the tethering of post-Golgi secretory vesicles to the plasma membrane during exocytosis [27–29]. The exocyst consists of eight conserved subunits: Sec3, Sec5, Sec6, Sec8, Sec10, Sec15, Exo70, and Exo84. This protein complex promotes the assembly of the exocytic soluble *N*-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) proteins complex, which in turn induces membrane fusion and the pore formation for the cargo to be released [30, 31]. Mutations of

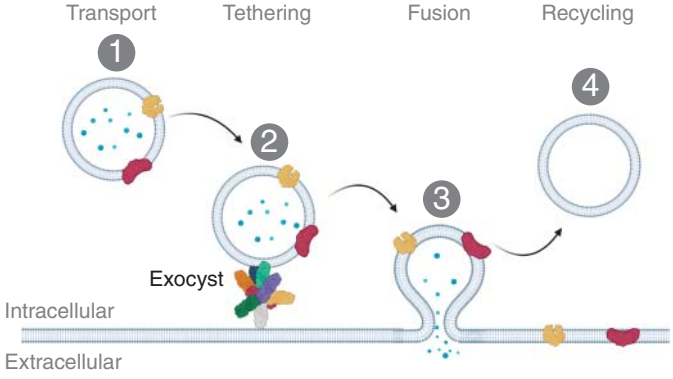


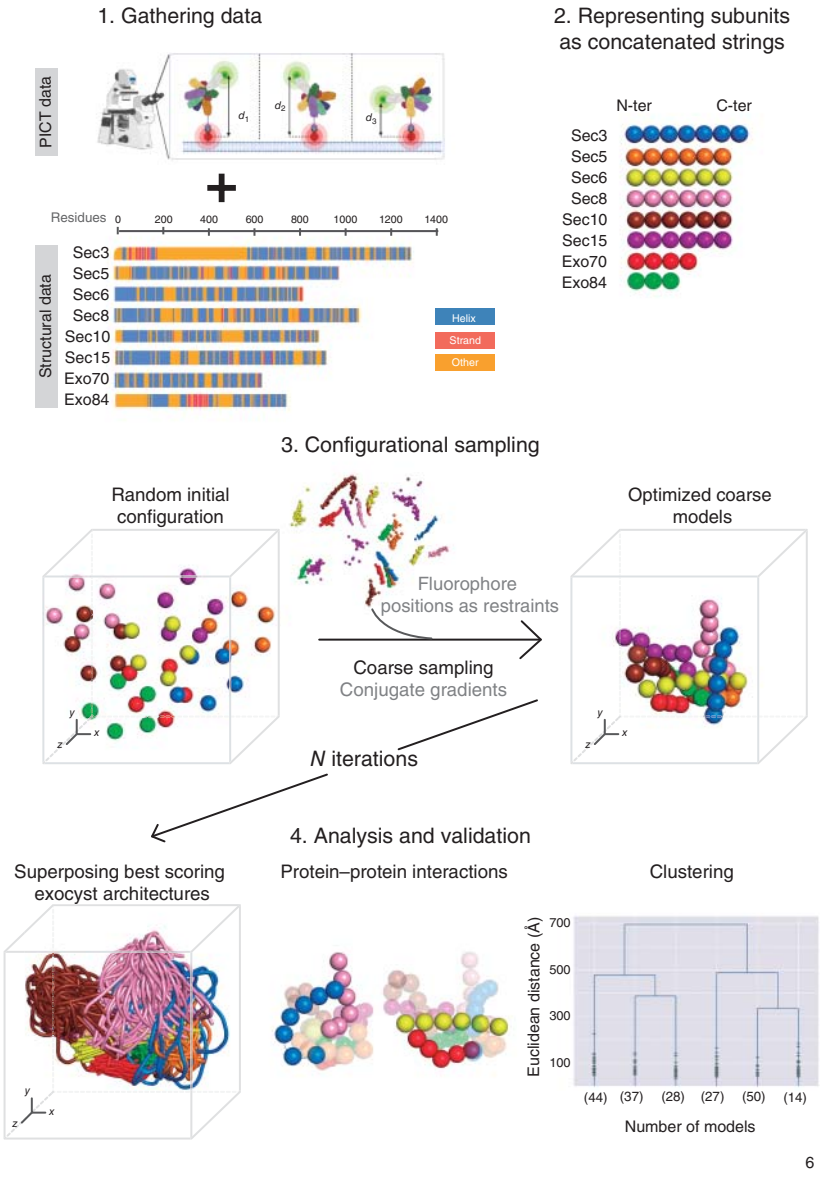
Figure 7.3 Main steps of exocytosis. Representation of the main steps of exocytosis: (1) vesicle transport from the Golgi apparatus to the cell membrane; (2) vesicle tethering to the membrane by the exocyst complex; (3) fusion pore formation and cargo delivery; (4) vesicle recycling (created with BioRender.com).

exocyst subunits result in the accumulation of exocytic vesicles in the cell, and its malfunction has been associated with a number of human pathologies including cancer, Joubert Syndrome, and Polycystic Kidney Disease [32–34]. Despite the importance of the exocyst, the lack of structural information has prevented addressing the molecular mechanisms of exocyst function and regulation.

In a recent study, PICT and structural modeling were combined to map, in living cells, the functional 3D architecture of the exocyst complex bound to a vesicle [11]. In each PICT measurement, one of the exocyst subunits was tagged at its N- or C-terminus with GFP. Therefore, N- and C-termini of the exocyst subunits were imaged with respect to the static anchoring sites, also tagged with RFP, subsequent to recruiting the ensemble in all possible orientations (i.e. using a different bait-FRB) (Figure 7.2, Step 1 and Figure 7.4, Step 1). Then, using localization microscopy, the distances between the anchor-RFP-FKBP and the GFP tags were estimated with a precision below 5 nm. Taking advantage of the IMP conjugate gradient optimization algorithm, distance measurements obtained by PICT allowed positioning of the fluorescent tags in the 3D space by trilateration (GFP fused to the exocyst subunits; RFP fused to the anchoring platform) (Figure 7.2). The selected solutions were clustered according to their similarity, and as expected, they converged in two populations of solutions that were the mirror images of each other and that fulfilled the distance measurement restraints (Figure 7.2, Step 3). Consequently, only one of the two cluster solutions was arbitrarily selected for further progress in the modeling pipeline.

As explained earlier, PICT distance measurements alone can only describe the relative position of the fluorescent tags that have been imaged. The architecture of the assembly was modeled by integrating additional structural information for each subunit (Figure 7.4, Step 1). At this point, as atomic structures were not available for most of the exocyst subunits, each subunit was represented as a flexible string of beads (Figure 7.4, Step 2). Structural features of the subunits retrieved

Figure 7.4 Integrative modeling workflow to compute the 3D architecture of the exocyst. (1) Fluorescent tag positions derived from PICT data were used to position the termini of the exocyst subunits they were fused to. Structural information of exocyst subunits was used for the model representation (made with https://predictprotein.org/). (2) Each subunit was represented as a flexible string of beads, where the length and spacing between beads varied according to the structural features derived from each subunit sequence. Source: Adapted from Picco et al. [11]. (3) Initially each bead was randomly positioned in space (left). Each subunit was imposed to be a concatenated string of beads, whose termini localized according to the positions of the fluorescent tags as derived from the PICT data. Clashes between beads were penalized by using excluded volume restraint. The positions of the beads within each subunit were optimized several times using conjugate gradients, resulting in an optimized model of the exocyst architecture (right). This process was iterated $N = 50{,}000$ times to cover all the space of possible solutions that were in agreement with the input restraints. (4) Solutions for the exocyst architecture were filtered to identify those models with best IMP score and that fulfilled all the input restraints. Exocyst subunits are represented as bundles to help visualize the resulting ensemble of solutions (left). Analysis and validation were performed by assessing biochemical data (such as protein–protein interactions between the C-termini of Exo70-Sec6 and Sec3-Sec8) (center). Structural comparison of models was performed by hierarchical clustering (right).



from deposited X-ray structures, homology modeling, and secondary structure predictions, were considered to define the number of beads representing each subunit (Figure 7.4, Steps 1 and 2). To determine the 3D architecture of the exocyst bound to the vesicle, tag positions were used as scaffold to locate the termini of each subunit in the 3D space (Figure 7.4, Step 3). Again, the modeling workflow was iterated to exhaustively explore the possible space of conformations for the subunits. After filtering best scoring models, it was possible to build the ensemble of best solutions

and reconstruct the 3D architecture of the assembly (Figure 7.4). The *in situ* 3D architecture of the exocyst complex provided relevant insights into the tethering of secretory vesicles. When it binds a secretory vesicle, the exocyst subunits adopt an extended conformation that allows them to interact with one of the termini at the core of the complex and to bind the vesicle with the other termini of Sec10 and Sec15. Given the position of those exocyst subunits that bind molecular landmarks (lipids and proteins) in the plasma membrane (i.e. Sec3 and Exo70), the model suggested that the vesicle establishes direct contact with the plasma membrane while it is sustained by the exocyst. At the same time, this work is a practical example of integrating live-cell imaging with other methods to overcome unresolved questions in cell biology. Unfortunately, data obtained by PICT cannot determine which of the mirror images correspond to the true architecture of the cellular exocyst complex. Additional hybrid approaches combining both *in situ* and *in vitro* are needed to overcome such technical limitations.

7.5 Comparing the *In Situ* Architecture of the Exocyst with a High-Resolution Cryo-EM Model

Consecutive to the exocyst reconstruction using PICT, another study combined cryo-EM, CL–MS, and *in silico* comparative modeling to build a reconstruction of the purified exocyst complex [35]. This approach captured structural features of the exocyst at a near-to-atomic resolution, including folding of the subunits and organization within the complex. For instance, this *in vitro* model showed that eight exocyst subunits fold in helical bundles and are organized in two modules of four subunits each (tetramers), in agreement with previous studies [27, 36]. While it was not possible to assess the functional state of the purified exocyst, the *in vitro* reconstruction allowed us to discern the mirror image ambiguity of the reconstruction in living cells (Figure 7.5A,B). This allowed exploiting the complementarity between the *in vitro* and *in situ* models to gain insight into the molecular mechanism of the exocyst [37].

The location and organization of the two tetrameric modules in the reconstruction done by PICT are in agreement with previous biochemical data and the cryo-EM structure (Figure 7.5B). Nonetheless, superimposition of both the *in situ* and the *in vitro* models showed relevant differences. Although both tetrameric modules superpose well with their counterpart, there is a significant deviation of their relative position. The reconstruction of the isolated exocyst shows a compact "closed" conformation with the two modules packed against each other, while the live-cell architecture of the exocyst bound to a vesicle presents an "open" conformation where module II has rotated about 69° with respect to the "closed" conformation found *in vitro* (Figure 7.5C). These differences in the arrangement of the two modules suggest that conformational dynamics are necessary for the exocyst to become functional. For instance, although Exo70 and Sec10 occupy equivalent locations within the *in vitro* and the *in situ* reconstructions, the two subunits adopt different conformations in each model. This suggests that, in

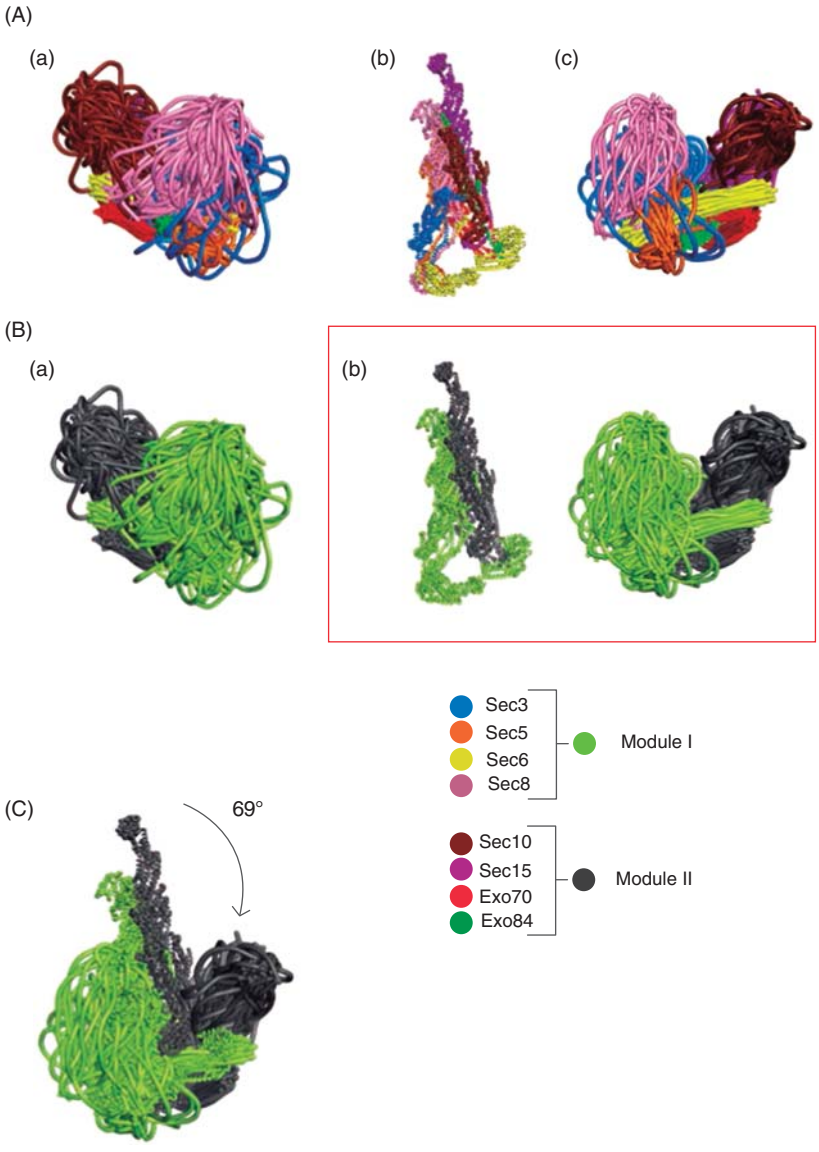


Figure 7.5 Comparative analysis of the exocyst architecture *in situ* and *in vitro*. (A) The two mirror images of the exocyst architecture (a, c) obtained from *in situ* PICT data [11] and the cryo-EM model (b) obtained *in vitro* [35]. (B) Representation of the exocyst complex with module I (green) and module II (gray) (a). The cryo-EM model identified the correct mirror image for the exocyst architecture reconstructed *in situ* (b). (C) Sec6 and Sec8 were used to superimpose *in situ* and *in vitro* models. Module II is rotated by 69° when comparing the two architectures.

addition to the rotation of one module, the tethering of secretory vesicles involves structural dynamics of Exo70 and Sec10. This hypothesis was further supported by a mutagenesis analysis that showed that Exo70 undergoes a conformational change necessary for the exocyst activation [38]. Furthermore, a recent integrative study corroborated fluctuations in some exocyst subunits, such as Sec6 and Exo70, when comparing their density model with the cryo-EM structure, suggesting that these subunits undergo conformational dynamics [39].

The analysis of the exocyst illustrates the complementarity between *in vitro* and *in situ* approaches and the mechanistic insight that can be derived from integrative methods. Together, the near-atomic structure of the isolated exocyst and the *in situ* architecture of the exocyst bound to a vesicle, suggest that the activation of the complex requires structural switches involving the rotation of one tetrameric module with respect to the other one and the rearrangement of Exo70 and Sec10 subunits.

7.6 Discussion and Future Perspectives

Our understanding of the cellular machinery relies on gathering sufficient data that are capable of capturing molecular mechanisms in detail. On the one hand, although *in vitro* techniques have succeeded in revealing the structure of large protein complexes, intrinsic limitations of these methods still prevent atomic-level observations in a near-physiological context. On the other hand, those techniques allowing to study cellular processes under near-to-native conditions (*in vivo* and *in situ* methods) do not provide enough structural detail to decipher the basis of molecular mechanisms (Figure 7.1). Approaches capable of bridging the gap between resolution and biological relevance remain to be discovered. Exploring the structure of protein complexes and their behavior in the cellular context may be achieved by combining different techniques. Integrative approaches have a number of advantages over traditional methods to resolve functional molecular mechanisms with detailed information. In the last decade, significant improvements in cryo-EM have revolutionized the way to obtain molecular structures at remarkable resolution.

Simultaneously, cryo-ET reconstructions and subtomogram averaging have permitted the structural analysis of complex structures in the cell, providing relevant contextual information of the studied proteins. These two techniques can be combined with light-based methods such as fluorescence microscopy and live-cell imaging, which complements high-resolution structures with quantification of the dynamics of cellular events *in situ*. Thus, the development of new hybrid techniques can be a good alternative in the absence of *in vivo* high-resolution data.

Despite integrative methods having the advantage of combining available information about the system of study, obtaining the best depiction is not always trivial. When combining information from different sources, the challenge remains in deciding which data should be included in the modeling process and in which way to avoid misinterpretations in the posterior analysis. Currently, we are placed at an inflection point where computational methods are representing an important piece in biological research. In the past decade, several initiatives have formed to provide user-friendly frameworks to easily perform integrative modeling pipelines, such as

IMP and HADDOCK. Further improvements in computing, optimizing, scoring, validating, visualizing, and dissemination must be achieved.

The accelerated improvement of machine learning-driven alignments, classification, and segmentation methods are opening new avenues to tackle the limitations in structural biology from innovative perspectives. Recent protein structure prediction approaches based on artificial intelligence (AI) have solved some challenges such as the problem of *ab initio* fold prediction. In the last biennial global competition called Critical Assessment of protein Structure Prediction (CASP), the last version of the deep-learning-based approach AlphaFold2 ([40] – DeepMind Technologies) succeeded to determine protein structures with an average error (root mean square deviation [RMSD]) of approximately 1.6 Å. This breakthrough may represent a revolution in the way of obtaining atomic models of protein complexes. However, machine learning approaches rely on learning datasets that are built from static *in vitro* structures, missing the dynamics of biological structures and the flexibility of large complexes. A single methodology is unlikely to time-resolved dynamic and short-lived protein assemblies at atomic resolution. Technical innovations, together with the development of novel hybrid approaches, will be the key to deciphering the molecular bases that orchestrate the cellular machinery in its entire form.

The case of the exocyst is a clear example of how integration of *in situ* and *in vitro* data may provide relevant insights on the structural dynamics that drive cellular processes, but also on the technical gaps that remain to be filled. Integration of *in situ* and *in vitro* structures proved to be efficient in retrieving mechanistic insight such as the conformational switches that mediate the activation of the exocyst. These are fundamental mechanistic details to understand the chain of events that control exocytosis. Nevertheless, the exocyst is a component of the larger protein network that regulates this process. The coordinated action of tens of proteins controls the transport, tethering, fusion, and subsequent delivery of biomolecules to the cell membrane and extracellular space. The structural basis that dictates the interplay between the exocyst and this complex and dynamic exocytic machinery remains to be elucidated. Integrative approaches capable of combining the latest technical advances in different fields of research will become indispensable to shed light on underlying mechanisms that explain cell biology, from exocytosis to all other cellular processes.

Acknowledgements

Oriol Gallego was funded by research grants from the Spanish funding agency (MINECO; ref: PGC2018-095745-B-I00 and EUR2019-103815). Baldo Oliva acknowledges support from MINECO (ref: BIO2017-85329-R). Damien P. Devos was funded by the Spanish Ministry of Economy and Competitiveness (ref: BFU2016-78326-P). Oriol Gallego and Baldo Oliva were supported by the "Unidad de Excelencia María de Maeztu", funded by MINECO (ref: MDM-2014-0370). Altair C. Hernandez was funded by MINECO (ref: PRE2019-088514). We thank Rodrigo Huertas for helping with the design of the figures.

References

1 Choudhary, D., Mossa, A., Jadhav, M., and Cecconi, C. (2019). Bio-molecular applications of recent developments in optical tweezers. *Biomolecules* 9 (1): 23.

2 Hellenkamp, B., Schmid, S., Doroshenko, O. et al. (2018). Precision and accuracy of single-molecule FRET measurements – a multi-laboratory benchmark study. *Nat. Methods* 15 (9): 669–676.

3 Quast, R.B. and Margeat, E. (2021). Single-molecule FRET on its way to structural biology in live cells. *Nat. Methods* 18 (4): 344–345.

4 Churchman, L.S., Okten, Z., Rock, R.S. et al. (2005). Single molecule high-resolution colocalization of Cy3 and Cy5 attached to macromolecules measures intramolecular distances through time. *Proc. Natl. Acad. Sci. U.S.A.* 102 (5): 1419–1423.

5 Wan, X., O'Quinn, R.P., Pierce, H.L. et al. (2009). Protein architecture of the human kinetochore microtubule attachment site. *Cell* 137 (4): 672–684.

6 Balzarotti, F., Eilers, Y., Gwosch, K.C. et al. (2017). Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science* 355 (6325): 606–612.

7 Huang, F., Sirinakis, G., Allgeyer, E.S. et al. (2016). Ultra-high resolution 3D imaging of whole cells. *Cell* 166 (4): 1028–1040.

8 Mund, M., van der Beek, J.A., Deschamps, J. et al. (2018). Systematic nanoscale analysis of endocytosis links efficient vesicle formation to patterned actin nucleation. *Cell* 174 (4): 884–896.

9 Gallego, O., Specht, T., Brach, T. et al. (2013). Detection and characterization of protein interactions in vivo by a simple live-cell imaging method. *PLoS One* 8 (5): e62195.

10 Chen, J., Zheng, X.F., Brown, E.J., and Schreiber, S.L. (1995). Identification of an 11-kDa FKBP12-rapamycin-binding domain within the 289-kDa FKBP12-rapamycin-associated protein and characterization of a critical serine residue. *Proc. Natl. Acad. Sci. U.S.A.* 92 (11): 4947–4951.

11 Picco, A., Irastorza-Azcarate, I., Specht, T. et al. (2017). The in vivo architecture of the exocyst provides structural basis for exocytosis. *Cell* 168 (3): 400–412.

12 Rout, M.P. and Sali, A. (2019). Principles for integrative structural biology studies. *Cell* 177 (6): 1384–1403.

13 Russel, D., Lasker, K., Webb, B. et al. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* 10 (1): e1001244.

14 De Vries, S.J., Van Dijk, M., and Bonvin, A.M.J.J. (2010). The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* 5 (5): 883.

15 Dominguez, C., Boelens, R., and Bonvin, A.M.J.J. (2003). HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* 125 (7): 1731–1737.

16 Ambrosetti, F., Jiménez-García, B., Roel-Touris, J., and Bonvin, A.M.J.J. (2020). Modeling antibody-antigen complexes by information-driven docking. *Structure* 28 (1): 119–129.

17 Roel-Touris, J., Jiménez-García, B., and Bonvin, A.M.J.J. (2020). Integrative modeling of membrane-associated protein assemblies. *Nat. Commun.* 11 (1): 6210.

18 Mirela-Bota, P., Aguirre-Plans, J., Meseguer, A. et al. (2021). Galaxy InteractoMIX: an integrated computational platform for the study of protein–protein interaction data. *J. Mol. Biol.* 433 (11): 166656.

19 Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. *Nat. Methods* 10 (1): 47.

20 Lasker, K., Förster, F., Bohn, S. et al. (2012). Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc. Natl. Acad. Sci. U.S.A.* 109 (5): 1380–1387.

21 Alber, F., Dokudovskaya, S., Veenhoff, L.M. et al. (2007). The molecular architecture of the nuclear pore complex. *Nature* 450 (7170): 695–701.

22 Kim, S.J., Fernandez-Martinez, J., Nudelman, I. et al. (2018). Integrative structure and functional anatomy of a nuclear pore complex. *Nature* 555 (7697): 475–482.

23 Rout, M.P., Aitchison, J.D., Suprapto, A. et al. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* 148 (4): 635–651.

24 Alber, F., Dokudovskaya, S., Veenhoff, L.M. et al. (2007). Determining the architectures of macromolecular assemblies. *Nature* 450 (7170): 683–694.

25 Heider, M.R. and Munson, M. (2012). Exorcising the exocyst complex. *Traffic* 13 (7): 898–907.

26 Martin-Urdiroz, M., Deeks, M.J., Horton, C.G. et al. (2016). The exocyst complex in health and disease. *Front. Cell Dev. Biol.* 4: 24.

27 Heider, M.R., Gu, M., Duffy, C.M. et al. (2016). Subunit connectivity, assembly determinants and architecture of the yeast exocyst complex. *Nat. Struct. Mol. Biol.* 23 (1): 59–66.

28 Hsu, S.C., Ting, A.E., Hazuka, C.D. et al. (1996). The mammalian brain rsec6/8 complex. *Neuron* 17 (6): 1209–1219.

29 TerBush, D.R., Maurice, T., Roth, D., and Novick, P. (1996). The exocyst is a multiprotein complex required for exocytosis in *Saccharomyces cerevisiae*. *EMBO J.* 15 (23): 6483–6494.

30 Wu, B. and Guo, W. (2015). The exocyst at a glance. *J. Cell Sci.* 128 (16): 2957–2964.

31 Yue, P., Zhang, Y., Mei, K. et al. (2017). Sec3 promotes the initial binary t-SNARE complex assembly and membrane fusion. *Nat. Commun.* 8: 14236.

32 Fogelgren, B., Lin, S.-Y., Zuo, X. et al. (2011). The exocyst protein Sec10 interacts with polycystin-2 and knockdown causes PKD-phenotypes. *PLoS Genet.* 7 (4): e1001361.

33 Seixas, C., Choi, S.Y., Polgar, N. et al. (2016). Arl13b and the exocyst interact synergistically in ciliogenesis. *Mol. Biol. Cell* 27 (2): 308–320.

34 Yamamoto, A., Kasamatsu, A., Ishige, S. et al. (2013). Exocyst complex component Sec8: a presumed component in the progression of human oral squamous-cell carcinoma by secretion of matrix metalloproteinases. *J. Cancer Res. Clin. Oncol.* 139 (4): 533–542.

35 Mei, K., Li, Y., Wang, S. et al. (2018). Cryo-EM structure of the exocyst complex. *Nat. Struct. Mol. Biol.* 25: 139–146.

36 Katoh, Y., Nozaki, S., Hartanto, D. et al. (2015). Architectures of multisubunit complexes revealed by a visible immunoprecipitation assay using fluorescent fusion proteins. *J. Cell Sci.* 128 (12): 2351–2362.

37 Irastorza-Azcarate, I., Castaño-Díez, D., Devos, D.P., and Gallego, O. (2019). Live-cell structural biology to solve biological mechanisms: the case of the exocyst. *Structure* 27 (6): 886–892.

38 Rossi, G., Lepore, D., Kenner, L. et al. (2020). Exocyst structural changes associated with activation of tethering downstream of Rho/Cdc42 GTPases. *J. Cell Biol.* 219 (2): e201904161.

39 Ganesan, S.J., Feyder, M.J., Chemmama, I.E. et al. (2020). Integrative structure and function of the yeast exocyst complex. *Protein Sci.* 29 (6): 1486–1501.

40 Senior, A.W., Evans, R., Jumper, J. et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 7792: 706–710.

8

Kinetics and Thermodynamics of Protein–Protein Encounter

Nicolas Künzel and Volkhard Helms

Saarland University, Center for Bioinformatics, Saarland Informatics Campus, Postfach 15 11 50, 66041 Saarbrücken, Germany

8.1 Introduction

This chapter follows up on two excellent ground-breaking review articles. In 1997, Gilson and colleagues laid out the statistical-thermodynamic basis for the computation of biomolecular binding affinities. These fundamental principles apply equally well to complexes where small molecules (e.g. drugs) bind to proteins and larger ensembles when two or more proteins bind to each other [1]. In 2009, Schreiber and colleagues focused their review on protein–protein association and discussed computational as well as experimental techniques and concepts [2]. All that content would be worth repeating here, but this is obviously not possible due to space constraints. Instead, this chapter reviews modern methodological developments and their application to study the energetics and kinetics of protein–protein association and dissociation.

8.2 Thermodynamic Ensembles and Free Energy

This section gives an introduction in the concept of thermodynamic ensembles and their connected state functions, the free energies. The focus is placed on the isothermal–isobaric ensemble in which most biological experiments are performed. Afterwards, the Jarzynski equality is explained that is a special equation allowing to calculate equilibrium free energies from non-equilibrium work values which are later on used when analyzing the performed alchemical simulations.

The field of statistical mechanics links the classical microscopic states of a system to the macroscopic observables which can be measured in experiments. These can be thermodynamic, structural, and dynamical properties. A key concept of statistical mechanics is thermodynamic ensembles [3] that characterize the probability distribution for all possible microscopic states of a system [4]. Ensembles can be defined for any possible set of external constraints. The best-known ones are the microcanonical (fixed number of particles N, volume V, and total energy E),

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

canonical (fixed N, V, and temperature T), the isoenthalpic-isobaric (fixed N, pressure P, and entropy S), the isothermal–isobaric (fixed N, P, T), often called NPT, and the grand canonical (fixed V, T, and chemical potential μ) ensembles. Because most experiments are performed in the isothermal–isobaric ensemble, we will focus on this one in the following.

As a start, we briefly summarize the laws of thermodynamics. The first law states that the internal energy change ΔU

$$\Delta U = Q + W \tag{8.1}$$

results from the work W the surrounding performs on the system and the heat Q that is added to the system during a process. This first law basically describes energy conservation, i. e. energy is only transformed from one form to another but can never be destroyed or created.

The second law of thermodynamics discusses the total entropy S (a state variable), also called disorder, of a system. It states that S cannot decrease in isolated systems, i. e. when there is no exchange of energy or matter with the surrounding. Another wording is that the heat from a warmer body will naturally flow to a colder one. The second law reads

$$\Delta S \geq \frac{Q}{T} \tag{8.2}$$

where T is the temperature. This means that the change in entropy equals the amount of heat Q added to the system with respect to the temperature for reversible processes. For irreversible processes the entropy exceeds it.

8.2.1 The Isothermal–Isobaric Ensemble and the Gibbs Free Energy

One can think of isobaric systems as being coupled to external pistons that enable to compress or expand the volume of the system to keep the internal pressure P constant. Analogously, isothermal systems can be viewed as being coupled to external thermal reservoirs exchanging heat with the system to keep the internal temperature T constant. The respective state function or thermodynamic potential of the isothermal–isobaric ensemble is the *Gibbs free energy* $G(N, P, T)$ that is connected to the particle number N, the pressure P, and the temperature T via

$$\mathrm{d}G = \mu \mathrm{d}N + V \mathrm{d}P - S \mathrm{d}T \tag{8.3}$$

where μ is the chemical potential, V is the volume, and S is the entropy. G is labeled "free" energy because it denotes the amount of energy that is "free" to perform work at a constant temperature T and pressure P [5]. The term was coined by Hermann von Helmholtz [6]. The Gibbs free energy is related to the isothermal–isobaric partition function Δ by

$$G(N, P, T) = -\frac{1}{\beta} \ln \Delta(N, P, T) \tag{8.4}$$

where $\beta = 1/(k_\mathrm{B} T)$ is the inverse temperature with the Boltzmann constant k_B. For a mathematical derivation and further background on ensembles, their partition functions and their interconnection using Legendre transformations, the reader

is referred to [3]. Knowing the Gibbs free energy G and its change ΔG_{AB} between two system states A and B directly enables one to calculate many macroscopic observables of a given isothermal–isobaric system and is thus key to understanding macroscopic systems and processes. These are, for example, the enthalpy $H = -\partial \ln(\Delta(N,P,T))/(\partial\beta)$, the heat capacity $C_P = \partial H/\partial T$, and the chemical potential $\mu = (\partial G/\partial N)_{P,T}$ to name just a few.

The change in free energy (the free energy difference) can tell if energy must be added to the system in order for a reaction to happen or if it occurs spontaneously. The free energy difference tells, for example, if a chemical compound is a promising drug candidate [3] or if two protein domains form a stable interaction. The standard binding free energy difference $\Delta G^0_{\mathrm{bind}}$ is also directly connected to experimental observables, such as the equilibrium dissociation constant K_{D}

$$\Delta G^0_{\mathrm{bind}} = -k_{\mathrm{B}} T \ln(c_0 K_{\mathrm{D}}) \tag{8.5}$$

with the standard concentration $c_0 = 1\ \mathrm{mol\,l^{-1}} \approx 1/1661\ \mathrm{\AA}^{-3}$.

It is possible to express the free energy as a function of reaction/generalized coordinates of the system, often also called collective variables (CVs). Depending on the process one is interested in, these can be, for example, angles, distances, and root mean squared deviations (RMSDs), or also a parameter of the system Hamiltonian, such as Lennard-Jones interactions. These coordinates spanfree energy hypersurfaces, which contain information about stable conformers of the system (with respect to the chosen coordinates) and their relative stability, barriers, and minimum-free-energy paths in-between them. The derivative of the free energy along with a CV is the ensemble-averaged force [7], i.e. the force along with a CV averaged over all configurations of a system, whereby the free energy along with a CV is often called *potential of mean force* (PMF). The average force is the first part of the instantaneous force acting along with the CV, the second part is a random force with zero average. It includes the fluctuations of all additional degrees of freedom and thus enforces the progression of the CV. These dynamics take place along with the time-independent PMF [8]. Generally, one can obtain the free energy along with a reaction coordinate ξ from the probability density function $P(\xi)$ of the CV using [9]

$$G(\xi) = -k_{\mathrm{B}} T \ln P(\xi) \tag{8.6}$$

if the sampling of the phase space is sufficient, i.e. the simulation is sufficiently long and was not trapped inside local free energy minima. However, this does not mean that the resulting free energy is a meaningful quantity. If ξ was chosen poorly then $G(\xi)$ will be meaningless in terms of describing the actual system states. There are two major requirements for CVs in molecular dynamics simulations. The relevant metastable states as well as the transition states between them have to be distinct regions in CV space. Thus, they have to be energetically separate regions in the chosen space. In contrast, if different metastable states are projected onto the same CV space, energy barriers are integrated out and major sampling problems occur when using CV-based sampling methods discussed below [10].

8.3 Overview of Computational Methods to Determine Binding Free Energies

In this section, we will introduce several computational methods to determine binding free energies of protein complexes. These are, often, also applicable to studying the binding of small molecules and peptides to proteins or DNA. If you are interested in a specific discussion of the binding of small molecules to proteins, please refer to Chapter 14 by Michael Hutter. Furthermore, we will concentrate on methods to determine binding free energies for complexes where the binding modes are already known. For methods to determine binding modes by molecular docking please refer to Chapter 4 by Martin Zacharias.

In principle, it is possible to obtain binding free energies using unbiased long molecular dynamics (MD) simulations. Most commonly, however, the time scales of binding and unbinding are too long to be sampled on common simulation hardware. In recent years, specialized hardware was developed that allowed to conduct millisecond long trajectories [11, 12] of solvated protein systems in atomistic detail. Notably, such simulations have been recently applied to study the reversible association and dissociation kinetics of the five protein–protein complexes; barnase–barstar, insulin dimer, ras-raf RBD, RNase HI-SSB-Ct, and TYK2-pseudokinase by atomistic MD simulations in explicit solvent [13]. The authors performed dozens of conventional MD simulations with aggregated simulation times of hundreds of microseconds. In addition, they performed "tempered" binding simulations, whereby "the strength of interactions between the protein monomer atoms, and sometimes between the protein monomer and solvent atoms, [was] scaled at regular time intervals using a simulated Hamiltonian tempering framework" [13]. This scaling was adjusted to allow dissociation from long-lived bound states to occur within hundreds of microseconds rather than days. Observing reversible binding and unbinding events yielded the following general picture. One may have thought that associating proteins could in principle form an encounter complex at an arbitrary interface and proceed from there to the native interface without dissociating by means of an extensive search [13]. Instead, the authors observed that "in successful association events the encounter complexes tended to form rather close to the native interface." On the other hand, encounter complexes that would not later reach the native interface formed in a wide variety of relative orientations. At present, this specialized hardware is too expensive for the normal scientist to work with and other methods have been developed to deal with the sampling problem. In the following, we will give an overview of different simulation methods, so-called enhanced sampling methods, which allow the system to overcome barriers in free energy, thus moving out of local or the global minima, which then enables one to calculate binding free energies *in silico*.

In comparison to association and dissociation of protein–small molecule, protein–peptide, or protein–RNA complexes, the association and dissociation of protein–protein complexes formed from globular monomers often have the advantage that the individual proteins have relatively stiff overall structures. Thus, fewer degrees of freedom have to be sampled to obtain reasonable binding free

energies from simulations. Yet, in recent years, more and more interactions between intrinsically disordered regions of proteins with other proteins have been detected, which leads to a much greater number of degrees of freedom that have to be sampled. A further disadvantage is the size of the system that has to be simulated leading to larger simulation boxes and thus more particles have to be taken into account.

There exist two major forms of binding free energy methods. Methods that are used to calculate absolute binding free energies, i. e. the change in free energy that results when a ligand binds to a receptor, and methods used to calculate relative binding free energies, i. e. the difference in absolute binding free energy between two ligands binding to the same receptor. Even though both are usually called free energies, the former is a free energy difference between the bound and unbound states and the latter is even a difference of free energy differences.

8.3.1 Coarse Graining

One way to speed up simulations is to reduce the number of degrees of freedom of the system. This is the reason behind coarse-grained simulations, where a single particle imitates the physicochemical properties of a group of atoms [14] thus reducing the amount of particles in the system and additionally lowering the bond vibration frequencies enabling to use a larger time step in the simulations. In the popular Martini force-field [15, 16], four non-hydrogen atoms are combined into one particle [17]. Such force fields exist for proteins, biomembranes, nucleic acids and carbohydrates. An important shortcoming of coarse-grained models is that they are insufficient in describing atomic details and can, thus, not explain certain detailed aspects of protein–ligand as well as protein–protein binding even though coarse-grained force fields are getting better in describing these interactions allowing to also study protein–ligand interactions [18]. A common coarse-graining method is called Brownian Dynamics (BD) and will be explained in Section 8.3.1.1.

8.3.1.1 Brownian Dynamics

At large separation distances, the relative motion of two proteins can be described as diffusive motion subject to their relative interaction. The Brownian Dynamics (BD) method is a suitable method to describe the diffusive motion of multiple interacting particles by iteratively propagating the Ermak–McCammon algorithm [19]. Here, the translational Brownian motion of two or more interacting proteins is simulated as the displacement Δr of the particle positions r during a time step Δt according to the relation

$$\Delta r = \frac{D\Delta t}{k_\mathrm{B}T}\mathbf{F} + \mathbf{R}, \text{ with } \langle \mathbf{R} \rangle = 0 \text{ and } \langle \mathbf{R}^2 \rangle = 6D\Delta t \tag{8.7}$$

where $\mathbf{F}$ is the systematic interparticle force, k_B is the Boltzmann constant, T is the temperature, and $\mathbf{R}$ is the stochastic displacement arising from collisions of the proteins with solvent molecules that are not represented explicitly. Analogous formulas

$$\Delta w_i = \frac{D_{iR}\Delta t}{k_\mathrm{B}}\mathbf{T}_{ij} + \mathbf{W}_i, \text{ with } \langle \mathbf{W}_i \rangle = 0 \text{ and } \langle \mathbf{W}_i^2 \rangle = 6D_{iR}\Delta t \tag{8.8}$$

are used to generate the rotational motions of the two proteins in terms of rotation angle $\mathbf{w}_j = (w_{1j}, w_{2j}, w_{3j})$, torque $\mathbf{T}_{ij}$ acting on protein i due to protein j, and rotational diffusion constant D_{iR} of each protein i $(i,j = 1,2, i \neq j)$, where $\mathbf{W}_i$ is again a stochastic term. The BD method was adapted by the McCammon and Wade groups to study the relative motion of one protein around another protein that is kept fixed as a diffusive motion of a rigid body subject to external forces [20, 21]. A popular software package for performing BD simulations is the SDA package by the Wade group [22].

From the observed dynamics, one computes association rates for the diffusive encounter of two particles using

$$k = k(b)\beta_\infty \tag{8.9}$$

with the rate $k(b)$ at which the reacting particles reach a distance b for the first time [23]. This distance is taken large enough so that the *potential of mean force* $U(r)$ between the two particles is only a function of their distance, not of their orientation. In this case, $k(b)$ can be determined straightforwardly from

$$k(b) = 4\pi D\left[\int_b^\infty \frac{dr}{r^2}\exp\left[\frac{U(r)}{k_B T}\right]\right]^{-1} \tag{8.10}$$

The factor β_∞ indicates how many of the particles located on a spherical surface with distance b to the binding interface will actually bind. It can be derived from explicit BD simulations. In principle, the kinetics of protein–protein interaction is determined by their diffusive properties. On the other hand, proteins that carry electrostatically complementary charges or dipoles may reach 100–1000 times higher association rates k_{on} than what is expected from the Smoluchowski equation for a purely diffusive motion. This behavior is termed "electrostatic steering".

BD has been shown to successfully reproduce experimental k_{on} rates for the association of electrostatically complementary protein–protein pairs [24]. In this regard, a successful binding event is typically detected as soon as 2 – 4 out of a given set of critical inter-protein contacts are established [24]. Commonly, BD techniques apply a continuum model for the solvent and explicitly treat intermolecular electrostatic interactions between the diffusing particles. To be able to describe subtle effects resulting from the molecular fine structure of the solvent at close distances from the diffusing particles, the SDA package also implements terms for short-range electrostatic desolvation interactions and for short-ranged hydrophobic desolvation interactions.

Using a large ensemble of BD trajectories, one may characterize the underlying protein–protein interaction free energy landscape by post-analysis of the trajectories. The basic procedure was presented in ref. [25] and then applied to the association of wild-type and mutant barnase–barstar complexes [26]. The method was also implemented into the SDA package. Recently, Ötztürk and Wade have presented BD simulations of the diffusional association of wild-type and mutants of the globular domain of the linker histone H1 from mouse to a nucleosome [27]. From

these simulations, they reported bimolecular association rate constants (k_{on}), the Gibbs binding free energy (ΔG), and the dissociation rate constant (k_{off}) for the formation of a diffusional encounter complex between the nucleosome and the histone. They found that the BD simulations were able to help in predicting the relative effects of single point mutations on fluorescence recovery after photobleaching (FRAP) recovery times related to protein binding.

The BD method can of course also handle more than two proteins. For example, McGuffee and Elock applied BD simulations to study diffusional dynamics of more than 1000 proteins reflecting the most abundant macromolecules of the *Escherichia coli* cytoplasm [28]. By calibrating the BD simulations to reproduce the translational diffusion coefficients of green fluorescent protein (GFP) observed in vivo, the authors used snapshots of the simulation trajectories to compute the cytoplasm's effects on the thermodynamics of protein folding, association, and aggregation events and found that their simulation model successfully described the relative thermodynamic stabilities of proteins measured in *E. coli*. Subsequently, the group of Michael Feig has pushed brute-force all-atom explicit solvent MD simulations to study systems of related size (100 million atoms and more) on timescales of hundreds of nanoseconds [29].

The benefits of BD simulations over atomistic simulations today are the ability to generate superior sampling of relative particle positions and orientations, the ability to extend system sizes to dimensions that are currently not reachable in explicit solvent simulations, and the beauty of a simplistic description that can well describe many interesting biological phenomena.

8.3.2 Endpoint Methods

So-called endpoint methods sample the bound and unbound states of a system and then calculate the binding free energy difference between these two states using approximations of the system energy. The simplest method is the linear response approximation or linear interaction energy (LIE) [30, 31]. It is usually applied to obtain protein–ligand free energies and is not discussed further here. Popular methods for evaluating protein–protein and protein–ligand binding free energies are the molecular mechanics with Poisson–Boltzmann and surface area solvation (MM/PBSA) [32, 33] and the molecular mechanics generalized Born surface area (MM/GBSA) [32, 33] methods, which will be explained in Section 8.3.2.1.

8.3.2.1 MM/PBSA/MM/GBSA

The MM/PBSA [32, 33] and MM/GBSA [32, 33] methods combine molecular mechanics force fields with continuum solvation models to estimate protein–protein and protein–ligand binding affinities. The free energy of a system state is evaluated using [32–34]

$$G = E_{\text{bonded}} + E_{\text{electrostatic}} + E_{\text{vdW}} + G_{\text{polar}} + G_{\text{non-polar}} - TS \tag{8.11}$$

The terms E_{bonded} (bond, angle and dihedral energy), $E_{\text{electrostatic}}$ (electrostatic energy), and E_{vdW} (van der Waals interactions) are the corresponding standard molecular mechanics (MM) energy terms. G_{polar} and $G_{\text{non-polar}}$ form the solvation free energy. The distinction between MM/PBSA and MM/GBSA is that the former one uses the Poisson–Boltzmann (PB) equation and the latter the generalized Born (GB) model to estimate the polar solvation term. The non-polar solvation is estimated using the solvent-accessible surface area (SASA). The last term includes the system temperature T as well as the entropy S. This is analyzed using normal-mode analysis of the systems vibrational frequencies for a standard state to be comparable to experimental values [35]. To obtain binding free energies, it is necessary to calculate the free energies of the complex, the unbound ligand and the unbound receptor based on (8.11) [36]. Commonly only the complex is simulated and the ensemble averages for the free ligand and free receptor are obtained by removing the relevant atoms from the system before the analysis. This leads to

$$\Delta G_{\text{bind}} = \langle G_{\text{receptor+ligand}} - G_{\text{receptor}} - G_{\text{ligand}} \rangle_{\text{receptor+ligand}} \,. \tag{8.12}$$

and to more precise results due to cancellation of intramolecular terms and reduces the required simulation time but ignores the energetic effects of relevant structural changes of the receptor and/or ligand upon binding [34].

Overall, MM/PBSA and MM/GBSA often give better results than LIE, docking, and scoring, but worse ones compared to more advanced techniques, e.g. pathway methods. Depending on the system, reasonably good binding free energy values can be obtained, but for some systems, the methods fail. It has been shown that reasonable results can already be achieved using less than 100 repeated simulations with a length of around 200 ps each [34, 37] and therefore much faster than using pathway methods discussed below. The overall coefficient of determination for the whole PDBbind database was shown to be $r^2 = 0.3$ but the individual results differed strongly, $r^2 = 0.0 - 0.8$ [34, 38]. MM/PBSA and MM/GBSA highly depend on the choice of the dielectric constant for the electrostatic energy and the used force field. Additionally, the binding free energy is calculated as a difference of large values, and therefore, the precision of the result is very low when the standard deviations of the individual terms are high and ligands with similar binding affinities cannot be compared successfully. For a deeper discussion of the MM/PBSA and MM/GBSA methods, their application and possible issues, the reader is referred to [34].

8.3.3 Potential of Mean Force/Pathway Methods

Contrary to the just discussed endpoint methods, free energies of protein–protein binding can also be computed using so-called pathway methods where the free energy is expressed as a function of geometrical reaction coordinates. Such simulations involve a considerably larger computational effort. In most cases, they are more accurate than the former. Most of the following methods are not only used to accelerate binding and unbinding processes but are, often, also used in simulations of protein folding.

Multiple reviews for these methods and their comparison have been published [10, 39–41].

8.3.3.1 Thermodynamic Integration

One of the oldest pathway methods is the so-called thermodynamic integration (TI) [42], which overcomes barriers in free energy by freezing the chosen CV at different values while sampling along all other degrees of freedom at these fixed points along with the reaction coordinate. A free energy profile or PMF is obtained by integrating the mean force, i.e. the derivative of the free energy with respect to the CV [43]. It is possible to slowly move the constraint instead of simulating the CV at fixed values. This is called slow growth [44]. Both methods need comparably long simulation times to reach sufficient sampling along with the degrees of freedom of the system and thus a converged PMF. In these methods, the momentum in the direction of the reaction coordinate is constrained, and thus, they do not fully sample the momentum space.

8.3.3.2 Umbrella Sampling (US)

Umbrella sampling (US) [45] differs from TI by replacing the fixed constraints using restraining biasing potentials, allowing to sample the full momentum space [43]. Here, a series of windows is selected along with the CV of interest so that the window distributions overlap sufficiently. Usually, one uses harmonic potentials of the form

$$w_i(\xi) = \frac{k_i}{2}(\xi - \xi_i^{\text{ref}})^2 \tag{8.13}$$

with center points ξ_i^{ref} and spring constants k_i as biasing restraints for umbrella sampling, but other choices can be imagined. It is also possible to choose an adaptive bias that tries to match the negative of the free energy at each point of the CV ξ. This method is, thus, called adaptive umbrella sampling [46]. It can be extended to periodically interacting with multiple walkers and on-the-fly resampling to sample neglected (undersampled) regions. The method is then called adaptive biasing force (ABF) [47] and will be discussed below. If the bias potential is moved or pulled along with the CV instead of using a finite number of fixed windows, the method is called steered MD (SMD) or force-probe MD [43, 48]. It will be explained in more detail in Section 8.3.3.3.

The most critical part of umbrella sampling is the correct choice of the CVs and the spring constants k_i defining the strength and thus the width of the biasing potentials. They have to be chosen in advance of the simulations. An advantage of umbrella sampling is that the MD simulations of different windows are completely independent of each other so that they can be executed in parallel. It is even possible to later insert additional windows with larger spring constants if the overlap between originally chosen windows is not sufficient.

Results from umbrella sampling simulations are usually combined either by the umbrella integration [43, 49] or by the popular weighted histogram analysis method (WHAM) [50].

For a more detailed description of umbrella sampling and its analysis, the reader is referred to [43].

When it comes to studying the assembly or dissociation of protein complexes, simple umbrella sampling simulations are potentially facing huge sampling problems due to the sheer number of possible relative orientations of the two binding partners, which have to be sampled in every window of the simulation [51]. This is less of a problem if the two proteins each have strongly dipolar character so that they will adopt a preferred orientation relative to each other during association and dissociation processes. The Brownian Dynamics simulations of Spaar et al. discussed above showed that this is the case, for example, for the barnase–barstar complex. When barstar binds to barnase, it approaches barnase "from the right side" due to favorable electrostatic interactions. Also, its binding interface is pre-oriented toward the binding interface like a spaceship that plans to land on the moon. In such cases, it appears plausible to employ a one-directional reaction coordinate to describe protein–protein association and dissociation, e.g. the distance between the proteins' center of masses. The direction of approach can be taken parallel to the vector connecting the COMs in the bound complex assuming that this is known either from structural studies or from docking. When comparing the unbinding of the three complexes barnase–barstar, cytochrome *c* – cytochrome *c* peroxidase, and enzyme 1 – histidine phosphocarrier, it turned out that the PMF computed via umbrella potential simulations had a monotonous uphill profile without transition states [52] (Figure 8.1). In all cases, the two proteins attracted each other up to distances of about 1.4–1.5 nm. Afterward, the PMF curve was flat reflecting that the protein interaction was shielded by the solvent. Beyond such distances, the relative orientation of the two proteins is not relevant anymore. The same calculations were also performed for a dissociation process starting from a nonspecific short-lived

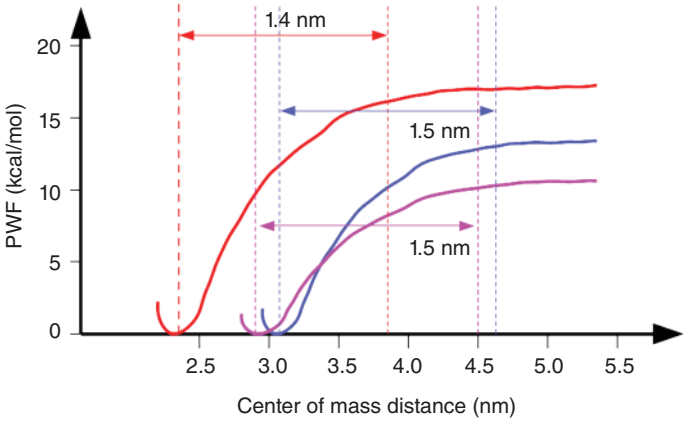


Figure 8.1 Potential of mean force obtained from umbrella sampling simulations, see [52]. Colored in red is the PMF for the formation of the specific Barnase: Barstar complex, in blue that of Cytochrome *c*: Cytochrome *c* peroxidase, and in magenta that for the complex between the N-terminal domain of enzyme I: histidine-containing phosphocarrier. In each case, the left dashed line parallel to the *y*-axis represents the center of mass (COM) distance in the specific complex and the right dashed line indicates the cutoff distance that separates the bound region from the unbound region (right), respectively. Source: Adapted from Ulucan et al. [52].

contact of the same protein pairs [53]. In that case, the PMF profiles had the same shape, but a shorter attraction basin of only about 0.8 nm, and the contact conformation was only about half as stable as the specific complex.

In order to treat sampling issues it is possible to use additional external restraints. Gumbart et al. [51, 54] applied geometrical and conformational restraints that are enforced on the overall relative protein motions and also to constituent amino acids forming the interface region. They showed that the resulting binding free energy of the studied barnase–barstar complex compares well to the experimental value and that the statistical error of the method is low for a system of this complexity. The time the method needs to reach convergence is also much lower than for umbrella sampling without additional restraints. One has to note that the contributions of the external restraints to the overall binding free energy difference were calculated using the adaptive biasing force method. Only the main separation PMF was calculated using umbrella sampling. In a recent study Suh et al. [55] showed that various advanced methods including their newly developed "String Method for Protein–Protein Binding Free-Energy Calculations" can lead to converged results far off from the experimental binding free energy of barnase–barstar. They discuss this in much detail and compare these results to the one obtained by Gumbart et al. [51, 54].

Another recent example of restraint umbrella sampling simulations of protein–protein interactions is found in [56].

8.3.3.3 Steered MD (SMD)

As already described, steered molecular dynamics (steered MD) or force-probe MD utilizes moving bias potentials to push the simulation system over barriers in the free energy. The mean force and thus the PMF can be sufficiently sampled and thus estimated if the movement of the potential is slow compared to the relaxation times of the system [43]. Thus, usually, very slow movements have to be chosen that drastically increase the amount of sampling to converge these simulations. An interesting feature of SMD is its equivalency to atomic-force microscopy [43]. Fast SMD is often used to obtain approximate starting positions for umbrella sampling simulations. If the results of steered MD are evaluated using nonequilibrium analysis methods, such as the Jarzynski equality [57] or the Bennet acceptance ratio (BAR) [58], the movement of the potential can be performed much faster than the relaxation time scales of the system. In order to obtain a suitable ensemble average of the PMF, these simulations have to be performed repeatedly. A few examples for the usage of steered MD to obtain protein–protein binding free energies are given in [59, 60].

8.3.3.4 Metadynamics

Metadynamics [61], like all other related methods, strongly depends on the choice of the used reaction coordinates. An advantage of metadynamics is the fact that it is possible to sample multiple CVs at the same time, which is not as easily possible in other methods. Usually, only two or three CVs are chosen because larger numbers would necessitate much more sampling.

In order to overcome barriers in free energy and to escape minima on the free energy surface, metadynamics uses history-dependent bias potentials, which are often called hills. These are Gaussian functions dependent on the CVs $\boldsymbol{\xi}(\mathbf{x}) = (\xi_1(\mathbf{x}), \cdots, \xi_n(\mathbf{x}))$ which themselves depend on the coordinates $\mathbf{x}$ of the system. These are added to the potential energy V of the study system at a chosen frequency. The overall resulting bias potential is

$$V_G(\boldsymbol{\xi}(\mathbf{x}), t) = \sum_{t'<t} W(t') \exp\left(-\sum_{i=1}^{n} \frac{(\xi_i(\mathbf{x}) - \xi_i(\mathbf{x}_G(t')))^2}{2\sigma_i^2} \right) \tag{8.14}$$

where $t' = \tau_G, 2\tau_G, 3\tau_G, \cdots$ are multiples of the hill-deposition time τ_G, where $\xi(\mathbf{x}_G(t))$ is the trajectory of the system subject to the action of $V + V_G$, where W is the height of the Gaussian potentials, and where the σ_i are the widths of the Gaussian potentials in the respective CV [17, 61, 62]. As is clear from (8.14) the hills are adding up during the course of the simulation and thus fill up the free energy minima over time. Higher and wider hills increase the speed of convergence but reduce the sharpness of details on the free energy surface.

In metadynamics, the free energy surface cannot be calculated from (8.6) because canonical sampling is hindered by the bias potential and thus it has to be calculated differently. After a certain time, the biasing potential has completely filled all minima on the free energy surface and the effective potential becomes flat. At this point, convergence is reached. The free energy surface is then simply the negative of the biasing potential

$$G(\xi(\mathbf{x})) = -V_G(\xi(\mathbf{x})) \, . \tag{8.15}$$

Advancement of the original metadynamics approach is the so-called well-tempered metadynamics [63]. Here, the height of the hills is decreased while the bias potential is accumulated, resulting in high hills at the beginning and lower hills at the end of the metadynamics simulations [17], thus strongly improving convergence speed while keeping the sharpness of details on the free energy surface. Furthermore, the well-tempered metadynamics approach leads asymptotically to an exact free energy surface [64], which is highly advantageous over the original approach.

Various extensions to the metadynamics method have been developed since it was first published. These are, for example, funnel metadynamics [65], volume-based metadynamics [66], multiple walkers metadynamics [67], and flux-tempered metadynamics [68], to name just a few [69]. By using a method called infrequent metadynamics [70, 71], it is possible to also calculate the dynamics of a given system, e.g. the kinetic rate constants [69] k_{on} and k_{off}.

Metadynamics in combination with parallel tempering [72] has been successfully used for protein oligomerization/dissociation studies [73, 74].

There is also a range of other methods utilizing history-dependent potentials. One example is the accelerated weight histogram method [75].

For recommendable tutorials on how to perform metadynamics simulations and to choose collective variables, the reader is referred to articles describing the tool Plumed [76, 77].

8.3.3.5 Adaptive Biasing Force (ABF)

The adaptive biasing force (ABF) method [47] tries to retain the dynamics of the system along with the PMF, including the random force described in Section 8.3, while additionally leveling the PMF to easily move along with the PMF, because barriers in free energy are removed. This leads to an acceleration of the passage between the relevant states along with the CV [8] and thus improves sampling along with the CV. This is achieved by calculating the mean force along with a reaction coordinate ξ and removing it via an external biasing force, which is exactly the negative of the current estimate of the mean force. This results in uniform sampling along ξ [9]. ABF has been successfully used to obtain reasonable protein–protein binding free energies, e.g. in [51] and [78]. For a detailed explanation of the ABF method, a suitable choice of the reaction coordinates, error analysis, and extended methods the reader is referred to [8].

8.3.4 Replica-Exchange Methods

In replica-exchange methods for molecular dynamics [79], the system is simulated in different system states at the same time. At regular time intervals, individual simulations may exchange properties, such as temperature or coordinates at certain steps so that barriers in free energy can be overcome more easily. Subsequently, replica-exchange methods were combined with pathway methods to be used in the calculation of free energies and PMFs. A great advantage of many parallel/replica-exchange methods is the possibility of simulating in parallel on multiple computer nodes because the interconnection is only required for the exchange steps for which one does not need high-speed connections between the individual nodes.

8.3.4.1 Parallel Tempering

Increasing the simulation temperature is an obvious way to more easily overcome the free energy barriers along with the CVs. Arrhenius law tells us that reaction rates increase with temperature because an increased number of particles have an energy greater than the minimum energy needed for the reaction at increased temperature [10]. Usually, simulated tempering methods [80] are performed as follows: First, the system is propagated at a fixed temperature T_i for a number of time steps. Second, the acceptance for switching between two temperatures T_i and T_j is evaluated as a Monte Carlo step with the acceptance probability of

$$\alpha = \min\left(1, \frac{Z_j}{Z_i}\exp\left[-\frac{U(x)}{k_\mathrm{B}T_j} + \frac{U(x)}{K_\mathrm{B}T_i}\right]\right) \tag{8.16}$$

where i is the index of the present temperature and j the temperature of the new one. In general, it is nontrivial to choose the weights Z_i in a suitable way so that all values of i are equivalently sampled [10].

In order to overcome the issue of finding the correct weights, it is possible to simulate multiple replicas of the system at the same time at different temperatures.

Here not the temperature of a single system is changed but rather the coordinates of two replicas are exchanged with the acceptance probability

$$\alpha = \min\left(1, \exp\left[\left(\frac{1}{k_B T_j} - \frac{1}{k_B T_i}\right)\left(U(x_j) - U(x_i)\right)\right]\right) \tag{8.17}$$

which is independent of the weights Z_i and Z_j. Equal sampling of each index i is achieved by only allowing pairwise swapping [10].

One example of a pathway method extended with parallel tempering is the combination of metadynamics with parallel tempering [72], which has been used in refs. [73, 74] to study protein oligomerization/dissociation. For further information on parallel tempering, the reader is referred to refs. [10, 81].

8.3.4.2 Generalized/Hamiltonian Replica-Exchange Methods

In the Section 8.3.4.1, the replicas differed in temperature. It is, however, also possible to change other parameters of the system, such as parts of the Hamiltonian [82], or combine various changes, e.g. replicas can have different temperatures and Hamiltonians at the same time and their acceptance probability then reads [10].

$$\alpha = \min\left(1, \frac{\exp\left[-\left(\frac{U_i(x_j)}{k_B T_i} + \frac{U_j(x_i)}{k_B T_j}\right)\right]}{\exp\left[-\left(\frac{U_i(x_i)}{k_B T_i} + \frac{U_j(x_j)}{k_B T_j}\right)\right]}\right). \tag{8.18}$$

8.3.5 Additional Pathway Methods

The methods explained in Sections 8.3.3 to 8.3.4 are just a brief summary of the most popular computational methods to calculate binding free energies of protein–protein interactions. Many more interesting methods, combinations of the aforementioned ones, as well as combinations with machine learning techniques have been successfully used to study association and dissociation of protein–protein systems. It was also beneficial to combine pathway methods with or derive CVs from experimental data, such as in small angle X-ray scattering (SAXS)-guided metadynamics [83].

8.3.6 Relative Binding Free Energies

In principle, it is also possible to calculate relative binding free energies for protein–protein interactions. Usually, one is interested e.g. in computing the free energy difference between two different charge states of a protein, or between wild-type and an amino acid mutation, etc. Various methods that were specifically established to compute relative binding free energies often used so-called alchemical approaches. An example is found in [84].

References

1 Gilson, M.K., Given, J.A., Bush, B.L. et al. (1997). The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.* 72: 1047–1069.

2 Schreiber, G., Haran, G., and Zhou, H.-X. (2009). Fundamental aspects of protein protein association kinetics. *Chem. Rev.* 109 (3): 839–860.

3 Tuckerman, M. (2010). *Statistical Mechanics: Theory and Molecular Simulation*, *Oxford Graduate Texts*. OUP Oxford. ISBN 9780191523465. https://books.google.de/books?id=Lo3Jqc0pgrcC.

4 Gibbs, J.W. (1902). *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundations of Thermodynamics*. C. Scribner's Sons. https://books.google.de/books?id=IGMSAAAAIAAJ.

5 Baierlein, R. (1999). *Thermal Physics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511840227.

6 Deutsche Akademie der Wissenschaften zu Berlin (1882). *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin*, volume Jan-Mai 1882. Berlin: Deutsche Akademie der Wissenschaften zu Berlin. https://www.biodiversitylibrary.org/item/93362.

7 Berendsen, H.J.C. (2007). *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815348.

8 Comer, J., Gumbart, J.C., Hénin, J. et al. (2015). The adaptive biasing force method: everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B* 119 (3): 1129–1151. https://doi.org/10.1021/jp506633n. PMID: 25247823.

9 Chipot, C. and Pohorille, A. (2007). *Free Energy Calculations - Theory and Applications in Chemistry and Biology*, *Springer Series in Chemical Physics*. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-540-38447-2. http://www.springer.com/de/book/9783540384472.

10 Abrams, C. and Bussi, G. (2014). Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy* 16 (1): 163–199. https://doi.org/10.3390/e16010163.

11 Shaw, D.E., Deneroff, M.M., Dror, R.O. et al. (2007). Anton, a special-purpose machine for molecular dynamics simulation. *SIGARCH Comput. Archit. News* 35 (2): 1–12. https://doi.org/10.1145/1273440.1250664.

12 Shaw, D.E., Grossman, J.P., Bank, J.A. et al. (2014). Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. *SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 41–53. https://doi.org/10.1109/SC.2014.9.

13 Pan, A.C., Jacobson, D., Yatsenko, K. et al. (2019). Atomic-level characterization of protein–protein association. *Proc. Natl. Acad. Sci. U.S.A.* 116 (10): 4244–4249. https://doi.org/10.1073/pnas.1815431116.

14 Tozzini, V. (2005). Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* 15 (2): 144–150. https://doi.org/10.1016/j.sbi.2005.02.005. Theory and simulation/Macromolecular assemblages.

15 Marrink, S.J., Risselada, H.J., Yefimov, S. et al. (2007). The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* 111 (27): 7812–7824. https://doi.org/10.1021/jp071097f. PMID: 17569554.

16 Monticelli, L., Kandasamy, S.K., Periole, X. et al. (2008). The MARTINI coarse grained force field: extension to proteins. *J. Chem. Theory Comput.* 4 (5): 819–834. https://doi.org/10.1021/ct700324x. PMID: 26621095.

17 Spiwok, V., Sucur, Z., and Hosek, P. (2015). Enhanced sampling techniques in biomolecular simulations. *Biotechnol. Adv.* 33 (6, Part 2): 1130–1140. https://doi.org/10.1016/j.biotechadv.2014.11.011. BioTech 2014 and 6th Czech-Swiss Biotechnology Symposium.

18 Souza, P.C.T., Thallmair, S., Conflitti, P. et al. (2020). Protein–ligand binding with the coarse-grained martini model. *Nat. Commun.* 11 (1): 3714. https://doi.org/10.1038/s41467-020-17437-5.

19 Ermak, D.L. and McCammon, J.A. (1978). Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.* 69 (4): 1352–1360. https://doi.org/10.1063/1.436761.

20 Madura, J.D., Briggs, J.M., Wade, R.C. et al. (1995). Electrostatics and diffusion of molecules in solution: simulations with the university of Houston Brownian dynamics program. *Comput. Phys. Commun.* 91 (1): 57–95. https://doi.org/10.1016/0010-4655(95)00043-F.

21 Gabdoulline, R.R. and Wade, R.C. (1998). Brownian dynamics simulation of protein-protein diffusional encounter. *Methods* 14 (3): 329–341. https://doi.org/10.1006/meth.1998.0588.

22 Martinez, M., Bruce, N.J., Romanowska, J. et al. (2015). SDA 7: A modular and parallel implementation of the simulation of diffusional association software. *J. Comput. Chem.* 36 (21): 1631–1645. https://doi.org/10.1002/jcc.23971.

23 Northrup, S.H., Allison, S.A., and McCammon, J.A. (1984). Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *J. Chem. Phys.* 80 (4): 1517–1524. https://doi.org/10.1063/1.446900.

24 Gabdoulline, R.R. and Wade, R.C. (2001). Protein-protein association: investigation of factors influencing association rates by Brownian dynamics simulations 1 edited by B. Honig. *J. Mol. Biol.* 306 (5): 1139–1155. https://doi.org/10.1006/jmbi.2000.4404.

25 Spaar, A. and Helms, V. (2005). Free energy landscape of protein-protein encounter resulting from Brownian Dynamics simulations of Barnase:Barstar. *J. Chem. Theory Comput.* 1 (4): 723–736. https://doi.org/10.1021/ct050036n. PMID: 26641694.

26 Spaar, A., Dammer, C., Gabdoulline, R.R. et al. (2006). Diffusional encounter of barnase and barstar. *Biophys. J.* 90 (6): 1913–1924. https://doi.org/10.1529/biophysj.105.075507.

27 Öztürk, M.A. and Wade, R.C. (2020). Computation of FRAP recovery times for linker histone–chromatin binding on the basis of Brownian dynamics simulations. *Biochim. Biophys. Acta, Gen. Subj.* 1864 (10): 129653. https://doi.org/10.1016/j.bbagen.2020.129653.

28 McGuffee, S.R. and Elcock, A.H. (2010). Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.* 6 (3): 1–18. https://doi.org/10.1371/journal.pcbi.1000694.

29 Yu, I., Mori, T., Ando, T. et al. (2016). Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *eLife* 5: e19274. https://doi.org/10.7554/eLife.19274.

30 Åqvist, J., Medina, C., and Samuelsson, J.-E. (1994). A new method for predicting binding affinity in computer-aided drug design. *Protein Eng. Des. Sel.* 7 (3): 385–391. https://doi.org/10.1093/protein/7.3.385.

31 Hansson, T., Marelius, J., and Åqvist, J. (1998). Ligand binding affinity prediction by linear interaction energy methods. *J. Comput.-Aided Mol. Des.* 12 (1): 27–35. https://doi.org/10.1023/A:1007930623000.

32 Srinivasan, J., Cheatham, T.E., Cieplak, P. et al. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J. Am. Chem. Soc.* 120 (37): 9401–9409. https://doi.org/10.1021/ja981844+.

33 Kollman, P.A., Massova, I., Reyes, C. et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* 33 (12): 889–897. https://doi.org/10.1021/ar000033j. PMID: 11123888.

34 Genheden, S. and Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discovery* 10 (5): 449–461. https://doi.org/10.1517/17460441.2015.1032936.

35 Gohlke, H. and Case, D.A. (2004). Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* 25 (2): 238–250. https://doi.org/10.1002/jcc.10379.

36 Swanson, J.M.J., Henchman, R.H., and McCammon, J.A. (2004). Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.* 86 (1): 67–74. https://doi.org/10.1016/S0006-3495(04)74084-9.

37 Genheden, S. and Ryde, U. (2011). A comparison of different initialization protocols to obtain statistically independent molecular dynamics simulations. *J. Comput. Chem.* 32 (2): 187–195. https://doi.org/10.1002/jcc.21546.

38 Sun, H., Li, Y., Tian, S. et al. (2014). Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. *Phys. Chem. Chem. Phys.* 16: 16719–16729. https://doi.org/10.1039/C4CP01388C.

39 Bernardi, R.C., Melo, M.C.R., and Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta, Gen. Subj.* 1850 (5): 872–877. https://doi.org/10.1016/j.bbagen.2014.10.019; http://www.sciencedirect.com/science/article/pii/S0304416514003559. Recent developments of molecular dynamics.

40 Luitz, M., Bomblies, R., Ostermeir, K., and Zacharias, M. (2015). Exploring biomolecular dynamics and interactions using advanced sampling methods. *J. Phys. Condens. Matter* 27 (32): 323101. https://doi.org/10.1088/0953-8984/27/32/323101.

41 Siebenmorgen, T. and Zacharias, M. (2020). Computational prediction of protein-protein binding affinities. *WIREs Comput. Mol. Sci.* 10 (3): e1448. https://doi.org/10.1002/wcms.1448.

42 Carter, E.A., Ciccotti, G., Hynes, J.T., and Kapral, R. (1989). Constrained reaction coordinate dynamics for the simulation of rare events. *Chem. Phys. Lett.* 156 (5): 472–477. https://doi.org/10.1016/S0009-2614(89)87314-2.

43 Kästner, J. (2011). Umbrella sampling. *WIREs Comput. Mol. Sci.* 1 (6): 932–942. https://doi.org/10.1002/wcms.66.

44 Hermans, J. (1991). Simple analysis of noise and hysteresis in (slow-growth) free energy simulations. *J. Phys. Chem.* 95 (23): 9029–9032. https://doi.org/10.1021/j100176a002.

45 Torrie, G.M. and Valleau, J.P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.* 23 (2): 187–199. https://doi.org/10.1016/0021-9991(77)90121--8.

46 Mezei, M. (1987). Adaptive umbrella sampling: self-consistent determination of the non-Boltzmann bias. *J. Comput. Phys.* 68 (1): 237–248. https://doi.org/10.1016/0021-9991(87)90054-4.

47 Adamson, S., Kharlampidi, D., and Dementiev, A. (2008). Stabilization of resonance states by an asymptotic coulomb potential. *J. Chem. Phys.* 128 (2): 024101. https://doi.org/10.1063/1.2821102.

48 Izrailev, S., Stepaniants, S., Isralewitz, B. et al. (1999). Steered molecular dynamics. In: *Proceedings of the 2nd International Symposium on Algorithms for Macromolecular Modelling* (ed. P. Deuflhard, J. Hermans, B. Leimkuhler et al.), 39–65. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-642-58360-5. https://doi.org/10.1007/978-3-642-58360-5.

49 Kästner, J. and Thiel, W. (2005). Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "umbrella integration". *J. Chem. Phys.* 123 (14): 144104. https://doi.org/10.1063/1.2052648.

50 Kumar, S., Rosenberg, J.M., Bouzida, D. et al. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* 13 (8): 1011–1021. https://doi.org/10.1002/jcc.540130812.

51 Gumbart, J.C., Roux, B., and Chipot, C. (2013). Efficient determination of protein-protein standard binding free energies from first principles. *J. Chem. Theory Comput.* 9 (8): 3789–3798. https://doi.org/10.1021/ct400273t.

52 Ulucan, O., Jaitly, T., and Helms, V. (2014). Energetics of hydrophilicprotein-protein association and the role of water. *J. Chem. Theory Comput.* 10 (8): 3512–524.

53 Ulucan, O. and Helms, V. (2015). How hydrophilic proteins form nonspecific complexes. *J. Phys. Chem. B* 119 (33): 10524–10530. https://doi.org/10.1021/acs.jpcb.5b05831. PMID: 26218591.

54 Gumbart, J.C., Roux, B., and Chipot, C. (2013). Standard binding free energies from computer simulations: what is the best strategy? *J. Chem. Theory Comput.* 9 (1): 794–802. https://doi.org/10.1021/ct3008099. PMC3685508[pmcid].

55 Suh, D., Jo, S., Jiang, W. et al. (2019). String method for protein-protein binding free-energy calculations. *J. Chem. Theory Comput.* 15 (11): 5829–5844. https://doi.org/10.1021/acs.jctc.9b00499. PMID: 31593627.

56 Siebenmorgen, T. and Zacharias, M. (2019). Evaluation of predicted protein-protein complexes by binding free energy simulations. *J. Chem. Theory Comput.* 15 (3): 2071–2086. https://doi.org/10.1021/acs.jctc.8b01022. PMID: 30698954.

57 Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* 78: 2690–2693. https://doi.org/10.1103/PhysRevLett.78.2690.

58 Bennett, C.H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* 22 (2): 245–268. https://doi.org/10.1016/0021-9991(76)90078-4.

59 Cuendet, M.A. and Michielin, O. (2008). Protein-protein interaction investigated by steered molecular dynamics: the TCR-pMHC complex. *Biophys. J.* 95 (8): 3575–3590. https://doi.org/10.1529/biophysj.108.131383.

60 Rodriguez, R.A., Yu, L., and Chen, L.Y. (2015). Computing protein-protein association affinity with hybrid steered molecular dynamics. *J. Chem. Theory Comput.* 11 (9): 4427–4438. https://doi.org/10.1021/acs.jctc.5b00340. PMID: 26366131.

61 Laio, A. and Parrinello, M. (2002). Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* 99 (20): 12562–12566. https://doi.org/10.1073/pnas.202427399.

62 Laio, A., Rodriguez-Fortea, A., Gervasio, F.L. et al. (2005). Assessing the accuracy of metadynamics. *J. Phys. Chem. B* 109 (14): 6714–6721. https://doi.org/10.1021/jp045424k. PMID: 16851755.

63 Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* 100: 020603. https://doi.org/10.1103/PhysRevLett.100.020603.

64 Dama, J.F., Parrinello, M., and Voth, G.A. (2014). Well-tempered metadynamics converges asymptotically. *Phys. Rev. Lett.* 112: 240602. https://doi.org/10.1103/PhysRevLett.112.240602.

65 Limongelli, V., Bonomi, M., and Parrinello, M. (2013). Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci. U.S.A.* 110 (16): 6358–6363. https://doi.org/10.1073/pnas.1303186110.

66 Capelli, R., Carloni, P., and Parrinello, M. (2019). Exhaustive search of ligand binding pathways via volume-based metadynamics. *J. Phys. Chem. Lett.* 10 (12): 3495–3499. https://doi.org/10.1021/acs.jpclett.9b01183.

67 Raiteri, P., Laio, A., Gervasio, F.L. et al. (2006). Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B* 110 (8): 3533–3539. https://doi.org/10.1021/jp054359r. PMID: 16494409.

68 Singh, S., Chiu, C.-c., and de Pablo, J.J. (2011). Flux tempered metadynamics. *J. Stat. Phys.* 145 (4): 932–945. https://doi.org/10.1007/s10955-011-0301-0.

69 Valsson, O., Tiwary, P., and Parrinello, M. (2016). Enhancing important fluctuations: rare events and metadynamics from a conceptual viewpoint. *Annu. Rev. Phys. Chem.* 67 (1): 159–184. https://doi.org/10.1146/annurev-physchem-040215-112229. PMID: 26980304.

70 Tiwary, P. and Parrinello, M. (2013). From metadynamics to dynamics. *Phys. Rev. Lett.* 111: 230602. https://doi.org/10.1103/PhysRevLett.111.230602.

71 Salvalaglio, M., Tiwary, P., and Parrinello, M. (2014). Assessing the reliability of the dynamics reconstructed from metadynamics. *J. Chem. Theory Comput.* 10 (4): 1420–1425. https://doi.org/10.1021/ct500040r. PMID: 26580360.

72 Bussi, G., Gervasio, F.L., Laio, A., and Parrinello, M. (2006). Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.* 128 (41): 13435–13441. https://doi.org/10.1021/ja062463w. PMID: 17031956.

73 Barducci, A., Bonomi, M., Prakash, M.K., and Parrinello, M. (2013). Free-energy landscape of protein oligomerization from atomistic simulations. *Proc. Natl. Acad. Sci. U.S.A.* 110 (49): E4708–E4713. https://doi.org/10.1073/pnas.1320077110.

74 Banerjee, P., Mondal, S., and Bagchi, B. (2018). Insulin dimer dissociation in aqueous solution: a computational study of free energy landscape and evolving microscopic structure along the reaction pathway. *J. Chem. Phys.* 149 (11): 114902. https://doi.org/10.1063/1.5042290.

75 Lindahl, V., Lidmar, J., and Hess, B. (2014). Accelerated weight histogram method for exploring free energy landscapes. *J. Chem. Phys.* 141 (4): 044110. https://doi.org/10.1063/1.4890371.

76 Bonomi, M., Branduardi, D., Bussi, G. et al. (2009). PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* 180 (10): 1961–1972. https://doi.org/10.1016/j.cpc.2009.05.011.

77 Tribello, G.A., Bonomi, M., Branduardi, D. et al. (2014). PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* 185 (2): 604–613. https://doi.org/10.1016/j.cpc.2013.09.018.

78 Mottin, M., Souza, P.C.T., and Skaf, M.S. (2015). Molecular recognition of PPARγ by kinase CDK5/p25: insights from a combination of protein-protein docking and adaptive biasing force simulations. *J. Phys. Chem. B* 119 (26): 8330–8339. https://doi.org/10.1021/acs.jpcb.5b04269. PMID: 26047365.

79 Sugita, Y. and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314 (1): 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9.

80 Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett. (EPL)* 19 (6): 451–458. https://doi.org/10.1209/0295-5075/19/6/002.

81 Earl, D.J. and Deem, M.W. (2005). Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* 7: 3910–3916. https://doi.org/10.1039/B509983H.

82 Sugita, Y. and Okamoto, Y. (2000). Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.* 329 (3): 261–270. https://doi.org/10.1016/S0009-2614(00)00999-4.

83 Kimanius, D., Pettersson, I., Schluckebier, G. et al. (2015). SAXS-guided metadynamics. *J. Chem. Theory Comput.* 11 (7): 3491–3498. https://doi.org/10.1021/acs.jctc.5b00299. PMID: 26575782.

84 Aldeghi, M., Gapsys, V., and de Groot, B.L. (2018). Accurate estimation of ligand binding affinity changes upon protein mutation. *ACS Cent. Sci.* 4 (12): 1708–1718. https://doi.org/10.1021/acscentsci.8b00717.

9

Markov State Models of Protein–Protein Encounters

Simon Olsson

Chalmers University of Technology, Department of Computer Science and Engineering, Rännvägen 6, 41258 Göteborg, Sweden

Notation

Symbol	Explanation
$\mathbb{P}[x]$	the probability of an event x
$p(x)$	a probability density function
$x \mid y$	event x given y. This may occur in probability densities or probabilities of events
$\mathcal{P}_\tau$	Markov Propagator. A "continuous space equivalent" of a Markov state model
$\langle\rangle$	Ensemble average with respect to the stationary distribution (Boltzmann distribution)
$\mu(x)$	Boltzmann distribution

9.1 Introduction

The encounter of proteins is critical to countless biological processes and may span several lengths- and time-scales [1]. For example, insulin binds the α-subunits of insulin receptors, thereby activating the tyrosine-kinase β-subunit auto-phosphorylation triggering a signal transduction cascade, leading to a broad range of responses from the molecular, over cellular, and to the physiological scales [2]. Every step along this cascade involves protein–protein interactions between different proteins or multiple copies of the same protein chain. This example is just one of many illustrating why mapping out the details of protein–protein encounters at the atomistic and molecular scale is critical to understanding these processes, what goes wrong in disease states, and inform intervention strategies to remedy or reverse pathological conditions [3].

Indeed, massive-scale efforts have attempted to characterize protein–protein interaction networks using high-throughput experimentation [4], and insights gained from these endeavors have undoubtedly been incredibly impactful [5, 6].

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

However, the strengths of these proteomic approaches lie in their broad scope but not in their resolution. Currently, only biophysical and molecular simulation techniques allow us to dissect the intimate structural, thermodynamic, and kinetic details [7–12]. For example, cryogenic electron microscopy and X-ray crystallography may potentially give us high-resolution snapshots of the encounter process at various stages [11, 13, 14]. Single-molecule fluorescence resonance energy transfer (FRET) spectroscopy can give structural and kinetic insights into protein–ligand binding [15]. Finally, nuclear magnetic resonance (NMR) spectroscopy enables detailed characterizations of protein–protein encounters, possibly giving us structural, thermodynamic, and kinetic insights, given favorable experimental conditions [7, 16–18]. Molecular dynamics simulations with explicit solvation uniquely give us a fully spatiotemporally resolved view of protein dynamics [19–26] including the encounter mechanism [27–31]. Advances in software and hardware technology enable us to routinely reach aggregate simulation timescales that overlap with experimental timescales for small protein–protein systems, especially when using kinetic modeling approaches, such as Markov state models (MSMs) [32–34].

This chapter will outline how molecular dynamics simulations, experimental data, and MSMs can synergize to map out the mechanism of protein–protein association and dissociation. Further, I will discuss whether we can currently estimate accurate rates and thermodynamics of critical metastable states. First, I motivate MSMs in the light of molecular dynamics theory. Then I outline the practical aspects of applying MSMs to studying protein–protein encounters and show some successful examples from the literature. I will further discuss how to use experimental data to validate and augment MSMs estimated from molecular simulation data. I will close with a few examples of emerging technologies that may improve the computational study of protein–protein encounters in the future.

9.2 Molecular Dynamics and Markov State Models

When applying molecular dynamics simulations, we aim to understand biomolecular processes. Ideally, our understanding must build on statistically robust scientific observations. The key observables of interest are:

1. Important structures,
2. their thermodynamic weights,
3. and the transition probabilities amongst them, or their interconversion rates.

Robust identification of these three properties allows us to directly connect MD results to experimental data, including NMR spectroscopy and single molecule FRET (sm-FRET) [35–38]. Comparisons such as these may serve as an important complementary means of validating the simulation models and can help drive robust scientific hypotheses and models.

Analysis of MD simulations, however, often relies on visually inspecting simulation trajectories one by one. Alternatively, we follow the simulation trajectories projected onto a few order parameters (or collective variables) derived from chemical

intuition about the process of interest or some global structural property [39–43]. Inspecting structures and following certain order parameters is an integral part of any analysis of molecular dynamics simulations. However, these strategies alone do not guarantee a statistical relevance of events observed, and the overall approach becomes increasingly time-consuming with growing datasets. Furthermore, limiting ourselves to these analyses may still overlook rare events important for biological function. So ultimately, conclusions drawn from these kinds of analyses may be misleading [32].

Statistical models to analyze data from MD simulations are enjoying increased attention in recent years [44–52]. This popularity is a necessary consequence of growing datasets enabled by improvements in software efficiency and large-scale investment into consumer-grade GPU (graphical processing units)-based compute resources by many academic groups. Another important factor is community-driven, cloud-based supercomputers such as Folding@Home [53] and GPUgrid (www.gpugrid.net) that generate enormous volumes of simulation data whose analysis critically relies on a systematic and principled framework. Markov state models (MSMs) are one prominent example of statistical models for analyzing molecular dynamics simulation, which fits the bill [32, 44, 46, 54].

This section will briefly discuss the motivation and theoretical basis of MSMs and some important mathematical properties of MSMs. With this text, I do not attempt to discuss these topics comprehensively but instead, provide a guiding primer and to enable the reader to build some intuition about the theory – in general, the text is based upon the references cited in this section. However, I intentionally minimize technical language and equations and avoid specific details in the notation for clarity. For a more detailed MSM theory treatment, I refer to the excellent review by Prinz et al. [32]. For a more comprehensive historical overview of MSMs, I refer to the review of Husic and Pande [33]. A recent tutorial for step-by-step MSM building is also available [34].

9.2.1 Markov State Models: Theory and Properties

Above, I outlined how we need to minimize the subjectivity going into analyzing data from MD simulations. Such subjectivity may stifle our ability to detect transient intermediate, or off-pathway, states, parallel protein–protein association pathways, and other intricate kinetic features. Consequently, we need simplification of the conformational space to enable human interpretation of the results. However, we should achieve this in a manner that supports our goals to extract as much kinetic and thermodynamic information from our simulation data as possible. MSMs provide a framework for achieving this goal.

But what *is* a MSM? – A MSM is an $N \times N$ matrix where each element encodes the conditional probability of ending in a state i from state j after a constant time, τ [44]. The N states each represent a different disjoint segment of the configurational space. Therefore, the MSM gives us an *ensemble view* of the molecular dynamics, where each trajectory corresponds to a sample from a distribution of dynamics trajectories [32] (Figure 9.1). This view is exactly analogous to that taken in statistical mechanics

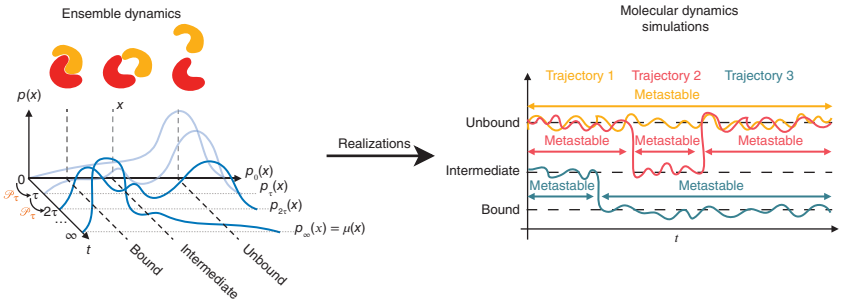


Figure 9.1 Illustration of the relationship between the ensemble view of dynamics and the individual realizations we obtain from molecular dynamics simulations.

and thermodynamics: the accuracy at which we can characterize the important properties of a molecular system is limited by how well we can estimate the statistical distribution of the system's dynamics. Any given trajectory will typically be too short to be representative of the full system dynamics. However, estimating this statistical distribution would allow us to pinpoint important structures and compute their thermodynamic and kinetic properties. The statistical distribution further allows us to predict how a nonequilibrium initial condition — prepared in an experiment — relaxes back to equilibrium or to predict experimentally measurable spectroscopic observables [35–37].

At first glance, estimating this statistical distribution may seem completely infeasible: the distribution domain is all possible temporal trajectories of a molecular system with all-atom detail. To make this estimation tractable, we rely on the following assumptions:

1. Time-homogenous Markovian dynamics
2. Ergodicity
3. Reversibility.

The first assumption restricts the dynamics we can consider to one where the transition probabilities from x_{t_M-t} to x_{t_M} after some time t are independent of what happened before. These transition probabilities do not change with time. More formally, we can simplify the conditional probability of arriving in x_{t_N} given all prior states, $x_{t_0}, \ldots, x_{t_M-t}$, by,

$$\mathbb{P}[x_{t_N} \mid x_{t_0}, \ldots, x_{t_M-t}, t] = \mathbb{P}[x_{t_M} \mid x_{t_M-t}, t]$$

that is, the probability of arriving in a state at time t_N only depends on the state the system was in at $t_M - t$ and that this probability is *invariant* to a time-shift – *homogeneous*. A trajectory of a systems dynamics is represented here by the states the system adopts $x_{t_0}, \ldots, x_{t_M-t}$ at a sequence of time points sampled uniformly in time $t_0, \ldots t_M - t$.

The second assumption tells us that we can reach any point in configuration space from any other point in configuration space within some finite time. There is a nonvanishing probability of arriving at any state x' from any other state x in a finite time. This assumption ensures the configuration space to be *dynamically connected*.

The final assumption ensures that the probability flux between points x and x' in configuration space is the same in either direction. In physical terms, this means that energy is not extracted or generated in any state. Formally, this corresponds to the fulfillment of the *detailed balance* condition

$$\mathbb{P}[x']\mathbb{P}[x \mid x'] = \mathbb{P}[x]\mathbb{P}[x' \mid x]$$

where $\mathbb{P}[x]$ is the stationary — equilibrium – probability of state x, typically given by the Boltzmann probability $\mathcal{Z}^{-1}\exp(-\beta U(x))$ for molecular systems at thermal equilibrium. I have suppressed the time dependence of the conditional transition probability for notational brevity. Strictly, this final assumption is unnecessary as many simulation setups involve doing work on the molecular system. In such scenarios, other factors will drive the system beyond the thermal fluctuations, and, in general, the system will not be in thermal equilibrium. Nevertheless, the fulfillment of the detailed balance condition leads to the symmetry of the joint probability $\mathbb{P}[x, x'] = \mathbb{P}[x', x]$, which we will see allows for a more statistically efficient estimation of MSMs in many cases.

Are all of these assumptions fulfilled in any practical cases? – Yes! Most of the common thermostatting algorithms used are consistent with the assumptions I outline above in molecular simulations. Prinz et al. discuss notable exceptions [32].

Remember, our original goal was to arrive at an *ensemble view* of molecular dynamics. This view describes the time evolution of many copies of the same molecular system. The copies are independent, and do not interact with each other, and are distributed according to the Boltzmann distribution, when at equilibrium, $\mu(x) = \mathcal{Z}^{-1}\exp(-\beta U(x))$. $\mathcal{Z}$ is the partition coefficient, $U(\cdot)$ is the system potential energy at the experimental conditions, and the inverse temperature $\beta = 1/k_B T$, with k_B and T being Boltzmann's constant and the system temperature, respectively. There is a rigorous theoretical framework to treat systems in such a way, however, we will here limit the discussion to the time-discrete cases, as these most directly relate to the MSM framework. Time-continuous models discussed elsewhere e.g. in Ref. [47], have analogous results [32].

The object of interest here is a *propagator*, $\mathcal{P}_\tau$. The propagator is an "integral operator," that acts on a probability density function, $p_t(\cdot)$, over – in our case – conformational space and returns the resulting probability density function on the same space after a time, τ. Formally,

$$p_{t+\tau}(x) = [\mathcal{P}_\tau p_t](x) = \int p(x \mid x', \tau) p_t(x')\, \mathrm{d}x'$$

where, $p(x \mid x', \tau)$ is the transition probability density function from x' to x after a time τ. If $p_t(x)$ is equal to the equilibrium distribution (Boltzmann distribution), then $p_{t+\tau}(x) = p_t(x) = \mu(x)$. In general, if we apply the propagator to some initial distribution $p_0(x)$ infinitely many times we arrive at the distribution $p_\infty(x) = \mu(x)$. In other words, the propagator describes how an initial condition, $p_0(x)$, relaxes to equilibrium.

This observation reminds us of an eigenvalue problem, where the Boltzmann distribution is a solution (eigenfunction), with the corresponding eigenvalue 1. Indeed, the propagator has infinitely many eigenfunctions, ϕ_i, whose eigenvalues

are bounded $1 > |\lambda_i|$ for reversible dynamics. Ergodic dynamics further ensure that only one eigenfunction has eigenvalue 1: namely the Boltzmann distribution $\phi_1 = \mu$.

The eigenvalues of $\mathcal{P}_\tau$ are the autocorrelations of the eigenfunctions ϕ_i, which follow single-exponential decays $c_{\phi_i}(\tau) = \lambda_i = \exp(-\kappa_i \tau)$, as $\mathcal{P}_\tau$ is first-order Markovian. $\kappa_i \geq 0$ are exchange rates, and $1/\kappa_i$ is often referred to as an implied timescale (ITS) [32]. We immediately notice that the implied timescale for μ is equal to ∞, which is consistent with our understanding that the Boltzmann distribution is stationary: it does not change with time under fixed conditions. Simultaneously, this observation suggests that all $\lambda_i < 1$ approach 0 for large τ, meaning that they – together with their corresponding eigenfunctions – encode information about the dynamics of our molecular system. These eigenfunctions (for $|\lambda_i| < 1$) describe what regions of conformational space exchange, on the timescale $1/\kappa_i$. The negative and positive signs of an eigenfunction, ϕ_i, define two regions of conformation space that are exchanging on the timescale $1/\kappa_i$.

So, the more eigenfunction-eigenvalue pairs we know the more we know about the ensemble thermodynamics (μ) and dynamics ($\phi_{i>1}$, $\lambda_{i>1}$) of our system – but how do we deal with the infinite amount of these pairs? — This question is key to a central assumption made when using MSMs: we are only interested in a small number, M, of pairs that correspond to those with the M largest eigenvalues. The larger the eigenvalue the slower the timescale – consequently, we focus our attention on slow dynamics. Immediately, this focus makes a lot of sense, since long timescales often are associated with biological function, including allosteric regulation and protein–protein binding. Simultaneously, long timescales remain challenging to study with unbiased MD compared to fast dynamics. The success of this approach lies in how representative the M largest eigenvalue-eigenfunction pairs are for the dynamics as a whole. Fortunately, for many systems, there are only a handful of eigenvalues that are close to 1, while the rest are close to 0.

Recall that the dynamics in the continuous space is Markovian by construction. To approximate the dynamics of a system from finite MD data, it is an advantage to discretize conformational space. The Markov state model (MSM) approach emerges naturally from this approximation. An MSM aims to *approximate* continuous space dynamics via a discrete space jump-process on a partition of the configuration space into N disjoint segments. The discretization of the space and the $N \times N$ transition probability matrix, T_τ, describing the "jump-process" constitutes the approximation. Since T_τ is an approximation of the continuous space dynamics, its eigenvectors and eigenvalues will—if properly built—approximate their corresponding quantities in the continuous space dynamics. The eigen-decomposition of T_τ, takes the form

$$T_\tau = \sum_{i=1}^{N} \lambda_i \boldsymbol{l}_i^\top \mathbf{r}_i \tag{9.1}$$

where $\boldsymbol{l}_i$ and $\boldsymbol{r}_i$ are orthonormal left and right eigenvectors, respectively. The left eigenvectors are given by $\boldsymbol{l}_i = \boldsymbol{\mu} \circ \boldsymbol{r}_i$, and $\circ$ is the element-wise product between two vectors. In this expression, we see more explicitly how eigenvectors with smaller numerical eigenvalues (faster timescales) contribute less numerically to the transition probability matrix.

Key message: The full-space dynamics contain essential thermodynamic and kinetic information that we need to characterize, for example, a protein–protein encounter process. How well we reduce the full-space into a set of discrete states controls the quality of our model. A sound reduction of the full-space minimizes the error of the eigenfunction corresponding to the largest eigenvalues of $\mathcal{P}_\tau$.

9.3 Strategies for MSM Estimation, Validation, and Analysis

As we saw above, building MSMs relies on discretizing conformational space into N disjoint segments. These segments need to provide a good basis for approximating the propagator eigenfunctions to ensure that we achieve the best possible approximation of the full space dynamics. The continuous space dynamics is typically high-dimensional, for all but the simplest systems, so it is not practical to place a fine grid on all dimensions. Placing such a grid would require enormous computer memory and simulation data to be successful. We are facing what in statistics is called the curse of dimensionality. In practice, building a MSM involves a sequence of four steps [34],

- Featurization – selecting a suitable representation of the molecular system
- Dimension reduction – reducing the representation of the molecular system
- Clustering – discretization of the representation
- Transition matrix estimation – estimation of the MSM.

9.3.1 Variational Approach for Conformational Dynamics and Markov Processes (VAC and VAMP)

When we build MSMs, we express the molecules' thermodynamic and kinetic properties on a discrete set of disjoint states. Adopting this strategy means that we approximate the eigenfunctions using a combination of indicator functions – functions that return one if we are in a certain area of configuration space and zero everywhere else. However, this is just one way of approximating the eigenfunctions, and we are free to approximate them with any function we like. The variational approach for conformational dynamics (VAC) [55, 56] gives us a principle to select the function that best approximates a molecular system's slow dynamics from a set of trial functions. Here, I briefly outline the idea – more detailed treatments are available elsewhere.

VAC uses that the eigenvalues of $\mathcal{P}_\tau$ are bounded and the eigenfunctions form an orthonormal basis. Consequently, if f_α is an approximation of the α'th eigenfunction of $\mathcal{P}_\tau$ the autocorrelation is given by

$$c_\tau(f_\alpha) = \int f_\alpha(x)\mu^{-1}(x)\mathcal{P}_\tau f_\alpha(x)\, \mathrm{d}x \leq \lambda_\alpha$$

where the equality holds if and only if $f_\alpha(x)$ is *exactly* the α'th eigenfunction of $\mathcal{P}_\tau$. Hence the variational principle tells us that we will always approximate the autocorrelation of an exact eigenfunction $\phi_\alpha(x)$ from below. Practically, this means we

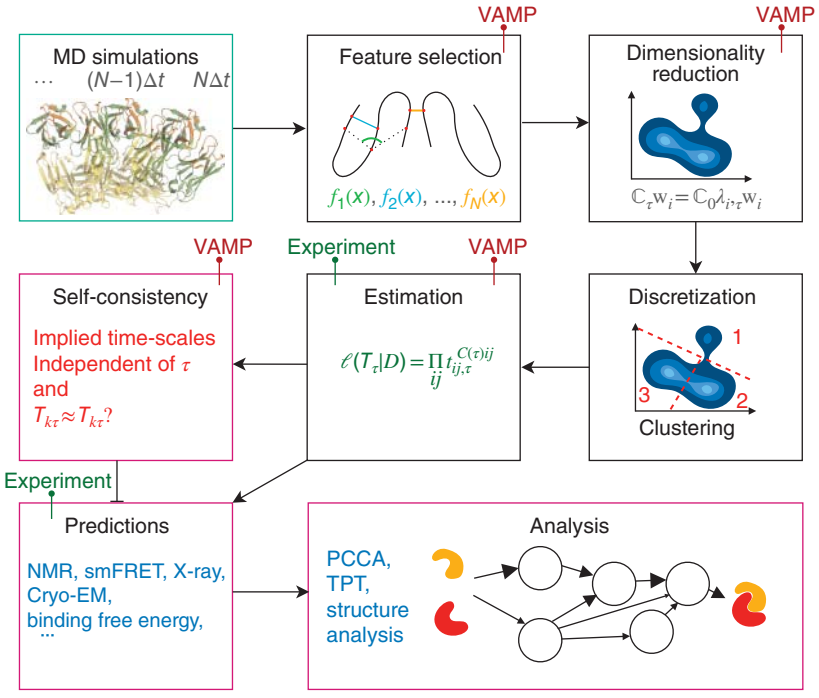


Figure 9.2 Flow-chart of Markov state modeling from molecular dynamics simulations to final model and analysis. Boxes are colored to indicate data collection (cyan), data processing and model estimation (black), and analysis and validation (magenta). I highlight steps that may benefit from specific techniques or experimental data by colored pins.

can devise algorithms to approximate a set of orthonormal approximations of the eigenfunctions of $\mathcal{P}_\tau$.

A more general variational approach for Markov processes (VAMP) [57] extends VAC to nonreversible dynamics, nonequilibrium data, and allows us to define scores that can be used for hyperparameter optimization, cross-validation, and model selection when building MSMs. These VAMP-scores summarize the autocorrelations on a set of basis functions (features), which best approximate the underlying dynamics, and therefore how well they represent slow dynamics. We can use the VAMP-scores at every step of the MSM building process to evaluate how well our modeling decisions will allow us to represent the slow dynamics of a molecular system.

9.3.2 Feature Selection

To facilitate the estimation of MSMs, we will need to arrive in a sufficiently low-dimensional space to enable effective discretization. However, the space has to include sufficient detail to capture the interesting slow processes in our dataset.

Fortunately, we frequently have a clear idea of what kinds of processes we are interested in resolving, or more specifically, what features we are not interested in resolving. For example, in many cases, we are not directly interested in studying the influence of solvation or the rotational and translational motion of the solutes. This focus leaves us with studying different internal coordinates or intermolecular coordinates when selecting features for building MSMs. These internal coordinates – or *features* – typically include contacts, distances, angles, and torsions between atoms or atom groups.

While the considerations outlined above refine our choice of possible structural features to use in our model building, it still leaves open an enormous set of potential structural features. To further narrow down this ambiguity, there are two different strategies:

1. Manual feature selection by selecting features based on chemical, biological, or physical insights that give us some information about possible slow processes
2. Algorithmic feature selection strategies.

It is difficult to approach the first strategy in a general and systematic way. Typically, this strategy involves manually refining the selection of features such that the model is robust and provides the necessary predictive and descriptive power envisaged for the project. The second approach is typically more systematic and generalizable and will normally be the best choice if we know little about the system beforehand. Several methods provide automated feature selection specifically designed with MSM building in mind: Scherer et al. illustrate use of VAMP in this respect [58], and Chen et al. use a genetic algorithm-based method for feature selection [59]. The former method works directly on the features, whereas the latter approach relies on subsequent modeling steps and their associated hyperparameters to evaluate the selected features. Therefore, the latter method is sensitive to model decisions beyond the feature selection, and special care must be taken when using this approach.

9.3.3 Dimensionality Reduction

Usually, preselecting several features (distances/contacts, angles, features, etc.) using the strategies outlined above is insufficient to sufficiently reduce the space to enable effective discretization of the conformational space. Alternatively, we may not know much about the system before starting our analysis, and we may want to identify structural features that characterize the molecular dynamics well. To face this problem, we can use dimensionality reduction techniques. These methods remove dependencies in the input data through linear (or nonlinear) combinations learned utilizing a range of different optimality criteria, thereby allowing us to represent the original data in a lower-dimensional space while keeping the optimality criteria used as small as possible. Dimensionality reduction techniques have their origin in machine learning and statistics in a branch which is now broadly referred to as unsupervised learning.

In the context of MSM, principal component analysis (PCA) [60–63] and time-lagged independent component analysis (TICA) [64–66] are the most widely used. Here I limited the discussion to these two.

PCA seeks to define a linear projection,

$$\mathbf{Y} = \mathbf{X}\mathbf{W}$$

of the set of input features, $\mathbf{X} \in \mathbb{R}^{N_{\text{numframes}} \times N_{\text{numfeats}}}$, to *maximize the variance* of each of the dimensions of $\mathbf{Y} \in \mathbb{R}^{N_{\text{numframes}} \times N_{\text{reduceddim}}}$, by learning $\mathbf{W} \in \mathbb{R}^{N_{\text{numfeats}} \times N_{\text{reduceddim}}}$, subject to an orthonormality constraint on the columns of $\mathbf{W}$ to ensure that each dimension in $\mathbf{Y}$ is uncorrelated and normalized. Consequently, PCA gives us a new set of features that best capture the variance of our input features and is an appropriate choice if we are interested in studying processes characterized by large-scale structural fluctuations.

TICA similarly seeks to find a linear projection as for PCA. However, instead of maximizing the variance, TICA uses the variational principle of conformational dynamics to determine projections with the slowest auto-correlation. Consequently, TICA is the appropriate choice if slow motions are of interest when studying a molecular system. Recall, slow dynamics is what constitute the dominating part of the propagator. Practically, we compute TICA by solving the generalized eigenvalue equation (subject to appropriate normalizations),

$$\mathbb{C}_\tau \mathbf{w}_i = \mathbb{C}_0 \lambda_{i,\tau} \mathbf{w}_i$$

where $\mathbb{C}_\tau = \frac{1}{N_{\text{numframes}} - \tau} \mathbf{X}^\top_{:\mathbf{numframes}-\boldsymbol{\tau}} \mathbf{X}_{\tau:}$ and $\mathbb{C}_0 = \frac{1}{N_{\text{numframes}}} \mathbf{X}^\top \mathbf{X}$ are the time-lagged and instantaneous covariance matrices, respectively. $\mathbb{C}_\tau$ computes the covariance between features spaced in time by τ and the indices $: \text{numframes} - \tau$ and $\tau :$ mean all but the last τ frames and the all but the first τ frames, respectively. At this point, τ is an integer with a time unit of the spacing interval between the frames in your MD trajectory data. We can use the *independent components*, $\mathbf{w}_i$, which solve this equation and correspond to the largest eigenvalues $|\lambda_{i,\tau}| < 1$ to project the data on to a lower-dimensional space, $\mathbf{y}_i = \mathbf{X}\mathbf{w}_i$, which conserves the slowest dynamic modes in the system. We use the total kinetic variance ς_τ^2 to quantify how much dynamic is preserved in the d-dimensional projection ($d < N_{\text{numfeats}}$),

$$\varsigma_\tau^2 = \frac{\sum_{i=2}^{d} \lambda_{i,\tau}^2}{\sum_{j=2}^{N_{\text{numfeats}}} \lambda_{j,\tau}^2}.$$

Both PCA- and TICA-based dimension reduction methods are part of the major MSM software packages PyEMMA [34, 67] and MSMBuilder [68].

Recent surveys discuss the use of nonlinear dimensionality reduction techniques in the context of MSM estimation. While promising, these methods have not seen broad adoption so far.

9.3.4 Clustering

MSMs rely on discretizing the configurational space into disjoint configurational states – *micro-states*. Clustering is the step where the grouping of molecular

configurations into discrete states happens. The most commonly applied algorithm towards this purpose is *k*-means clustering, yet several other methods perform this task with a variety of different strategies [69]. As for feature selection, we can use VAMP scores and cross-validation to evaluate our clustering quality. Below, I expand on other considerations that are important for clustering the states when studying protein–protein encounters.

9.3.5 Model Estimation and Validation

Following clustering, we can assign every molecular configuration to a Markov state. Trajectories now realize a jump process on a set of, N, discrete states, each of the states are connected back to a molecular configuration. We call these *discrete trajectories*, $D = \{d_1, \dots, d_M\}$. The task of estimating a Markov state model corresponds to computing the most likely transition probabilities, $t_{ij,\tau}$ between any two states i and j after a lag-time of τ. Recall, we assume the dynamics are Markovian, so the *likelihood* of observing our data, D, is equal to the product of all the transition probabilities,

$$\begin{aligned}\mathscr{\ell}(T_\tau \mid D) &= \prod_{d\in D} p(d[\tau] \mid d[0], \tau) p(d[2\tau] \mid d[\tau], \tau) \dots p(d[M] \mid d[M-\tau], \tau) \\ &= \prod_{d\in D} t_{d[0]d[\tau],\tau} t_{d[\tau]d[2\tau],\tau} \cdots t_{d[M-\tau]d[M],\tau} \\ &= \prod_{ij} t_{ij,\tau}^{C(\tau)_{ij}}\end{aligned}$$

$C(\tau)$ is the count matrix where each element $C(\tau)_{ij}$ is the number of transitions between states i and j, with a time lag, τ, observed in all the trajectories, D. Estimating a MSM then corresponds to finding the transition probabilities, given the observed transition counts in the count matrix, $C(\tau)$. We can either do maximum likelihood estimation [32, 70], or Bayesian posterior sampling of the transition probabilties [38, 71]. The first approach gives us the one most likely model, whereas the latter approach gives us a distribution of models that we can use to compute properties as well as their uncertainties.

Major MSM software packages implement algorithms to perform inference via either mode, with options to enforce constraints such as detailed balance [38, 71] or a fixed stationary distribution [72]. As outlined above, the detailed balance constraint ensures that a reversible MSM is estimated and reduces the number of degrees of freedom to be estimated. Adding constraints to the estimation when possible is often desirable, as it may increase robustness of the results.

We can estimate a reversible MSM of a slow process even if we only have data reversibly sampling transitions between intermediate steps. Consequently, we can get quantitative information about very slow processes by partitioning them into multiple faster sequential steps. However, this may not always be possible.

Choosing the lag time when building a MSM decides the effective time-resolution of the resulting model [73]. Consequently, we want to keep this number small, to preserve as much of the information in our data as possible. However, since we reduce a high-dimensional space down to a lower-dimensional one to enable discretization,

there is no guarantee that the projected dynamics will be Markovian at short lag times [74–76]. We check the "Markovianity" of the projected dynamics by computing the ITS as a function of lag time and ensuring no systematic change in the ITS as a function of lag time considering the statistical uncertainty. A good choice of lag time is then one which is as short as possible, yet shows no significant change in the ITS when increased or decreased slightly. This analysis is typically facilitated by an ITS plot, showing the ITS as a function of lag time.

Having selected an appropriate lag time, we can test the resulting MSM for self-consistency with the simulation data via the Chapman–Kolmogorov (CK) test [20, 32]. This test makes use of the time-discrete Chapman-Kolmogorov equation

$$T_{k\tau} = T_{\tau}^{k}$$

which predicts that the transition probabilities of a model estimated with a lag time $k\tau$ should be equal to the transition probabilities of a model estimated with a lag time τ to the power of k. We typically visualize the CK test by comparing the values on either side of the equation with error bars as a function of integer multiples of the MSM lag-time. As for the ITS analysis, we here aim to see agreement within statistical uncertainty. Usually, only a reduced set of states, or a coarse-grained model, is used to facilitate analysis.

9.3.6 Spectral Gaps and Coarse-Graining

It is not uncommon that MSMs end up having hundreds or thousands of microstates. The large number of microstates helps us to bring down the error when approximating eigenfunctions. However, it can stifle the subsequent analysis. Consequently, we often coarse-grain the MSMs into a handful of metastable macrostates, which summarize the slow dynamics. Coarse-graining here should not be confused with coarse-grained simulations, where beads represent multiple atoms. We have to decide how many states, and that number may not be evident from the start. In many cases, we can use the spectral gap in the eigenvalue spectrum of the MSM to decide on how many states we need to coarse-grain a MSM to, to ensure that we represent the slow dynamics.

Suppose we sort the eigenvalues-eigenvectors pairs of a MSM by the amplitude of the eigenvalue, and plot them. In that case, we often see one or more drops in the amplitude with increasing index (decreasing eigenvalue). These drops are spectral gaps and pinpoint separations between fast and slow dynamics in the molecular system represented by the MSM. We can use these spectral gaps to decide on how many states to use for a coarse-graining, as every eigenfunction specifies what two regions of conformation space are exchanging on the ITS that can be computed from the eigenvalue. Consequently, if we have n eigenvalues that are less than 1 above a spectral gap, a $n + 1$ state coarse-graining will be appropriate.

Perron Cluster–Cluster Analysis (PCCA) [77, 78] is a method that groups together microstates based on the sign structure of the eigenvectors of a MSM. It is the most common way to identify important metastable macrostates sampled during molecular dynamics simulations. Two related algorithms, PCCA+ and PCCA++, find an optimal linear transformation of the eigenvector coordinates onto a probability simplex [79].

Hidden Markov state models (HMMs) are an alternative to both MSMs and PCCA [80]. HMMs avoid the assumption of MSMs of Markovian dynamics in the reduced space by estimating a "hidden" Markov chain observed indirectly via the trajectory data on the discrete microstates. A HMM, therefore, estimates a transition probability matrix, and an "emission matrix", E. The first matrix is responsible for modeling the dynamics, and the latter models the observation process: given we are in hidden state i, we will be in microstate j with probability $p(j \mid i) = E_{ij}$. Consequently, the emission matrix tells us what states exchange rapidly, given we are in a specific metastable configuration. We can use this to simplify the many states into just a few states. HMMs have a range of other theoretical advantages but are also more challenging to estimate than MSMs. There are other alternatives to defining lower-dimensional models to facilitate analysis of slow dynamics in terms of a few metastable states. However, their performance in the context of protein–protein encounters is currently unknown [81, 82].

9.3.7 Adaptive and Enhanced Sampling Strategies

The quality of the molecular dynamics simulation data ultimately determines the quality of the estimated MSMs. Here, quality means the number of transitions sampled between configurational states of interest for the molecular system. An advantage of MSM analysis is that we do not necessarily need to sample transitions between all states of interest in every trajectory but sample only a subset of the possible transitions. However, in practical cases, we still have to make the most of limited resources – blindly or naively running numerous simulations may not be the most effective.

Adaptive sampling strategies (semi) automatically decide how multiple simulations run in parallel and over several "epochs". These strategies have to balance exploration and exploitation: sampling new states and refining sampling statistics between previously visited states [83–85]. Several groups have proposed strategies using different assumptions about what is important for characterizing molecular systems [27, 49, 83, 86–90]. A complementary set of strategies aim to sample transitions between known states [91–93]. However, due to the relatively high computational cost of studying protein–protein encounters, these methods are yet to be compared in rigorous benchmarks.

Enhanced sampling methods bias molecular dynamics simulations intending to speed up sampling processes of interest, such as protein–protein binding and unbinding [39, 41, 94, 95]. Unfortunately, introducing the right biases to enhance the sampling of a process of interest remains a labor-intensive process. Nevertheless, methods are available to recover stationary properties from biased simulations, yet proving more difficult for dynamic properties. However, combining unbiased and biased simulation data via recent MSM estimation techniques can significantly improve the estimated models' robustness [29, 96–100]. In Section 9.4 we highlight the successful use of adaptive and enhanced sampling techniques to study protein–protein binding-unbinding equilibria.

9.3.8 Practical Consideration for Studying Protein–Protein Encounters

The procedures outlined above generally apply to molecular systems. However, there are additional aspects that we need to take into account when modeling protein–protein encounters.

We can use macroscopic variables, including concentration, temperature, pressure, and mutations, to control the molecular system's ensemble, including the population of bound and unbound states and their kinetics of exchange. As we have discussed, when following experimental observables as a function, these variables allow us to quantify essential properties such as affinities, rate constants, and structural information about the complex formation process.

Computationally, we often have to settle on a single – or a few – macroscopic setting(s) of variables to study. This limitation is due to the large computational requirements associated with sampling each condition, even using the advanced simulation strategies, including those outlined above. A notable exception is the temperature, which is leveraged in enhanced sampling techniques to improve sampling efficiency. When analyzed together with regular MD simulation data using appropriate statistical estimators, they may improve MSM estimation. However, using these data on their own makes it challenging to get insights about exchange kinetics between conformational states.

The primary differences we face when studying protein–protein encounters, compared with studying the molecular dynamics in a single protein molecule, are stoichiometry and concentration. Practically, the simulation volume is limiting: when we increase the volume, we need to simulate larger systems, usually comprising an increasing number of water molecules. This fact makes simulations with high protein concentrations the only computational viable strategy currently.

A high concentration in molecular dynamics has some practical consequences that may make it practically difficult to study certain mechanisms of protein–protein encounters. Let us consider a case of the conformational selection mechanism, where a low-population state A^* of unbound protein A binds the protein B to form the complex $A^* : B$ in the following reversible chemical kinetic relation

$$A \rightleftarrows A^* + B \rightleftarrows A^* : B$$

where $k_{A\to A^*} \ll k_{A^*\to A}$ with both rates being independent of concentration. We assume that the protein B does not undergo conformational changes, which perturb this relation directly. The on-rate, $k_{A^*+B\to A^*:B}$, is proportional to the protein concentration and the population of the unbound state of A^*. So as concentrations increase, the probability of observing binding events increase. In an alternative binding mechanism (induced fit) binding happens before conformational change in the protein A,

$$A + B \rightleftarrows A^* : B.$$

Here, the on-rate (the rate of binding), $k_{A+B\to A^*:B}$, depends on the protein concentration and the population of the highly populated state of protein A. In many reported cases both mechanisms are possible, consequently, we seek to understand

the balance of these two mechanisms – more generally, we seek to characterize the association–dissociation path ensemble [101, 102]. However, since we are at high concentrations we may have $k_{A^*+B\to A^*:B} \ll k_{A+B\to A^*:B}$, and we may even have $k_{A\to A^*} \ll k_{A^*\to A} \ll k_{A+B\to A^*:B}$. So with only finite MD simulation data, we may severely undersample or completely miss certain mechanisms, even if they are important. In other words, high protein concentrations in MD simulations may increase the free energy of the unbound state to the point where the association is barrier free, and the unbound state is not metastable [8].

More concretely, given the competition between these mechanisms, sampling the induced-fit mechanism is much more likely than sampling the conformational selection mechanism. Even conformational sampling of protein A is much less likely than sampling binding via induced fit. Practically, these conditions mean that we will have an intrinsic preference to observe a certain biophysical binding mechanism and may over-sample mechanisms that are not relevant at physiological protein concentrations, including unspecific binding events. As a result, we would need to acquire more simulation data to ensure statistically sufficient sampling of alternative binding mechanisms and conformational mixing of the unbound states.

The fast on-rates at high concentrations may also influence our ability to distinguish unbound and bound states automatically. The timescale, t_i, of a process, i, between states a_i and b_i depends on the geometric average of the rates of the forward and backward process $t_i = \frac{1}{\kappa_i} = \frac{1}{k_{a_i\to b_i}+k_{a_i\to b_i}}$, which is numerically dominated by the faster (larger) rate. As a result, the timescale of the binding–unbinding process will be fast. Therefore, dimension reduction techniques resolving coordinates with slow exchange rates, may not resolve it. Consequently, a MSM based on a clustering defined only in this space will miss the process altogether. However, we can overcome this problem by explicitly separating bound and unbound states, such as a molecular feature that clearly distinguishes the unbound and bound states.

A final complication we face when studying protein–protein encounters is the choice of an appropriate force field model. This choice may significantly affect the sampled binding mechanisms and may be prone to deficiencies, such as strong unspecific binding [103]. Although efforts continuously improve these force field models and address their outstanding limitations, we often do not know how well a given force field will represent a new system of interest before we start simulations. In Section 9.4, we discuss strategies to validate MSMs built using potentially imperfect force-field models and possibly overcome some of the limitations.

9.3.9 Analysis of the Association–Dissociation Path Ensemble

The "mechanism" of binding is ultimately governed by the statistical distribution of different paths from the unbound state to the bound state. The importance of the different paths between the unbound and bound states is governed by the flux along that path. Transition path theory (TPT) [20, 37, 101, 102] provides us with a theoretical framework through which we can compute reactive flux-matrices from MSMs of protein–protein encounters. The "reaction" here refers to the transition from a set of "reactant states" (unbound) A to a set of "product states" B (bound).

TPT gives us tools to assign all intermediate states I (not bound, and not unbound), committor probabilities, q^+ and q^-, which gives us the probability of reaching the state B before A from an intermediate state $i \in I$ via forward committor q_i^+ and vice versa for the backward committor q_i^-. We can further use this framework to compute mean first passage times (MFPT), for example, the average on- and off-rates, as well as to dissect all the possible pathways from unbound to bound states. TPT is therefore an important tool for analysis of MSMs, in particularly when we want to understand a specific process. Major MSM softwares implement TPT analyses and plotting functions to visualize the results [34, 67, 68].

9.4 The Connection to Experiments

In favorable cases, appropriately validated MSMs predict molecular mechanisms with high temporal and spatial resolution. These insights can, of course, guide our understanding of important molecular phenomena associated with, for instance, protein–protein binding. However, so far, we have only discussed validation as statistical self-consistency and minimizing projection errors (optimizing variational scores). Since we generate the simulation data that we use to drive the estimation MSMs with imperfect classical empirical force field models, agreement with experimental data is not a given. In this section, I will outline how we can predict important biophysical observables to check for agreements, the limitations of these comparisons, and how we may integrate experimental data into the estimation of MSMs using the augmented Markov model framework, to bring experiment and simulation into alignment.

9.4.1 Experimental Observability, Forward Models, and Errors

What is an observable? – In our context, an experimental observable is a function of state; that means a function of the configurations adopted by a molecular system at specific experimental conditions. The definition encompasses both bulk experiments, where the cumulative signal of a very large number ($\sim 10^{23}$) of copies of identical systems are measured, and experiments where time-resolved trajectory signals from single molecules are measured. The manifestation of a particular observable is described by a physical model, $f(\cdot)$, describing the relationship between a configuration of the system, x, and the observed signal, o. Simple examples of f may be a ruler measuring the Euclidean distance between two atoms in a molecule, or a function computing the potential energy of the system configuration, x.

In a *stationary* bulk experiment at equilibrium we measure an expectation value of $f(\cdot)$ under the Boltzmann distribution,

$$\begin{aligned}\langle O_f \rangle &= \mathcal{Z}^{-1} \int f(x) \exp(-\beta U(x)) \, \mathrm{d}x \\ &= \int f(x) \mu(x) \, \mathrm{d}x \\ &= \mathbb{E}_{\mu(x)}[f(x)].\end{aligned}$$

We use $p(x)$ as shorthand for the normalized Boltzmann distribution for a given β, and the bracket to denote ensemble averages.

In an ergodic, *dynamic* bulk experiment at equilibrium we measure the autocorrelation,

$$\begin{aligned}\langle O_f(0)O_f(\Delta t)\rangle &= \iint f(x')p(x' \mid x;\Delta t)p(x)f(x)\,\mathrm{d}x\mathrm{d}x' \\ &= \mathbb{E}_{p(x)}[\mathbb{E}_{p(x'|x,\tau)}[f(x')f(x)]]\end{aligned}$$

where $p(x' \mid x;\Delta t)$ is the transition probability. Note that we may analogously define cross-correlation experiments by using two different models for the observables,

$$\langle O_{f_a}(0)O_{f_b}(\Delta t)\rangle = \mathbb{E}_{p(x)}[\mathbb{E}_{p(x'|x,\tau)}[f_b(x')f_a(x)]].$$

Some experimental setups will allow us to initialize an ensemble in a nonequilibrium ensemble p_0 and follow the relaxation process back to equilibrium [20, 37]. Such experiments include pressure and temperature jump, as well as stopped flow. Cross-correlation functions measured in such relaxation experiments can be expressed as

$$\langle O_{f_a}(0)O_{f_b}(\Delta t)\rangle_{p_0} = \iint f_b(x')p(x' \mid x;\Delta t)\frac{p_0(x)}{p(x)}f_a(x)\,\mathrm{d}x'\mathrm{d}x.$$

In single molecule experiments, observables are followed over time as trajectories, with some time resolution Δt, in its simplest form:

$$O_f = \{f(x(0)), f(x(\Delta t)), \ldots, f(x(N\Delta t))\}.$$

9.4.1.1 Sources of Errors and Uncertainty

In a typical setting, we have some set of experimental data, which may include any combination of the classes above, and we wish to compare these observables to the corresponding predictions made for stationary and dynamic properties of the molecular system represented by our computational model. Several sources of uncertainty and error may arise in this setting that we will have to be mindful of:

1. **Experimental noise (thermal noise, shot noise, etc.)**
   This category includes all contributions of stochastic noise due to limitations in the experimental setup, involving, e.g., imperfections in instrumentation measurements or sample (labeling) stability. In many cases, theoretical analyses of experiments are available, which may help decide how this error should be modeled.
2. **Systematic experimental errors/biases**
   These errors and biases arise due to imperfect referencing — in the case of, for example, relative experimental measurements — or unknown, or imprecise experimental conditions (temperature, concentrations, pressure, etc) or fluctuations of these parameters during data acquisition. This source of error is typically more challenging to systematically model or perfectly compensate for and often requires substantial knowledge of the experimental setup.

3. **Systematic error in the computational model of the molecular system dynamics**
   Computational models of $p(x)$ and $p(x' \mid x, \Delta t)$ are typically estimated using finite simulation data obtained from empirical force field models. The quantitative agreement of these with experiment is ever improving, however, still suffers significantly from systematic errors. The errors may arise from the classical approximations made of quantum mechanical interactions, to make simulations computationally tractable, or other approximations. Our efforts to compare simulation models to experiments are typically driven by the recognition of these issues, and a desire to understand the limitations and merits of a given moded. Another systematic error encompassed in this section is the sampling error, where we simply do not have enough simulation data to accurately estimate $p(x)$ and $p(x' \mid x, \Delta t)$.
4. **Modeling error of observable functions** $f(\cdot)$
   Like simulation models, the forward prediction of instantaneous (time-independent) experimental observables approximates complicated experimental setups or quantum mechanical phenomena in a computationally efficient manner. In many cases, quantifying errors and biases in these models is challenging.

9.4.2 Predicting Experimental Observables Using MSMs

The expressions given above for the experimental observables are general for any ergodic, Markovian dynamics, at, or relaxing to, a time-invariant equilibrium state. However, for molecular systems, these expressions involve intractable integrals over the configuration space. Fortunately, for MSMs, these integrals simplify to standard linear algebra operations, which we can compute efficiently.

The discretization of configurational space, $\mathcal{S} = \{S_1, \ldots, S_N\}$, associated with the MSM, leads to a discretization of the instantaneous experimental observable (predicted via $f(\cdot)$) as the vector $\mathbf{a} \in \mathbb{R}^N$, with elements,

$$a_i = \frac{1}{\int_{x \in S_i} p(x)\ \mathrm{d}x} \int_{x \in S_i} p(x) f_a(x)\ \mathrm{d}x \approx \frac{1}{N_{S_i}} \sum_{j \in S_i} f(x_j)$$

where S_i is a configurational space segment, or its finite sample approximation with N_{S_i} samples. We can extend this expression to vector-valued observables. Since this discretization replaces a function, which takes on arbitrary real valued numbers for different conformations, by a piecewise constant function, we hope to minimize the variance of $f(x)$ within each S_i to ensure a good approximation.

We use the stationary distribution π, of the transition matrix T_τ, along with discretized feature vector $\mathbf{a}$ to compute *stationary bulk experimental observables* as

$$\langle O_{f_a} \rangle = \pi \cdot \mathbf{a} = \sum_{i=1}^{N} \pi_i a_i.$$

The expression for *dynamic bulk experiments* can be formulated using the transition matrix,

$$\langle O_f(0) O_f(N\tau) \rangle = \mathbf{a}^\top \mathbf{\Pi} T(\tau)^N \mathbf{a}$$

where $\mathbf{\Pi} = \mathrm{diag}(\pi)$ is a matrix with stationary probabilities on the diagonal, and zeros elsewhere. N is an integer that expresses time in multiples of the MSM lag-time τ. Cross-correlation experiments can similarly be expressed as

$$\langle O_{f_a}(0) O_{f_b}(N\tau) \rangle = \mathbf{b}^\top \mathbf{\Pi} T(\tau)^N \mathbf{a}$$

where $\mathbf{b}$ is defined as $\mathbf{a}$, but for an observable predicted by $f_b(\cdot)$. In general, MSMs predict auto- and cross-correlation function as mixture of exponential decays. We see this more directly by considering the spectral decomposition of the transition matrix,

$$\begin{aligned}
\langle O_f(0) O_f(N\tau) \rangle &= \mathbf{a}^\top \mathbf{\Pi} \left(\sum_{i=1}^{N} \lambda_i^N \boldsymbol{l}_i^\top \mathbf{r}_i \right) \mathbf{a} \\
&= \sum_{i=1}^{N} \lambda_i^N \left(\mathbf{a}^\top \cdot \boldsymbol{l}_i^\top \right) \left(\boldsymbol{l}_i \cdot \mathbf{a} \right) \\
&= \sum_{i=1}^{N} \exp\left(-\frac{N\tau}{t_i} \right) \left(\boldsymbol{l}_i \cdot \mathbf{a} \right)^2 \\
&= (\pi \cdot \mathbf{a})^2 + \sum_{i=2}^{N} \left(\boldsymbol{l}_i \cdot \mathbf{a} \right)^2 \exp\left(-\frac{N\tau}{t_i} \right).
\end{aligned}$$

Similar expressions can be written down for cross-correlation and relaxation experiments [37, 104]. This simple form of the auto- and cross-correlation functions from MSMs facilitates analytical expression of several experimental observables including those from NMR spectroscopy, dynamic neutron scattering, and FRET spectroscopy.

9.4.3 Integrating Experimental and Simulation Data into Augmented Markov Models

As mentioned above, systematic errors in the empirical force field models used for molecular simulations to build MSMs lead to statistically robust yet systematic errors in our predictions. Using the equations above, we can quantify, but not remedy, these biases. A wealth of methods have been introduced to bias MD simulations [105–112], or reweight simulation data *a posteriori* [111, 113–119], to match experimental data, using different inference philosophies – several excellent reviews discuss these approaches in more detail [120–122]. In the context of MSMs, we already have simulation data available or are in the process of adaptively acquiring it. Consequently, adopting an approach that would alter the ensemble of our MD simulations is undesirable – excluding the use of experimental data to bias simulations. On the other hand, reweighing MD trajectories generally sacrifices the dynamic information from our simulation data.

The augmented Markov models (AMM) [115] framework allows us to balance experimental and simulation data when building Markov models of molecular kinetics. AMMs, therefore, achieve better agreement with experimental data while preserving the dynamic information from molecular simulations. To estimate AMMs the log-likelihood function of MSMs is augmented with a term to balance systematic

discrepancies between experimental and simulation data via a set of Lagrange multipliers λ,

$$\ell(T_\tau, \lambda \mid \mathbf{C}(\tau), \mathbf{O}, \boldsymbol{\sigma}) \propto \sum_{ij} c_{ij} \log t_{ij,\tau} - \sum_k \frac{(\hat{m}_k - o_k)^2}{2\sigma_k^2} \tag{9.2}$$

where $t_{ij,\tau}$ is the i,j'th element of T_τ, c_{ij} the corresponding element in the count matrix $\mathbf{C}(\tau)$, and o_k and σ_k are the k'th experimental observable and its experimental uncertainty respectively. The prediction of the experimental expectation value from T_τ is $\hat{m}_k = \mathbf{a}_k \cdot \hat{\pi}$ where

$$\hat{\pi}_i = \frac{\pi_i \exp(\sum_\nu \lambda_\nu a_{\nu_i})}{\sum_j \pi_j \exp(\sum_\nu \lambda_\nu a_{\nu_j})}$$

models the experimental Boltzmann distribution via a maximum entropy perturbation of the simulation ensemble π computed from T_τ. λ_ν is the ν'th Lagrange multiplier corresponding to experimental observable o_ν and its back-prediction for a Markov state i as a_{ν_i}. Optimizing Eq. (9.2) subject to detailed balance constraints yields an AMM.

We motivate the use of a maximum entropy perturbation as it provides a model as close as possible – in the Kullback-Leibler sense – to the simulation ensemble. The critical assumption is therefore that the simulation ensemble provides a reasonable starting point to model the experimental data, including covering all metastable configurations necessary to accurately predict the experimental observables.

Other approaches similarly allow for the integration of experimental data into MSM estimation. One approach enables the gradual adjustment of MSM stationary distributions against target observables [123]. Matsunaga and Sugita present a method to integrate single-molecule FRET data and molecular simulation using a stepwise HMM estimation procedure [124]. Brotzakis et al. propose a maximum entropy and maximum caliber approach to reweigh trajectory ensembles against bulk observables [116].

Although the integration of experimental data and simulation data has been an active area of research for several decades, several problems remain open. In particular, some data still cannot be included in the AMMs framework, including single-molecule FRET trajectories or dynamic bulk observables.

9.5 Protein–Protein and Protein–Peptide Encounters

Several groups have reported kinetic models of protein–protein and protein–peptide encounters using molecular dynamic simulations and MSMs. As yet, tightly binding complexes (small dissociation constants, K_D) with slow association–dissociation kinetics dominate the literature, as they constitute the biggest challenge for molecular simulations. While slow macroscopic kinetics and large free energy differences characterize these systems, microscopically, these protein–protein and protein–peptide encounters may happen via multistep processes. Consequently, we can sample rare events on the seconds to minutes timescale by connecting the

bound and unbound states via sampling the much more likely transitions between intermediate states. MSMs excel in cases such these: transitions sampled between intermediate steps along binding–unbinding paths can be combined into a model that predicts the slow macroscopic dynamics of the full binding–unbinding process inaccessible for direct simulation.

The first reported MSM study of a fully reversible protein–protein binding by all-atom molecular dynamic simulations was for the inhibitory complex of ribonuclease barnase and barstar. The barnase:barstar complex is an excellent benchmark system due to its extensive experimental charaterization by multiple biophysical methods. Plattner et al. collected molecular simulations with 2 ms of aggregate simulation length [27]. The data was distributed between 1.7 ms of independent simulations initiated from dissociated states and 0.3 ms using adaptive sampling. The adaptive scheme enabled the authors to sample barnase:barstar association with a few microseconds of aggregate simulation time, while the equilibrium binding rate is tens of microseconds. Unbinding similarly benefitted from adaptive sampling, with unbinding events being sampled in a few hundred microseconds, while the equilibrium off-rate is expected to be on the hours timescale. The authors estimated a HMM to compute the thermodynamics, kinetics, and important structural states of the protein–protein encounter process. They compared predictions of macroscopic thermodynamics and kinetic observables against experiment: binding free energy 12–19 kcal/mol against the experimental 16.8 kcal/mol and the dissociation rate 3×10^{-6} to 10^{-1} compared to the experimental range of 8×10^{-5} to 5.0×10^{-4} s^{-1}. The large uncertainties illustrate how MSMs and HMMs quality, accuracy, and precision for these parameters critically rely on the number of binding and unbinding events sampled in the aggregate simulation data. Nevertheless, the on-rate could be predicted with high accuracy and the most stable state coincided with the crystallographic structure (pdb: 1BRS). Further, perturbation theory allowed for accurate prediction of binding free energy changes upon mutation within statistical uncertainty. The resulting barnase:barstar HMM predicts a binding mechanism, where barstar can associate to all points of barnase's surface, early intermediates preferably bind opposite to the native binding groove, and late intermediates states and a "trap" state bind close to the binding grove, but in non-native orientations. Later still in the process, the complex passes through late intermediates into a prebound, loosely bound, and then finally the native bound state. The rate-limiting step is the prebound state that is stabilized by electrostatic and hydrophobic interactions between the two protein domains.

Two studies report MSMs of protein–peptide encounters involving the p53-antagonist MDM2 [29, 125]. The first study investigated one antagonistic pathway of MDM2 via its binding of the p53 transactivation domain (TAD) [125]. The study modeled this interaction via a TAD peptide and uses extensive, unbiased molecular dynamics simulations, as for the barnase:barstar study discussed above. The second study instead reported the binding of MDM2 to an inhibitory peptide PMI via integrating unbiased simulations and data from enhanced sampling [29]. The first study reported a MSM that predicts quantities with an accuracy comparable or worse than that observed for the barnase:barstar case above. Qualitatively

accurate on-rates, yet off-rates do not agree with experiments – this is likely due to the relatively small data set used here of 831 μs in aggregate length and force field inaccuracies. Nevertheless, the authors can identify important structural states and investigate possible binding mechanisms. Their model favors an induced fit binding mechanism, where TAD first binds MDM2 and then folds into the native complex structure.

For the second study [29], and in a follow-up study, the authors used multi-ensemble Markov models (MEMMs) to quantitatively predict binding thermodynamics and kinetics of MDM2 to the PMI peptide with high precision. MEMMs define MSMs over multiple thermodynamic states, such as those used in enhanced sampling techniques, including replica exchange and umbrella sampling. The advantage of this approach is that it relies on less simulation data (approximately 102 μs of Hamiltonian replica exchange and 500 μs of unbiased MD). This advantage depends on designing an effective enhanced sampling strategy for the system of interest, which may be challenging to achieve without substantial trial-and-error and extensive human intervention.

To summarize, MSMs and related kinetic modeling approaches are currently the only available strategy to gain comprehensive microscopic insights into the thermodynamics *and* kinetics of protein–protein and protein–peptide encounters. These analyses can help us to distinguish between different binding mechanisms, and, combined with perturbation approaches, offer qualitative insights into the influence of point mutations on the binding. However, collecting sufficient data remains a serious challenge when applying these methods in practice. At simulation rates of around 400–500 ns per day and GPU collecting millisecond-sized datasets may take GPU-years to complete. New adaptive sampling strategies and the integration of experimental data and enhanced sampling simulations may help to lower the demands on unbiased simulations. Nevertheless, studies have focused on relatively small protein–protein and protein–ligand systems and interactions with high affinity. Further improvements in computing power, simulation, and analysis methods are needed to ensure that these analyses can benefit structural biology more broadly. Finally, how these approaches will fare on low-affinity complex systems with a less clear separation of timescales also remain to be understood.

9.6 Emerging Technologies

As we saw above, Markov state models are emerging as an important tool in characterizing the thermodynamics and kinetics of protein–protein encounters, in ideal cases providing detailed mechanistic models at atomic resolution. We are steadily progressing towards better methods for featurization, dimension reduction, clustering, and adaptive sampling strategies; these advances contribute to minimizing the computational and labor effort needed to build high-quality MSMs. However, we remain reliant on access to state-of-the-art computing resources and, in many cases, the extensive manual intervention of highly skilled researchers. Simultaneously, our ambitions are increasing, and we want to study larger and larger systems—our

growing ambitions introduce two challenges: simulating enough to establish statistically sufficient models and increasingly expensive simulations. A broad range of machine learning methods is currently emerging that directly address the challenges faced by MSMs.

Recall, the fundamental use of MSMs is to build a low-dimensional approximation of an infinite-dimensional molecular dynamics operator. This task relies on several preprocessing steps in sequential succession: featurization, dimension reduction, clustering, and model estimation. The success at each stage depends on the careful adjustment of hyperparameters against an optimality criterion. An error or suboptimal choice made early on in the sequence may negatively impact our final model's quality. Yet, identifying and resolving such problems is often not straightforward and relies on extensive testing and manual intervention. With VAMPnets, Mardt et al. illustrate how we may, in principle, replace the entire sequence of preprocessing steps and the model estimation by an artificial neural network [126]. The key idea is to input all-atom coordinates into a neural network that outputs a categorical distribution representing a conformation membership to N metastable states. The optimization objective of the neural network is a VAMP score computed between pairs of simulation frames with a time lag of τ. The neural network learns the complicated function from molecular dynamics trajectories to a molecular kinetics model in a single step through this procedure. Extensions of VAMPnets are already emerging to improve data efficiency and impose further constraints on the modeled dynamics.

The number of states a molecular system may potentially adopt grows exponentially with its size. So, in addition to declining simulation rates with system size, we must sample a much larger conformational space. In other words, our simulations get slower, and we have to simulate more to ensure statistically sufficient models. The latter of these two problems arises from how we represent the metastable states as global configurations. Consequently, we need to explicitly account for every metastable state, even if differences between these states are only minor structural changes.

Dynamic graphical models (DGM) replace the global representation of metastable states with local sub-systems [127]. These subsystems are spatially localized and can be a sidechain rotamer or a whole protein domain. A DGM aims to encode the conformational states of all the subsystems and how they influence each other's evolution in time. Like MSMs, DGMs approximate the transition probability densities between all possible configurations of our subsystems, yet without the need to enumerate them all explicitly. Their indirect representation of global configurations allows DGMs to rely on fewer parameters than MSMs, lowering simulation data demands. A recent study shows that this strategy can be very effective when modeling molecular dynamics, quantitatively predicting the thermodynamics and kinetics of molecular systems [127]. As DGMs are generative models, we may also predict realistic metastable states not seen during the model's estimation.

MSMs have come a long way. We now see regular application of this methodology to the quantitative study of complex problems such as protein folding, protein–protein interactions, and conformational dynamics. The growing community of

researchers, together with making improvements in simulation methodology and hardware, is rapidly expanding the scope of systems we can address. With the advent of powerful machine learning we can expect to see these developments accelerate further. These developments and their extensions bode well for Markovian models' future in the quantitative study of protein–protein encounters.

Acknowledgments

I thank Dr. Rocío Mercado for comments and input on drafts of this chapter. I thank Lillian Chong and Kresten Lindorff–Larsen for feedback on an early preprint. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1 Nooren, I.M. (2003). Diversity of protein–protein interactions. *EMBO J.* 22 (14): 3486–3492.
2 Czech, M.P. (1985). The nature and regulation of the insulin receptor: structure and function. *Annu. Rev. Physiol.* 47 (1): 357–381.
3 Zinzalla, G. and Thurston, D.E. (2009). Targeting protein–protein interactions for therapeutic intervention: a challenge for the future. *Fut. Med. Chem.* 1 (1): 65–93.
4 Blikstad, C. and Ivarsson, Y. (2015). High-throughput methods for identification of protein–protein interactions involving short linear motifs. *Cell Commun. Signal.* 13 (1): 38.
5 Gavin, A.-C., Aloy, P., Grandi, P. et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440 (7084): 631–636.
6 Krogan, N.J., Cagney, G., Yu, H. et al. (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440 (7084): 637–643.
7 Chakrabarti, K., Agafonov, R., Pontiggia, F. et al. (2016). Conformational selection in a protein–protein interaction revealed by dynamic pathway analysis. *Cell Rep.* 14 (1): 32–42.
8 Weikl, T.R. and Paul, F. (2014). Conformational selection in protein binding and function. *Protein Sci.* 23 (11): 1508–1518.
9 Palmer, A.G. (2004). NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* 104 (8): 3623–3640.
10 Min, W., English, B.P., Luo, G. et al. (2005). Fluctuating enzymes: lessons from single-molecule studies. *Acc. Chem. Res.* 38 (12): 923–931.
11 Frank, J. (2018). New opportunities created by single-particle cryo-EM: the mapping of conformational space. *Biochemistry* 57 (6): 888.
12 Boehr, D.D., Nussinov, R., and Wright, P.E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* 5 (11): 789–796.

13 Dementiev, A., Dobó, J., and Gettins, P.G. (2006). Active site distortion is sufficient for proteinase inhibition by serpins. *J. Biol. Chem.* 281 (6): 3452–3457.

14 Huntington, J.A., Read, R.J., and Carrell, R.W. (2000). Structure of a serpin–protease complex shows inhibition by deformation. *Nature* 407 (6806): 923–926.

15 Kim, E., Lee, S., Jeon, A. et al. (2013). A single-molecule dissection of ligand binding to a protein with intrinsic dynamics. *Nat. Chem. Biol.* 9 (5): 313–318.

16 Sugase, K., Dyson, H.J., and Wright, P.E. (2007). Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* 447 (7147): 1021–1025.

17 Bezsonova, I., Bruce, M.C., Wiesner, S. et al. (2008). Interactions between the three CIN85 SH3 domains and ubiquitin: implications for CIN85 ubiquitination†. *Biochemistry* 47 (34): 8937–8949.

18 Purslow, J.A., Khatiwada, B., Bayro, M.J., and Venditti, V. (2020). NMR methods for structural characterization of protein–protein complexes. *Front. Mol. Biosci.* 7: 9.

19 Raich, L., Meier, K., Günther, J. et al. (2021). Discovery of a hidden transient state in all bromodomain families. *Proc. Natl. Acad. Sci. U.S.A.* 118 (4): e2017427118.

20 Noé, F., Schütte, C., Vanden-Eijnden, E. et al. (2009). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U.S.A.* 106 (45): 19011–19016.

21 Barros, E.P., Demir, O., Soto, J. et al. (2021). Markov state models and NMR uncover an overlooked allosteric loop in p53. *Chem. Sci.* 12 (5): 1891–1900.

22 Lindorff-Larsen, K., Piana, S., Dror, R.O., and Shaw, D.E. (2011). How fast-folding proteins fold. *Science* 334 (6055): 517–520.

23 Shaw, D.E., Maragakis, P., Lindorff-Larsen, K. et al. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science* 330 (6002): 341–346.

24 Stelzl, L.S., Mavridou, D.A., Saridakis, E. et al. (2020). Local frustration determines loop opening during the catalytic cycle of an oxidoreductase. *ELIFE* 9: e54661.

25 Pietrucci, F., Marinelli, F., Carloni, P., and Laio, A. (2009). Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations. *J. Am. Chem. Soc.* 131 (33): 11811–11818.

26 Stanley, N., Esteban-Martín, S., and Fabritiis, G.D. (2014). Kinetic modulation of a disordered protein domain by phosphorylation. *Nat. Commun.* 5 (1): 5272.

27 Plattner, N., Doerr, S., Fabritiis, G.D., and Noé, F. (2017). Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* 9 (10): 1005–1011.

28 Pan, A.C., Jacobson, D., Yatsenko, K. et al. (2019). Atomic-level characterization of protein–protein association. *Proc. Natl. Acad. Sci. U.S.A.* 116 (10): 4244–4249.

29 Paul, F., Wehmeyer, C., Abualrous, E.T. et al. (2017). Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations. *Nat. Commun.* 8 (1): 1095.

30 Saglam, A.S. and Chong, L.T. (2019). Protein–protein binding pathways and calculations of rate constants using fully-continuous, explicit-solvent simulations. *Chem. Sci.* 10 (8): 2360–2372.

31 Piana, S., Lindorff-Larsen, K., and Shaw, D.E. (2013). Atomistic description of the folding of a dimeric protein. *J. Phys. Chem. B* 117 (42): 12935–12942.

32 Prinz, J.-H., Wu, H., Sarich, M. et al. (2011). Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* 134 (17): 174105.

33 Husic, B.E. and Pande, V.S. (2018). Markov state models: from an art to a science. *J. Am. Chem. Soc.* 140 (7): 2386–2396.

34 Wehmeyer, C., Scherer, M.K., Hempel, T. et al. (2019). Introduction to Markov state modeling with the PyEMMA software [article v1.0]. *Living J. Comput. Mol. Sci.* 1 (1): 5965.

35 Olsson, S. and Noé, F. (2016). Mechanistic models of chemical exchange induced relaxation in protein NMR. *J. Am. Chem. Soc.* 139 (1): 200–210.

36 Noe, F., Doose, S., Daidone, I. et al. (2011). Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proc. Natl. Acad. Sci. U.S.A.* 108 (12): 4822–4827.

37 Prinz, J.-H., Keller, B., and Noé, F. (2011). Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.* 13 (38): 16912.

38 Noé, F. (2008). Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.* 128 (24): 244103.

39 Huber, T., Torda, A.E., and van Gunsteren, W.F. (1994). Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J. Comput.-Aided Mol. Des.* 8 (6): 695–708.

40 Grubmüller, H. (1995). Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E* 52 (3): 2893–2906.

41 Laio, A. and Parrinello, M. (2002). Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* 99 (20): 12562–12566.

42 Rohrdanz, M.A., Zheng, W., and Clementi, C. (2013). Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annu. Rev. Phys. Chem.* 64 (1): 295–316.

43 Lange, O.F. and Grubmüller, H. (2006). Collective langevin dynamics of conformational motions in proteins. *J. Chem. Phys.* 124 (21): 214903.

44 Schütte, C., Fischer, A., Huisinga, W., and Deuflhard, P. (1999). A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.* 151 (1): 146–168.

45 Schütte, C., Noé, F., Lu, J. et al. (2011). Markov state models based on milestoning. *J. Chem. Phys.* 134 (20): 204105.

46 Bowman, G.R., Pande, V.S., and Noe, F. (eds) (2014). *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Springer Netherlands.

47 Buchete, N.-V. and Hummer, G. (2008). Coarse master equations for peptide folding dynamics†. *J. Phys. Chem. B* 112 (19): 6057–6069.

48 Swope, W.C., Pitera, J.W., and Suits, F. (2004). Describing protein folding kinetics by molecular dynamics simulations. 1. Theory†. *J. Phys. Chem. B* 108 (21): 6571–6581.

49 Doerr, S., Harvey, M.J., Noé, F., and Fabritiis, G.D. (2016). HTMD: High-throughput molecular dynamics for molecular discovery. *J. Chem. Theory Comput.* 12 (4): 1845–1852.

50 Sriraman, S., Kevrekidis, I.G., and Hummer, G. (2005). Coarse master equation from Bayesian analysis of replica molecular dynamics simulations†. *J. Phys. Chem. B* 109 (14): 6479–6484.

51 Zwanzig, R. (1983). From classical dynamics to continuous time random walks. *J. Stat. Phys.* 30 (2): 255–262.

52 Rao, F. and Caflisch, A. (2004). The protein folding network. *J. Mol. Biol.* 342 (1): 299–306.

53 Shirts, M. and Pande, V.S. (2000). Screen savers of the world unite!. *Science* 290 (5498): 1903–1904.

54 Chodera, J.D., Singhal, N., Pande, V.S. et al. (2007). Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* 126 (15): 155101.

55 Nüske, F., Keller, B.G., Pérez-Hernández, G. et al. (2014). Variational approach to molecular kinetics. *J. Chem. Theory Comput.* 10 (4): 1739–1752.

56 Noé, F. and Nüske, F. (2013). A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model. Simul.* 11 (2): 635–655.

57 Wu, H. and Noé, F. (2019). Variational approach for learning Markov processes from time series data. *J. Nonlinear Sci.* 30 (1): 23–66.

58 Scherer, M.K., Husic, B.E., Hoffmann, M. et al. (2019). Variational selection of features for molecular kinetics. *J. Chem. Phys.* 150 (19): 194108.

59 Chen, Q., Feng, J., Mittal, S., and Shukla, D. (2018). Automatic feature selection in Markov state models using genetic algorithm. *J. Comput. Sci. Educ.* 9 (2): 14–22.

60 Sittel, F., Jain, A., and Stock, G. (2014). Principal component analysis of molecular dynamics: on the use of Cartesian vs. internal coordinates. *J. Chem. Phys.* 141 (1): 014111.

61 García, A.E. (1992). Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* 68 (17): 2696–2699.

62 Ichiye, T. and Karplus, M. (1991). Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins Struct. Funct. Genet.* 11 (3): 205–217.

63 de Groot, B.L., Daura, X., Mark, A.E., and Grubmüller, H. (2001). Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds. *J. Mol. Biol.* 309 (1): 299–313.

64 Pérez-Hernández, G., Paul, F., Giorgino, T. et al. (2013). Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* 139 (1): 015102.

65 Schwantes, C.R. and Pande, V.S. (2013). Improvements in Markov state model construction reveal many nonnative interactions in the folding of NTL9. *J. Chem. Theory Comput.* 9 (4): 2000–2009.

66 Schwantes, C.R. and Pande, V.S. (2015). Modeling molecular kinetics with tICA and the kernel trick. *J. Chem. Theory Comput.* 11 (2): 600–608.

67 Scherer, M.K., Trendelkamp-Schroer, B., Paul, F. et al. (2015). PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* 11 (11): 5525–5542.

68 Harrigan, M.P., Sultan, M.M., Hernández, C.X. et al. (2017). MSMBuilder: Statistical models for biomolecular dynamics. *Biophys. J.* 112 (1): 10–15.

69 Keller, B., Daura, X., and van Gunsteren, W.F. (2010). Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* 132 (7): 074110.

70 Bowman, G.R., Beauchamp, K.A., Boxer, G., and Pande, V.S. (2009). Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* 131 (12): 124101.

71 Trendelkamp-Schroer, B., Wu, H., Paul, F., and Noé, F. (2015). Estimation and uncertainty of reversible Markov models. *J. Chem. Phys.* 143 (17): 174101.

72 Trendelkamp-Schroer, B. and Noé, F. (2013). Efficient Bayesian estimation of Markov model transition matrices with given stationary distribution. *J. Chem. Phys.* 138 (16): 164113.

73 Husic, B.E. and Pande, V.S. (2017). Note: MSM lag time cannot be used for variational model selection. *J. Chem. Phys.* 147 (17): 176101.

74 Feng, H., Costaouec, R., Darve, E., and Izaguirre, J.A. (2015). A comparison of weighted ensemble and Markov state model methodologies. *J. Chem. Phys.* 142 (21): 214113.

75 Suárez, E., Adelman, J.L., and Zuckerman, D.M. (2016). Accurate estimation of protein folding and unfolding times: beyond Markov state models. *J. Chem. Theory Comput.* 12 (8): 3473–3481.

76 Suárez, E., Wiewiora, R.P., Wehmeyer, C. et al. (2021). What Markov state models can and cannot do: correlation versus path-based observables in protein folding models. *J. Chem. Theory Comput.* 17 (5): 3119–3133.

77 Deuflhard, P. and Weber, M. (2005). Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.* 398: 161–184.

78 Kube, S. and Weber, M. (2007). A coarse graining method for the identification of transition rates between molecular conformations. *J. Chem. Phys.* 126 (2): 024103.

79 Röblitz, S. and Weber, M. (2013). Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Adv. Data Anal. Classif.* 7 (2): 147–179.

80 Noé, F., Wu, H., Prinz, J.-H., and Plattner, N. (2013). Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.* 139 (18): 184114.

81 Hummer, G. and Szabo, A. (2014). Optimal dimensionality reduction of multistate kinetic and Markov-state models. *J. Phys. Chem. B* 119 (29): 9029–9037.

82 Gerber, S., Olsson, S., Noé, F., and Horenko, I. (2018). A scalable approach to the computation of invariant measures for high-dimensional Markovian systems. *Scientific Reports* 8 (1): 1796.

83 Zimmerman, M.I. and Bowman, G.R. (2015). FAST conformational searches by balancing exploration/exploitation trade-offs. *J. Chem. Theory Comput.* 11 (12): 5747–5757.

84 Hruska, E., Abella, J.R., Nüske, F. et al. (2018). Quantitative comparison of adaptive sampling methods for protein dynamics. *J. Chem. Phys.* 149 (24): 244119.

85 Bowman, G.R., Ensign, D.L., and Pande, V.S. (2010). Enhanced modeling via network theory: adaptive sampling of Markov state models. *J. Chem. Theory Comput.* 6 (3): 787–794.

86 Zimmerman, M.I., Porter, J.R., Sun, X. et al. (2018). Choice of adaptive sampling strategy impacts state discovery, transition probabilities, and the apparent mechanism of conformational changes. *J. Chem. Theory Comput.* 14 (11): 5459–5475.

87 Wang, W., Cao, S., Zhu, L., and Huang, X. (2017). Constructing Markov state models to elucidate the functional conformational changes of complex biomolecules. *WIRES Comput. Mol. Sci.* 8 (1).

88 Hruska, E., Balasubramanian, V., Lee, H. et al. (2020). Extensible and scalable adaptive sampling on supercomputers. *J. Chem. Theory Comput.* 16 (12): 7915–7925.

89 Doerr, S. and Fabritiis, G.D. (2014). On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J. Chem. Theory Comput.* 10 (5): 2064–2069.

90 Pérez, A., Herrera-Nieto, P., Doerr, S., and Fabritiis, G.D. (2020). AdaptiveBandit: A multi-armed bandit framework for adaptive sampling in molecular simulations. *J. Chem. Theory Comput.* 16 (7): 4685–4693.

91 Chong, L.T., Saglam, A.S., and Zuckerman, D.M. (2017). Path-sampling strategies for simulating rare events in biomolecular systems. *Curr. Opin. Struct. Biol.* 43: 88–94.

92 Zuckerman, D.M. and Chong, L.T. (2017). Weighted ensemble simulation: review of methodology, applications, and software. *Annu. Rev. Biophys.* 46 (1): 43–57.

93 Ahn, S.-H., Jagger, B.R., and Amaro, R.E. (2020). Ranking of ligand binding kinetics using a weighted ensemble approach and comparison with a multiscale milestoning approach. *J. Chem. Inf. Model.* 60 (11): 5340–5352.

94 Zhou, R. Replica exchange molecular dynamics method for protein folding simulation. (eds. Yawen Bai, Ruth Nussinov) In: *Protein Folding Protocols*, 205–224. Humana Press.

95 Kumar, S., Rosenberg, J.M., Bouzida, D. et al. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* 13 (8): 1011–1021.

96 Chodera, J.D., Swope, W.C., Noé, F. et al. (2011). Dynamical reweighting: improved estimates of dynamical properties from simulations at multiple temperatures. *J. Chem. Phys.* 134 (24): 244107.

97 Wu, H., Paul, F., Wehmeyer, C., and Noé, F. (2016). Multiensemble Markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci. U.S.A.* 113 (23): E3221–E3230.

98 Mey, A.S., Wu, H., and Noé, F. (2014). xTRAM: Estimating equilibrium expectations from time-correlated simulation data at multiple thermodynamic states. *Phys. Rev. X* 4 (4): 041018.

99 Rosta, E. and Hummer, G. (2014). Free energies from dynamic weighted histogram analysis using unbiased Markov state model. *J. Chem. Theory Comput.* 11 (1): 276–285.

100 Stelzl, L.S., Kells, A., Rosta, E., and Hummer, G. (2017). Dynamic histogram analysis to determine free energies and rates from biased simulations. *J. Chem. Theory Comput.* 13 (12): 6328–6342.

101 Weinan, E. and Vanden-Eijnden, E. (2006). Towards a theory of transition paths. *J. Stat. Phys.* 123 (3): 503–523.

102 Metzner, P., Schütte, C., and Vanden-Eijnden, E. (2009). Transition path theory for Markov jump processes. *Multiscale Model. Simul.* 7 (3): 1192–1219.

103 Best, R.B., Zheng, W., and Mittal, J. (2014). Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* 10 (11): 5113–5124.

104 Keller, B.G., Prinz, J.-H., and Noé, F. (2012). Markov models and dynamical fingerprints: unraveling the complexity of molecular kinetics. *Chem. Phys.* 396: 92–107.

105 Olsson, S., Frellsen, J., Boomsma, W. et al. (2013). Inference of structure ensembles of flexible biomolecules from sparse, averaged data. *PLoS ONE* 8 (11): e79439, N. Fernandez-Fuentes, Ed.

106 Olsson, S., Vögeli, B.R., Cavalli, A. et al. (2014). Probabilistic determination of native state ensembles of proteins. *J. Chem. Theory Comput.* 10 (8): 3484–3491.

107 Bonomi, M., Camilloni, C., Cavalli, A., and Vendruscolo, M. (2016). Metainference: a Bayesian inference method for heterogeneous systems. *Sci. Adv.* 2 (1): e1501177.

108 Cavalli, A., Camilloni, C., and Vendruscolo, M. (2013). Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* 138 (9): 094112.

109 Pitera, J.W. and Chodera, J.D. (2012). On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* 8 (10): 3445–3451.

110 White, A.D. and Voth, G.A. (2014). Efficient and minimal method to bias molecular simulations with experimental data. *J. Chem. Theory Comput.* 10 (8): 3023–3030.

111 Hummer, G. and Köfinger, J. (2015). Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* 143 (24): 243150.

112 Olsson, S., Ekonomiuk, D., Sgrignani, J., and Cavalli, A. (2015). Molecular dynamics of biomolecules through direct analysis of dipolar couplings. *J. Am. Chem. Soc.* 137 (19): 6270–6278.

113 Olsson, S., Strotz, D., Vögeli, B. et al. (2016). The dynamic basis for signal propagation in human Pin1-WW. *Structure* 24 (9): 1464–1475.

114 Bottaro, S., Bengtsen, T., and Lindorff-Larsen, K. (2020). Integrating molecular simulation and experimental data: a Bayesian/maximum entropy reweighting approach. (ed. Zoltán Gáspári) In: *Methods in Molecular Biology*, 219–240. Springer US.

115 Olsson, S., Wu, H., Paul, F. et al. (2017). Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc. Natl. Acad. Sci. U.S.A.* 114 (31): 8265–8270.

116 Brotzakis, Z.F., Vendruscolo, M., and Bolhuis, P.G. (2020). A method of incorporating rate constants as kinetic constraints in molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* 118 (2): e2012423118.

117 Ge, Y. and Voelz, V.A. (2018). Model selection using BICePs: a Bayesian approach for force field validation and parameterization. *J. Phys. Chem. B* 122 (21): 5610–5622.

118 Salvi, N., Abyzov, A., and Blackledge, M. (2016). Multi-timescale dynamics in intrinsically disordered proteins from NMR relaxation and molecular simulation. *J. Phys. Chem. Lett.* 7 (13): 2483–2489.

119 Kümmerer, F., Orioli, S., Harding-Larsen, D. et al. (2020). Fitting side-chain NMR relaxation data using molecular simulations. *J. Chem. Theory Comput.* 17 (8): 5262–5275.

120 Orioli, S., Larsen, A.H., Bottaro, S., and Lindorff-Larsen, K. (2020). How to learn from inconsistencies: integrating molecular simulations with experimental data. (eds. Birgit Strodel, Bogdan Barz) In: *Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly*, 123–176. Elsevier.

121 Boomsma, W., Ferkinghoff-Borg, J., and Lindorff-Larsen, K. (2014). Combining experiments and simulations using the maximum entropy principle. *PLoS Comput. Biol.* 10 (2): e1003406, M. Levitt, Ed.

122 Bottaro, S. and Lindorff-Larsen, K. (2018). Biophysical experiments and biomolecular simulations: a perfect match?. *Science* 361 (6400): 355–360.

123 Rudzinski, J.F., Kremer, K., and Bereau, T. (2016). Communication: consistent interpretation of molecular simulation kinetics using Markov state models biased with external information. *J. Chem. Phys.* 144 (5): 051102.

124 Matsunaga, Y. and Sugita, Y. (2018). Linking time-series of single-molecule experiments with molecular dynamics simulations by machine learning. *ELIFE* 7: e32668.

125 Zhou, G., Pantelopulos, G.A., Mukherjee, S., and Voelz, V.A. (2017). Bridging microscopic and macroscopic mechanisms of p53-MDM2 binding with kinetic network models. *Biophys. J.* 113 (4): 785–793.

126 Mardt, A., Pasquali, L., Wu, H., and Noé, F. (2018). VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* 9 (1): 5.

127 Olsson, S. and Noé, F. (2019). Dynamic graphical models of molecular kinetics. *Proc. Natl. Acad. Sci. U.S.A.* 116 (30): 15001–15006.

10

Transcription Factor – DNA Complexes

Volkhard Helms

Saarland University, Center for Bioinformatics, Saarland Informatics Campus, Postfach 15 11 50, 66041 Saarbrücken, Germany

10.1 Introduction

Specific binding of proteins to DNA is a central step in many important processes in biological cells. Many different types of proteins can bind specifically to DNA involving transcription factors (TFs) with activating or repressive effects on gene expression, enzymes that pack or unpack chromatin structure, enzymes functioning in DNA repair or that place or remove chemical (epigenetic) modifications on and from DNA, topoisomerases that contribute DNA supercoiling or to unzip double-stranded DNA, etc. In this chapter, we will focus on the binding of eukaryotic TFs to DNA. Chapter 11 by Fischle and coworkers will discuss further enzymes involved in epigenetic processes.

Generally, two types of physicochemical interactions contribute to stabilizing protein–DNA interactions. On the one hand, protein–DNA contacts always involve an underlying electrostatic attraction between positively charged amino acids that are enriched at the protein-binding interface and the negatively charged phosphate backbone of DNA. As these interactions only involve atoms of the DNA backbone, they are not sequence specific. On the other hand, specific polar as well as nonpolar contacts are formed between some nucleotide bases of the DNA-binding motif and protein residues.

As TFs one generally considers proteins that may bind to DNA in a sequence-specific manner and regulate transcription [1]. To this end, eukaryotic TFs always contain one of about 100 known eukaryotic DNA-binding domains that are cataloged in the databases Pfam [2], SMART [3], or Interpro [4] together with hidden Markov models characteristic for these domain families. TFs usually contain at least one further structural domain, often an activation or effector domain that is sensitive to environmental or cellular conditions, such as the concentration of ions or cyclic AMP [5], and that may also bind to further proteins. For human, a recent compendium cataloged 924 effector domains belonging to 594 human TFs [6]. In this manner, the TFs enable fine-tuned regulation of gene expression depending on the respective condition of the biological cell. In eukaryotes, TF-binding motifs are

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

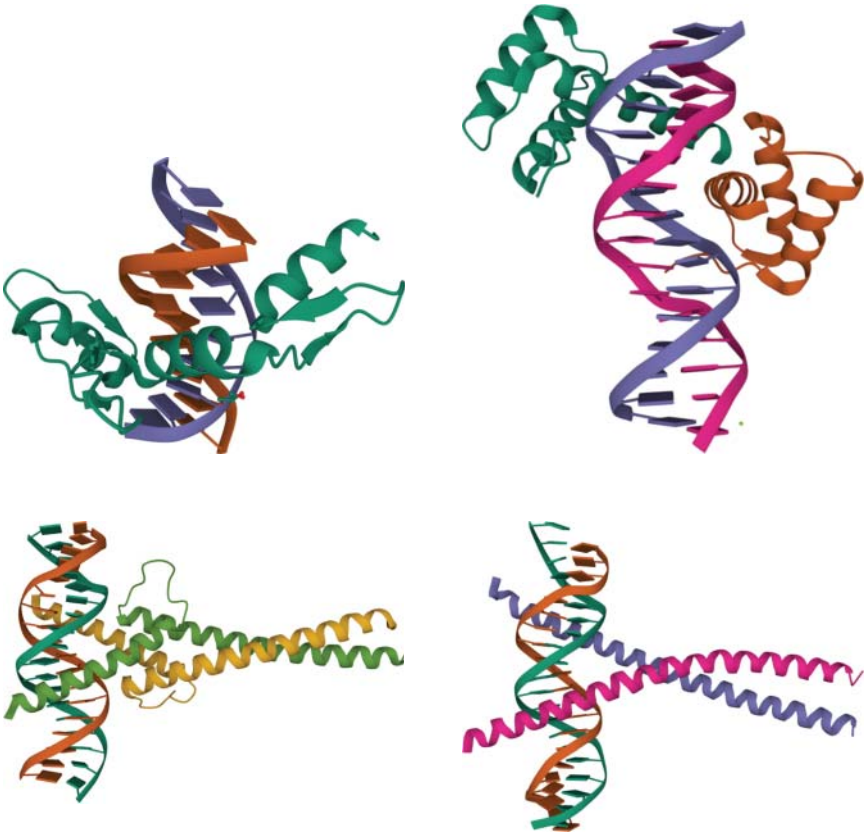


Figure 10.1 Top row: (left) Kruppel-like factor 4 (KLF4) bound to methylated DNA. KLFs belong to the family of C2H2-type Zinc-finger transcription factors, PDB code 4M9E, (right) WUS- Homeodomain from *Arabidopsis thaliana* bound to TGAA DNA. PDB-code 6YRD. Bottom row: (left) the Mad-Max dimer belongs to the family of basic helix-loop-helix TFs, PDB-code 1NLW. (right) C/EBP basic leucine zipper, PDB code 1NWQ. Source: Adapted from Sehnal et al. [8].

found both in the enhancer and promoter regions of their target genes. TFs contribute to either attracting or repelling RNA polymerase to or from a transcription start site so that the expression of the respective gene is either up- or downregulated.

Over the past decades, atomistic structures of many TFs could be determined both by X-ray crystallography and NMR [7]. Figure 10.1 illustrates several frequently observed structural topologies of eukaryotic TFs [9, 10], namely, the C2H2-zinc-finger proteins (ZF), basic helix-loop-helix (Bhlh) proteins, basic leucine zipper (Bzip), and the homeodomain.

Most organisms possess between a few hundred and a few thousand TFs. An upper estimate is set by identifying all protein-coding genes in the respective genome that have a DNA-binding domain. As just described, only a certain fraction of these proteins are TFs. For example, the genome of *Saccharomyces cerevisiae* contains 245 genes with known DNA-binding domains (about 4% of all

yeast genes). In the human genome, about 2600 genes contain one of the known DNA-binding domains (representing 11.8% of all human genes). Out of these, about three-quarters have been annotated with TF functionality [1]. Sonawane et al. analyzed the expression of a subset of 644 human TFs in 38 different tissue types sampled by the GTEx consortium [11]. They found that about one-third of them showed tissue-specific expression, which is a lower fraction than the average of all genes (41.6%). Hence, they suggested that TF expression is not the primary driver of tissue-specific functions. Besides recruiting RNA polymerase and thus promoting active transcription, TFs can also directly negatively affect transcription by blocking other proteins from binding to the same site. In most cases, eukaryotic TFs appear to recruit further cofactors as coactivators or corepressors.

Malfunction of transcription is known as one important driver of diseases. We will illustrate this here briefly on the example of the transcription factor p53, which is known as the "guardian of the cell." In case of DNA damage, p53 stops progression of the cell cycle and stimulates DNA repair processes. If the damages are too severe, p53 induces apoptosis. An X-ray structure of the p53 core domain bound to ds-DNA determined by the group of Nicola Pavletich revealed specific hydrogen bonding between several arginine residues of the protein and DNA bases (PDB-code 1TSR) [12]. Three of these arginine residues at the binding interface are indeed among the most frequently mutated p53 residues in cancer. Hence, the structural view of the binding mode perfectly matches the findings from mutagenesis analysis. Further examples of tumor mutations affecting transcription factors are described in detail in ref. [13].

10.2 Principles of Sequence Recognition

Figure 10.1 illustrates that many TF families interact with DNA via α-helices at their binding interfaces. The overall shape and dimensions of an α-helix can be accommodated in the major groove of DNA in multiple different ways. As mentioned before, establishing sequence-specific contacts with the DNA bases enables recognition of specific sequence motifs. Structural studies showed that contacts may involve direct hydrogen bonds between protein side chains and nucleic bases or between the polypeptide backbone and the bases, hydrogen bonds bridged by mediating water molecules, as well as hydrophobic contacts. About half of all hydrogen bonds involve atoms of the DNA backbone.

Subsequent to the seminal theoretical work of Berg et al. [14], experimental as well as computational and theoretical work showed that TFs find their specific binding motifs on the DNA by a mixture between times of free diffusion around the DNA strand and times where they bind nonspecifically to DNA. In the latter case, they may also slide along (scan) the DNA chain. Among the different options of either three-dimensional diffusion (jumping regime), one-dimensional motion (random-walk regime), or a combination of 3D and 1D motions (sliding regime), a theoretical analysis found that an optimal search dynamics results when the TFs explore both 1D and 3D pathways during the search for optimal-binding sites

on the DNA [15]. Munoz and colleagues recently showed by fluorescence correlation spectroscopy [16] that the transcription factor Engrailed from *Drosophila melanogaster* tracks its specific-binding sites on the DNA with the help of "DNA antennas" in the vicinity of these binding sites. The initial puzzling finding was that introducing severe mutations of the TAATTA recognition motif resulted in much smaller drops of the binding affinity than expected. This discrepancy could be resolved by assuming that EngHD also binds promiscuously to the flanking DNA sequence, thereby buffering the effect of mutations or even removal of the binding motif. This nonspecific buffering then showed an expected dependence on the ionic strength of the solution.

10.3 Dimerization of Eukaryotic TFs

The expression of a particular gene is typically regulated by the binding of multiple TFs whereby their binding sequence motifs are organized into so-called cis-regulatory modules. The simplest combinatorial element is the formation of homotypic dimers by many TFs of the bZIP, bHLH, MADS-box, NR, STATs, HD-ZIP, and NF-κB families [17]. In this way, TFs are able to form dimers having distinct biological properties (more than 500 dimers in humans and up to 2500 when alternative splicing is taken into consideration). However, as one TF monomer can in effect have several binding partners, it can form TF dimers with diverse properties and distinct regulatory effects. Which one of these dimers is established often depends on the levels of posttranslational modifications of the TFs and their binding affinities with each other. As an example, we will mention the pair of heterodimers Myc-Max and Mad-Max that both involve the ubiquitously expressed protein Max [17, 18]. When a Myc–Max heterodimer is assembled at the promoter element of one of its target genes, it will either recruit the SWI/SNF nucleosome remodeling complex or histone acetyltransferases (HATs). The SWI/SNF complex is known to break up the nucleosome structure, whereas HATs acetylate strongly conserved lysine residues in the disordered N-terminal tails of histone proteins. In both cases, the binding motifs of further TFs become exposed in the promoter region of the target gene. Instead, binding of a Mad–Max heterodimer will downregulate the target genes by recruiting histone deacetylases (HDACs) to the promoter element where it is bound. HDACs obviously play an inverse role to HATs.

In principle, N genes of a given TF family can give rise to N homodimers + $N\ (N-1)/2$ unique heterodimers, whereby we neglect potentially cooperative binding of the monomers, any cell-specific expression patterns, or splicing effects. Thus, the 51 human bZIPs, 118 bHLHs, and 48 NRs existing in humans could in principle form 1326, 7021, and 1176 unique dimers, respectively. Due to practical constraints of the involved binding interfaces, however, only certain specific monomer–monomer binding options are actually realized. For example, results from protein-array experiments led to an estimate of ∼350 unique bZIP dimers that are being formed, which is roughly a quarter of the theoretically possible combinations [17].

10.4 Detection of Epigenetic Modifications

It has been shown in mammals and plants that chromatin conformation is closely linked to the degree of CpG methylation [19] and the nature of epigenetic modifications placed on the nucleosomal histone proteins [20]. As illustrated in Figure 10.1, TFs typically pack tightly either into the major or minor groove of double-stranded DNA. As they have a non-negligible volume, it is reasonable to assume that TFs can best bind to dsDNA adopting an open chromatin conformation.

However, there also exist so-called pioneer transcription factors that may bind to condensed chromatin as well [21]. Taipale and colleagues systematically explored interactions between the nucleosome and 220 TFs representing diverse structural families by an adapted SELEX strategy [22]. In agreement with earlier findings, they observed that most of the studied TFs cannot access DNA that is wound around nucleosomes as easily as free DNA. The binding motifs derived from TFs that were either bound to nucleosomal or to free DNA were overall quite similar. However, in the steric context of the nucleosome, sequence motifs are only accessible from certain angles and locations, which places certain restrictions on the positioning of binding motifs. In fact, many pioneer TFs were found to bind near the ends of nucleosomal DNA. For example, competitive nucleosome-binding assays showed that the pioneer factor TP53 has a strong preference for sites outside the 100 bp region surrounding the nucleosome dyad [23].

Besides its effect on DNA conformation [24] that indirectly affects TF binding, DNA methylation at the C5 position in a CpG context may also directly facilitate TF binding by providing additional hydrophobic methyl groups as contact sites or repress TF binding by sterically blocking the original contacts with unmethylated cytosine bases. For example, the MBD domain of the human transcription factor MeCP2 binds specifically to cytosine-methylated DNA via an interplay of hydrogen bonding and cation–π interactions between two MBD arginines and the methylated cytosine bases. Schulten and coworkers [25] studied this complex through molecular dynamics simulations and showed that methylation favors binding of MBD to mDNA by increasing the hydrophobic interfacial area of mDNA and stabilizing the interaction between mDNA and MBD proteins. Shanak et al. studied the same complex and suggested that C5-cytosine methylation entropically favors binding of the MBD domain to the human MeCP2 protein, whereas binding enthalpy made no noticeable contribution [26].

Regulatory effects of DNA methylation on TF binding are actually quite widespread. Taipale and colleagues [27] studied the binding of 542 human TFs to unmethylated vs. CpG-methylated DNA by using a methylation-sensitive systematic evolution of ligands by exponential enrichment (SELEX) protocol. They found that mCpG inhibited binding of most major classes of transcription factors to DNA, including Bzip, HLH, and ETS. Contrary to this, some other transcription factors, e.g. belonging to the POU, homeodomain, and NFAT domains showed a higher propensity for binding to methylated DNA.

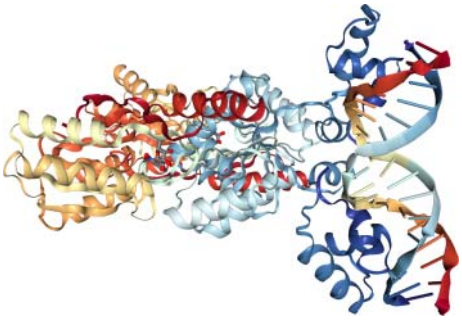


Figure 10.2 X-ray structure of Lac repressor from *E. coli* bound to 20-base pair symmetric operator segment of DNA and the anti-inducer orthonitrophenylfucoside (ONPF), PDB-code 1EFA.

10.5 Detection of DNA Curvature/Bending

The association of some TFs with DNA may result in strong bending of the DNA away from its canonical straight shape of B-DNA. Actually, different sequence contexts have different intrinsic stiffness against bending [28]. Hence, binding of TFs to bent DNA is another means to read out sequence-related aspects of the respective DNA segment besides any specific contacts established with nucleic acids mentioned above. Also, certain TF dimer pairs may only be able to bind to specifically bent DNA.

For example, X-ray crystallography showed that binding of a LacI dimer induces a 36° bending in DNA, see Figure 10.2. Liao et al. performed MD simulations of free LacI dimer and when it is complexed either to bent DNA as found in the X-ray structure of the complex, or to straight DNA (started from a modeled geometry) [29]. Even when LacI was modeled as close as sterically possible to straight DNA, the simulations did not show characteristic clamping of the LacI-binding helices around DNA, presumably due to steric clashes between LacI and the straight DNA. Furthermore, after LacI established contact with straight DNA, it started to slide along the DNA, reminiscent of a searching motion for a better position. The authors suggested that while the search process of a nonspecifically bound TF likely takes place along straight DNA, bending of the DNA promotes formation of a tightly bound specific (and possibly also nonspecific) LacI–DNA complex.

10.6 Modifications of Transcription Factors

Like most other cellular proteins, TFs may be subject to alternative splicing that often affects their activity levels and may even alter the direction of their activity. For example, Belluti et al. reviewed cases where alternative splicing isoforms of the TFs NFYA, STAT3, TCF4, and WT1 directly regulate specific transcriptional programs, leading to opposite cell fates [30]. Also, it was shown that two different isoforms of the transcription factor MeCP2 either promote pluripotency or drive stem cells into differentiation [31].

Adding and removing posttranslational modifications is another well-known mechanism to control the activity of cellular proteins. In effect, more than

two-thirds of the 21 000 proteins encoded by the human genome have been shown to be subject to phosphorylation [32]. Hence, phosphorylation is also a frequently used mechanism to link the activity of signaling pathways to the control of gene expression patterns [33]. For example, casein kinase I phosphorylate the pluripotency factor NANOG and thereby regulates self-renewal of embryonic stem cells [34]. Phosphorylation of retinoblastoma protein (Rb) drives the cell cycle [35]. Also, p53 can be modified by phosphorylation by a broad range of kinases, including ATM/ATR/DNA-PK and Chk1/Chk2 [36, 37].

10.7 Transcription Factor Binding Sites

DNA segments that establish specific physical interactions with individual TFs are named transcription factor binding sites (TFBSs). These commonly have a length between 8 and 20 base pairs (bp) and possess a core region of 5–8 bps of evolutionary highly conserved nucleotide bases. Neighboring positions around the core region may be more divergent. Double-stranded DNA has a periodicity of 10 bp. Hence, such core regions of short TFBS motifs are slightly longer than half a turn of dsDNA. TFs may also bind specifically to similar, but not identical DNA sequences that differ in a few nucleotide positions. Some TFs bind specifically to hundreds or even thousands of locations in the genome. As an example, Figure 10.3 shows the sequence motif preferred by TF Gata1. Such sequence logos are a convenient means to graphically display the degree of degeneracy in the TFBS.

10.8 Experimental Detection of TFBS

Still in use are several well-established experimental small-scale *in vitro* techniques to discover and analyze instances where a protein binds specifically to DNA or RNA. These methods include, for example, the electrophoretic mobility shift assay (EMSA) and the DNAse footprinting assay.

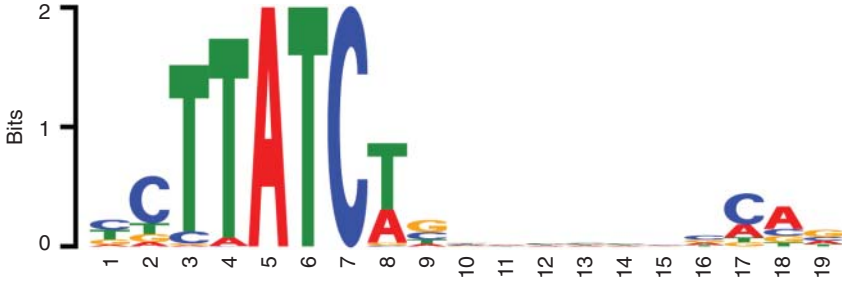


Figure 10.3 Binding motif of the human transcription factor Gata1 (according to www .factorbook.org. Source: Adapted from Wang et al. [38].

10.8.1 Protein-Binding Microarrays

One option to upscale these *in vitro* experiments to large numbers of DNA variants is protein-binding microarrays (PBMs) [39]. This DNA microarray technology enables one to measure under *in vitro* conditions one by one but in parallel the binding characteristics of multiple DNA-binding proteins. For this, one employs a microarray carrying different putative-binding motifs of double-stranded DNA in different wells, see Figure 10.4. The protein of interest is expressed and purified with an epitope tag and then added to the microarray. In a washing step, nonspecifically bound protein is removed from the solution. Then, all wells containing bound protein are read out in a labeling step by adding a fluorophore-conjugated antibody that binds specifically to the epitope tags exposed in populated wells. In this way, all spots carrying a significant amount of protein are identified. Based on the DNA sequences associated with these spots, one determines which DNA-binding site motifs are found to be enriched for the considered DNA-binding protein.

Biochemical solution assays, such as EMSA, DNAse footprinting, or protein-binding assays, are suitable to identify particular DNA motifs to which an individual TF prefers to bind. Based on the results, one generates a sequence logo (see Figure 10.3) or a position-specific scoring matrix (PSSM) for this TF (this will be explained below) and scans the genome sequence or parts of it for favorable binding positions based on the match of the considered segment to this TFBS motif. Unfortunately, such motifs have an overall quite short length and contain only few invariant positions. Hence, some motifs will be detected millions of times in the genome. Thus, although a TF could theoretically bind to any such motif instance *in vivo*, it is found only at about 1 in 500 sites in organisms having large genomes. For example, the mouse genome contains about 8 million segments with similarity

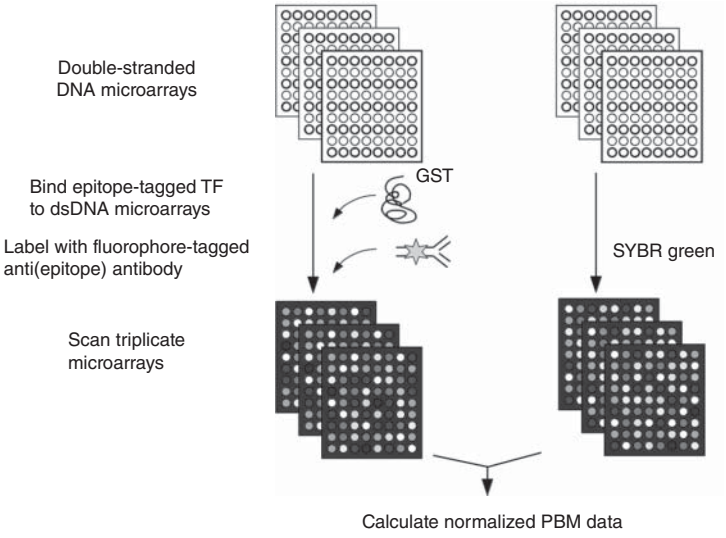


Figure 10.4 Main principles of protein-binding-microarray technique.

to the binding site motif of GATA-binding factor 1 (see Figure 10.3). However, in erythroid cells, GATA-1 was found to bind at only about 15 000 of them [40].

10.8.2 Chromatin Immunoprecipitation Assays

To overcome the drawbacks of *in vitro* assays just mentioned, modern parallel methods, such as ChIP-chip and ChIP-seq, are able to identify TF-binding sites *in vivo*. As indicated by their names, these methods utilize either DNA microarrays or new sequencing techniques, respectively.

A ChIP-seq experiment, see Figure 10.5, starts by purifying a cellular extract with the help of an antibody that binds to a particular TF. Then, the DNA sequences complexed to the TF are processed with a restriction enzyme. All leftover DNA can be assumed to be tightly coordinated by the respective TF. This TF is washed again from the DNA followed by sequencing of the DNA. All identified DNA reads belong to DNA segments that were bound to the TF before. Then, one uses a motif-search package, such as MEME [41], to characterize enriched sequence motifs among those

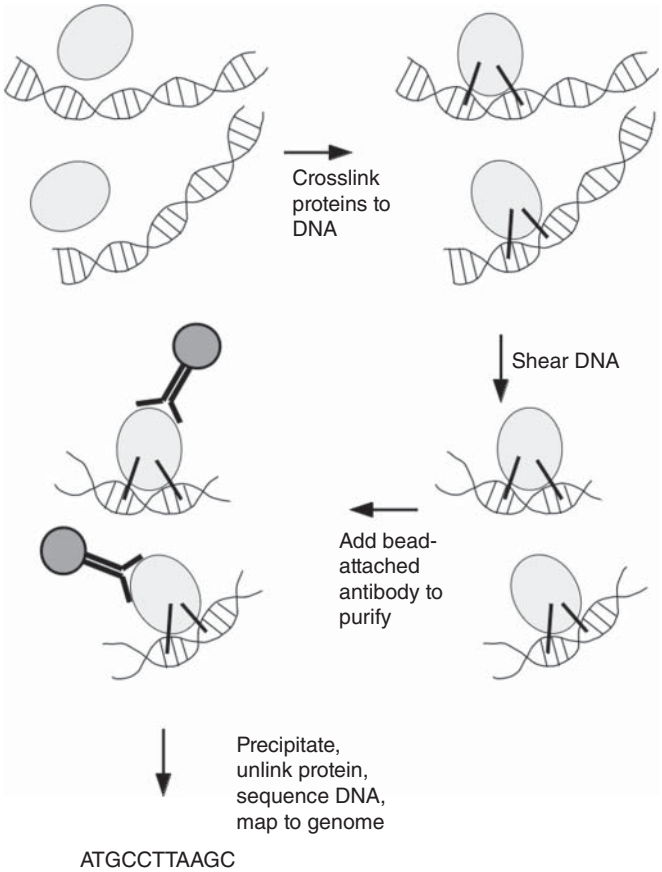


Figure 10.5 Main steps of a ChIP-seq experiment.

sequences. The resulting binding motif can be represented either as a PSSM or graphically as a sequence logo.

10.8.3 DamID Profiling of Protein–DNA Interactions

DamID profiling relies on the expression of *Escherichia coli* DNA adenine methyltransferase (abbreviated as Dam) that is fused with a chromatin interacting protein of interest (e.g. a TF) [42]. In a DamID-experiment, one monitors what adenine positions in the DNA get methylated to a significant extent compared to a control experiment where Dam is not present. Comparing the two outcomes then reveals to which DNA segments the chromatin interacting protein was bound (so that the fused Dam enzyme could methylate adenine bases in its spatial vicinity).

10.9 Position-Specific Scoring Matrices

A position-specific scoring matrix (PSSM) is often used to represent motifs (patterns) in biological sequences. In the case of TF-binding motifs, one identifies a number of DNA sequences able to bind a particular TF. Then one computes the frequency n_i^j of the 4 nucleotides in all relevant positions i, and from this the score matrix

$$s_i^j = \ln \frac{\left(n_i^j + p_i\right)/(N+1)}{p_i}$$

with N being the number of considered sequences. Adding the frequencies p_i in the numerator and dividing by $N+1$ resolves mathematically problematic cases where $n_i^j = 0$. A score $s_i^j = 0$ is assigned to all positions where the observed frequency matches what is expected randomly. Positive scores indicate enrichment of particular bases at these positions and negative scores vice versa.

Setting up a PSSM model implicitly assumes that there is no cooperativity between neighboring positions and each sequence position makes an independent contribution to the binding affinity that is approximated as a binding score. This is certainly a compromise. In reality, adjacent bases may clearly affect each other either directly or by affecting the local DNA conformation.

10.10 Molecular Modeling of TF–DNA Complexes

Molecular dynamics (MD) simulations described in Chapter 8 can also be used to study protein–DNA complexes. Such simulations provide detailed insight into the attractive interactions that drive the binding partners toward each other and that stabilize the atomic contacts formed in the specific complex. For example, Beuerle et al. reported that a copy of the transcription factor ERG that was laterally displaced by 2.3 nm from the DNA found its specific-binding site on the DNA sequence within 60 nanoseconds (ns) of simulation [43]. Huertas et al. studied the conformational dynamics of entire nucleosomes in MD simulations. They observed that the

DNA sequence segments containing binding sites for the pioneer TF Oct4 displayed increased local structural flexibility, which should essentially facilitate binding of Oct4 [44].

As described in Chapter 8, there also exist technologies to perform alchemical mutations within MD simulations, so that one nucleotide base can be mutated into another one, and the resulting binding free energy change can be reported. Gapsys and de Groot [45] utilized so-called free energy perturbation calculations to predict alterations of binding affinity upon 397 cases of nucleotide mutations that were experimentally reported for 16 different protein–DNA complexes. The authors reported that the computed binding affinity differences for single mutations deviated from experimentally measured binding affinities only by 5.6 kJ/mol on average with a correlation coefficient of 0.57. Hence, MD simulations can essentially be used to predict DNA mutation effects on TF-binding affinities in a quantitative manner.

Molecular dynamics simulations can also provide insight into larger-scale mobility. For example, Yu and colleagues [46] observed in all-atom MD simulations of the plant-transcription factor WRKY domain protein that the TF was able to process laterally on the DNA on a timescale of a few microseconds. Such simulations will eventually be able to pinpoint mechanistic details about TF–DNA recognition. Also, if combined with tailored force fields, molecular dynamics simulations can be applied to study partially disordered systems, such as the p53 complex [47], which are not easily amenable to structural biology.

10.11 Cis-Regulatory Modules

Although a typical eukaryotic cell expresses hundreds of TFs, a transcriptional code where individual TFs bind one by one to the promoter segments of target genes would not be sufficient to implement the required complex expression patterns of thousands of genes. Instead, expression of eukaryotic genes is typically controlled by simultaneous binding of multiple TFs to their promoter regions. Sometimes, the TFs can establish direct structural contacts among each other, e.g. illustrated in the X-ray structure of the Oct4-Sox2 dimer bound to DNA, see Figure 10.6. In such cases, it is well plausible that their mutual binding affinities are affected in a cooperative manner. For steric reasons, the distance between the TFBSs of adjacent TFs must then be restricted to a certain range.

Only in few cases, structural biology was able to capture the simultaneous binding of multiple TFs to nearby DNA segments. One such case involves the regulation of the interferon-beta (IFN-beta) gene. The expression of IFN-beta is initiated once the TFs ATF-2/c-Jun, IRF-3/IRF-7, and NFκB assemble cooperatively into a TF complex termed “enhanceosome” at the IFN-beta enhancer region [48]. This complex subsequently recruits further coactivators and chromatin-remodeling proteins to this sequence segment. Harrison and colleagues determined an X-ray structure of the DNA-binding domains of IRF-3, IRF-7, and NFκB, bound to one-half of the enhancer segment. Together with a previously determined structure

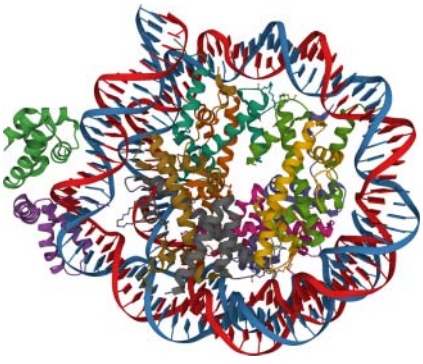


Figure 10.6 Two pluripotency factors, Oct4 (green) and Sox2 (violet), bound to a nucleosome (PDB structure 6YOV).

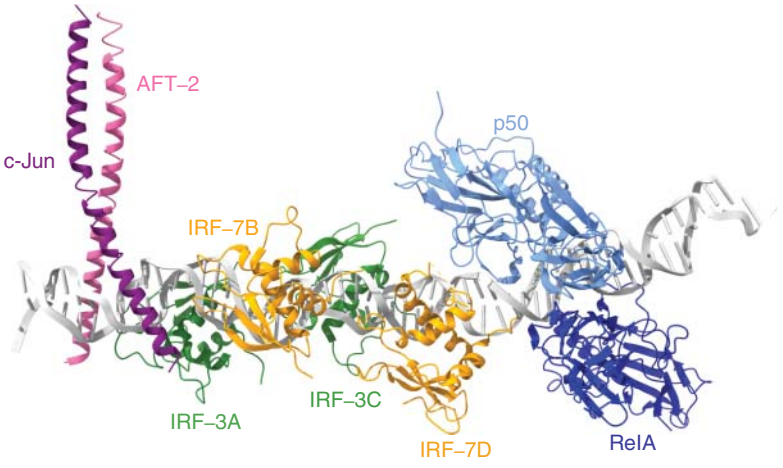


Figure 10.7 Structural model of an interferon-beta "enhanceosome" involving the transcription factors (c-Jun, AFT-2, IRF-3, IRF-7, p50, and Rel A). The model was generated by superposing chain B of the ATF-2/c-Jun/IRF-3/DNA complex (PDB-entry 1T2K) and chain F of another structure that includes four IRF-3 DNA-binding domains (2O6G). Afterward, parts of chain A of a structure of NFκB:IRF-7:IRF-3:DNA (2O61) were superposed with chains G and H of 2O6G. Of all doubly occurring proteins and DNA strands, one copy each was deleted. The superpositions and the graphics were created using UCSF ChimeraX [49].

of another assembly, this enabled them to generate a structural model of the complete enhanceosome architecture involving eight proteins in the vicinity of DNA. The model shown in Figure 10.7 reveals only few direct protein–protein contacts that led the authors to suggest that cooperative occupancy of the enhancer is due to binding-induced changes in DNA conformation on the one hand and interactions with additional proteins, such as CBP, on the other hand.

Such a cluster of multiple TFBS sites is termed a cis-regulatory module (CRM). Some regions of a CRM may promote binding of one or more TF complexes. In metazoans, a CRM may commonly extend more than 500 bp and may contain 10–50 TFBSs where between three and 15 different sequence-specific TFs may bind [50]. In case it contains multiple similar binding sites for an individual TF, this increases

the sensitivity for this TF, yields a more robust transcriptional response, or may simply favor the binding of a homo-oligomeric TF (e.g. NF-κB or p53). Some TF pairs have well-known binding partners, such as the TF pair Sox 2 and Oct4, illustrated in Figure 10.6. The ENCODE project mentioned in the next subsection found that 114 out of the 117 considered human TFs participate in about 3300 pairs of statistically co-associated factors. Given that 117 TFs could potentially form $117 \times 117/2 \approx 6000$ complexes, the finding that more than half of these actually exist in Nature is quite remarkable. These pairs included expected combinations, such as that of Fos and Jun, but also some unexpected novel combinations.

When a TF forms a complex with another TF of the same type, this is named a homotypic interaction. The Lac repressor dimer shown in Figure 10.2 is a well-studied example of such a homotypic assembly. Heterotypic interactions describe situations when one TF binds to another TF of a different type. Besides, DNA-binding TFs may also bind indirectly to other DNA-binding TFs by involving additional cofactors or bridge proteins. Chapter 5 explains the DACO algorithm for construction of protein complexes. One application scenario of this software is the identification of putative protein complexes that contain one or more transcription factors. When using TF complexes of *S. cerevisiae* as seed proteins, we identified a number of protein complexes containing two or three TFs [51]. We argued that binding of such multivalent complexes to the promoter regions of target genes may enable a much finer transcriptional regulation than binding of individual TFs.

10.12 Relating Gene Expression to Binding of Transcription Factors

As mentioned before, several experimental techniques can be used to detect binding of TFs to DNA segments. Depending on how comprehensive these data are, it enables setting up linear or nonlinear mathematical models of how the expression of a particular gene, a set of genes, or even all genes of an organism depend on the concentration levels and activation states of the available TFs.

A major advance in this field is due to the activities of the large-scale ENCODE consortium (short for **Enc**yclopedia **of D**NA **E**lements) that was funded between 2003 and 2012. Based on large-scale ChIP-seq experiments, functional regulatory elements were identified in the sequence of the human genome for a representative set of 147 different human cell types. In its main paper, ENCODE characterized the binding locations for 119 human DNA-binding proteins involving 87 of the 1600 known human sequence-specific TFs [52]. If one concatenates all identified transcription factor binding site motifs, 4.6% of the entire sequence is covered. Turned around, 95% of all genomic locations are located within 8 kb of a DNA-TF contact detected by ENCODE based on ChIP-seq. Furthermore, by classifying the genome into seven distinct chromatin states with the tool ChromHMM, the authors identified 400 000 regions having enhancer-like features and about 70 000 regions with promoter-like features, as well as hundreds of thousands of quiescent regions.

The detected gene expression levels spanned a wide dynamic range from 10^{-2} to 10^4 reads per kb per million reads (r.p.k.m.) for polyadenylated RNAs, and from 10^{-2} to 10^3 r.p.k.m. for non-polyadenylated RNAs. Several research groups, including the ENCODE team, presented linear regression models that relate the expression levels of individual genes to the occupancy of adjacent TFBSs. E.g for K562 cells predicted and observed expression levels predicted by the ENCODE team showed a Pearson correlation of 0.81 [52]. As the binding of TFs and epigenetic modifications are interrelated, linear regression models either based on TFBS occupancies or epigenetic modifications achieved similar predictive performance.

10.13 Summary

Protein–DNA interactions are among the most crucial biomolecular interactions in cells. Over the past decades, structural techniques, such as X-ray crystallography and NMR, were able to determine structures of the complexes of many transcription factors or other DNA-binding proteins when specifically bound to DNA. Their binding preference on the genome can be unraveled both by experimental *in vitro* assays or *in vivo* by ChIP-seq experiments. We discussed in this chapter how the binding affinity of transcription factors to linear DNA sequence motifs is modulated by conformational distortion of linear DNA, epigenetic modifications, and organizing multiple TFs into protein complexes that show cooperative binding to cis-regulatory sequence modules.

References

1 Lambert, S.A., Jolma, A., Campitelli, L.F. et al. (2018). The human transcription factors. *Cell* 172: 650–665. https://doi.org/10.1016/j.cell.2018.01.029.
2 Mistry, J., Chuguransky, S., Williams, L. et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49: D412–D419. https://doi.org/10.1093/nar/gkaa913.
3 Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* 49: D458–D460. https://doi.org/10.1093/nar/gkaa937.
4 Blum, M., Chang, H., Chuguransky, S. et al. (2020). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49: D344–D354. https://doi.org/10.1093/nar/gkaa977.
5 Babu, M.M. and Teichmann, S.A. (2003). Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* 31: 1234–1244. https://doi.org/10.1093/nar/gkg210.
6 Soto, L.F., Li, Z., Santoso, C.S. et al. (2022). Compendium of human transcription factor effector domains. *Mol. Cell* 82: 514–526. https://doi.org/10.1016/j.molcel.2021.11.007.

7 Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). The regulation of p53 by phosphorylation: a model for how distinct signals integrate into the p53 pathway. *Genome Biol.* 1, reviews001.1.

8 Sehnal, D., Bittrich, S., Deshpande, M. et al. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* 49: W432–W437. https://doi.org/10.1093/nar/gkab314.

9 Johnson, P.F. and McKnight, S.L. (1989). Eukaryotic transcriptional regulatory proteins. *Annu. Rev. Biochem.* 58: 799–839. https://doi.org/10.1146/annurev.bi.58.070189.004055.

10 Pabo, C.O. and Sauer, R.T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* 61: 1053–1095. https://doi.org/10.1146/annurev.bi.61.070192.005201.

11 Sonawane, A.R., Platig, J., Fagny, M. et al. (2017). Understanding tissue-specific gene regulation. *Cell Rep.* 21: 1077–1088. https://doi.org/10.1016/j.celrep.2017.10.001.

12 Cho, Y.J., Gorina, S., Jeffrey, P.D., and Pavletich, N.P. (1994). Crystal structure of p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* 265: 346–355. https://doi.org/10.1126/science.8023157.

13 Martínez-Jímenez, F., Muiños, F., Sentís, I. et al. (2020). A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* 20: 555–572. https://doi.org/10.1038/s41568-020-0290-x.

14 Berg, O.G., Winter, R.B., and von Hippel, P.H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* 20: 6929–6948. https://doi.org/10.1021/bi00527a028.

15 Shvets, A.A., Kochugaeva, M.P., and Kolomeisky, A.B. (2018). *Molecules* 23: 2106. https://doi.org/10.3390/molecules23092106.

16 Castellanos, M., Mothi, N., and Muñoz, V. (2020). Eukaryotic transcription factors can track and control their target genes using DNA antennas. *Nat. Commun.* 11 (540): https://doi.org/10.1038/s41467-019-14217-8.

17 Amoutzias, G.D., Robertson, D.L., Van de Peer, Y., and Oliver, S.G. (2008). Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem. Sci* 33: 220–229. https://doi.org/10.1016/j.tibs.2008.02.002.

18 Grandori, C., Cowley, S.M., James, L.P., and Eisenman, R.N. (2000). The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu. Rev. Cell Dev. Biol.* 16: 653–699.

19 Lee, D.-S., Chongyuan, L., Zhou, J. et al. (2019). *Nat. Methods* 16: 999–1006. https://doi.org/10.1038/s41592-019-0547-z.

20 Bannister, A.J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res.* 21: 381–395. https://doi.org/10.1038/cr.2011.22.

21 Zaret, K.S. (2020). Pioneer transcription factors initiating gene network changes. *Annu. Rev. Genet.* 54: 367–385. https://doi.org/10.1146/annurev-genet-030220-015007.

22 Zhu, F., Farnung, L., Kaasinen, E. et al. (2018). The interaction landscape between transcription factors and the nucleosome. *Nature* 562: 76–81. https://doi.org/10.1038/s41586-018-0549-5.

23 Yu, X. and Buck, M.J. (2019). Defining TP53 pioneering capabilities with competitive nucleosome binding assays. *Genome Res.* 29: 107–115. https://doi.org/10.1101/gr.234104.117.

24 Carvalho, A.T.P., Gouveia, L., Kanna, C.R. et al. (2014). Understanding the structural and dynamic consequences of DNA epigenetic modifications: computational insights into cytosine methylation and hydroxymethylation. *Epigenetics* 9: 1604–1612. https://doi.org/10.4161/15592294.2014.988043.

25 Zou, X., Ma, W., Solov'yov, I.A. et al. (2011). Recognition of methylated DNA through methyl-CpG binding domain proteins. *Nucleic Acids Res.* 40: 2747–2758. https://doi.org/10.1093/nar/gkr1057.

26 Shanak, S., Ulucan, Ö., and Helms, V. (2017). Methylation-targeted specificity of the DNA binding proteins R.DpnI and MeCP2 studied by molecular dynamics simulations. *J. Mol. Model.* 23 (152): https://doi.org/10.1007/s00894-017-3318-8.

27 Yin, Y., Morgunova, E., Jolma, A. et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* 356 (6337): https://doi.org/10.1126/science.aaj2239.

28 Geggier, S. and Vologodskii, A. (2010). Sequence dependence of DNA bending rigidity. *PNAS* 107: 15421–15426. https://doi.org/10.1073/pnas.1004809107.

29 Liao, Q., Lüking, M., Krüger, D.M. et al. (2019). Long time-scale atomistic simulations of the structure and dynamics of transcription factor-DNA recognition. *J. Phys. Chem. B* 123: 3576–3590. https://doi.org/10.1021/acs.jpcb.8b12363.

30 Belluti, S., Rigillo, G., and Imbriano, C. (2020). Transcription factors in cancer: when alternative splicing determines opposite cell fates. *Dells* 9 (760): https://doi.org/10.3390/cells9030760.

31 Lu, Y., Loh, Y.-H., Li, H. et al. (2014). Alternative splicing of MBD2 supports self-renewal in human pluripotent stem cells. *Cell Stem Cell* 15: 92–101. https://doi.org/10.1016/j.stem.2014.04.002.

32 Ardito, F., Giuliani, M., Perrone, D. et al. (2017). The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. *Int. J. Mol. Med.* 40: 271–280. https://doi.org/10.3892/ijmm.2017.3036.

33 Nardozzi, J.D., Lott, K., and Cingolani, G. (2010). Phosphorylation meets nuclear import: a review. *J. Cell Commun* 8 (32): https://doi.org/10.1186/1478-811X-8-32.

34 Mullin, N.P., Varghese, J., Colby, D. et al. (2020). Phosphorylation of NANOG by casein kinase I regulates embryonic stem cell self-renewal. *FEBS Lett.* 595: 14–25. https://doi.org/10.1002/1873-3468.13969.

35 Giacinti, C. and Giordano, A. (2006). RB and cell cycle progression. *Oncogene* 25: 5220–5227. https://doi.org/10.1038/sj.onc.1209615.

36 Kruse, J.-P. and Gu, W. (2009). Modes of p53 regulation. *Cell* 137: 609–622. https://doi.org/10.1016/j.cell.2009.04.050.

37 Maclaine, N.J. and Hupp, T.R. (2009). *Aging* 1: 490–502. https://doi.org/10.18632/aging.100047.

38 Wang, J., Zhuang, J., Iyer, S. et al. (2013). Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.* 41: D171–D176. https://doi.org/10.1093/nar/gks1221.

39 Berger, M.F. and Bulyk, M.L. (2006). Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol. Biol.* 338: 245–260. https://doi.org/10.1385/1-59745-097-9:245.

40 Hardison, R.C. and Taylor, J. (2012). Genomic approaches towards finding *cis*-regulatory modules in animals. *Nat. Rev. Genet.* 13: 469–483. https://doi.org/10.1038/nrg3242.

41 Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME suite. *Nucleic Acids Res.* 43: W39–W49. https://doi.org/10.1093/nar/gkv416.

42 Aughey, G.N. and Southall, T.D. (2016). Dam it's good! DamID profiling of protein-DNA interactions. *Wiley Interdiscip. Rev.: Dev. Biol.* 5: 25–37. https://doi.org/10.1002/wdev.205.

43 Beuerle, M.G., Dufton, N.P., Randi, A.M., and Gould, I.R. (2016). Molecular dynamics studies on the DNA-binding process of ERG. *Mol. Biosyst.* 12: 3600–3610. https://doi.org/10.1039/C6MB00506C.

44 Huertas, J., MacCarthy, C.M., Schöler, H.R., and Cojocaru, V. (2020). Nucleosomal DNA dynamics mediate Oct4 pioneer factor binding. *Biophys. J.* 118: 2280–2296. https://doi.org/10.1016/j.bpj.2019.12.038.

45 Gapsys, V. and de Groot, B.L. (2017). Alchemical free energy calculations for nucleotide mutations in protein–DNA complexes. *J. Chem. Theory Comput.* 13: 6275–6289. https://doi.org/10.1021/acs.jctc.7b00849.

46 Dai, L., Xu, Y., Du, Z. et al. (2021). Revealing atomic-scale molecular diffusion of a plant-transcription factor WRKY domain protein along DNA. *PNAS* 118: e2102621118. https://doi.org/10.1073/pnas.2102621118.

47 Demir, Ö., leong, P.U., and Amaro, R.E. (2017). Full-length p53 tetramer bound to DNA and its quaternary dynamics. *Oncogene* 36: 1451–1460. https://doi.org/10.1038/onc.2016.321.

48 Panne, D., Maniatis, T., and Harrison, S.C. (2007). An atomic model of the interferon-β enhanceosome. *Cell* 129: 1111–1123. https://doi.org/10.1016/j.cell.2007.05.019.

49 Pettersen, E.F., Goddard, T.D., Huang, C.C. et al. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25: 1605–1612. https://doi.org/10.1002/jcc.20084.

50 Wray, G.A., Hahn, M.W., Abouheif, E. et al. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20: 1377–1419. https://doi.org/10.1093/molbev/msg140.

51 Will, T. and Helms, V. (2014). Identifying transcription factor complexes and their roles. *Bioinformatics* 30: i415–i421. https://doi.org/10.1039/bioinformatics/btu448.

52 The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74. https://doi.org/10.1038/nature11247.

11

The Chromatin Interaction System

Sarah Kreuz, Stefan-Sebastian David, Lorena Viridiana Cortes Medina, and Wolfgang Fischle

King Abdullah University of Science and Technology, Division of Biological and Environmental Sciences and Engineering, Laboratory of Chromatin Biochemistry, Thuwal 23955-6900, Saudi Arabia

11.1 Chromatin Is a Special Interaction Platform

Inside the nucleus of each cell, genetic information in the form of DNA is packaged into chromatin. Various cellular pathways of homeostasis and responses to environmental stimuli are integrated on chromatin as an important signaling platform culminating in regulation of all DNA-dependent processes (e.g. transcription, replication, and repair). Nucleosomes are the repeating units of chromatin. These are nucleoprotein complexes of two copies each of the core histone proteins H2A, H2B, H3, and H4, wrapped by 147 bp of DNA, and connected to each other by linker DNA (Figure 11.1A). The canonical core histones are very basic (i.e. rich in amino acids lysine and arginine), globular proteins of 11–15 kDa. These have extended N- and C-terminal tails of 12–39 amino acids, which protrude from the nucleosome core particle. In addition to linker histones of the H1 type that associate with linker DNA, an extended view of chromatin also contains nonhistone proteins, RNAs, and small molecules, which are more or less tightly associated with the genetic material.

With its unique and complex composition, chromatin represents a special environment for protein interactions with a plethora of other molecules (Figure 11.1). Beyond the generic composition of chromatin, the number of possible distinct interactions is increased by several orders of magnitudes due to the locally restricted exchange of the canonical histones for so-called histone variants (distinct functionalities due to sequence variation) and not least by highly abundant chemical modifications of histones, nonhistone proteins, and nucleic acids (Figure 11.1A). These control local protein–chromatin interactions, affect enzymatic activities, and modulate protein structures.

In addition to its unique molecular and biochemical composition, chromatin also has special biophysical properties and is a hotspot for phase separation events. It is thought that these can have significant effects on protein interactomics.

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

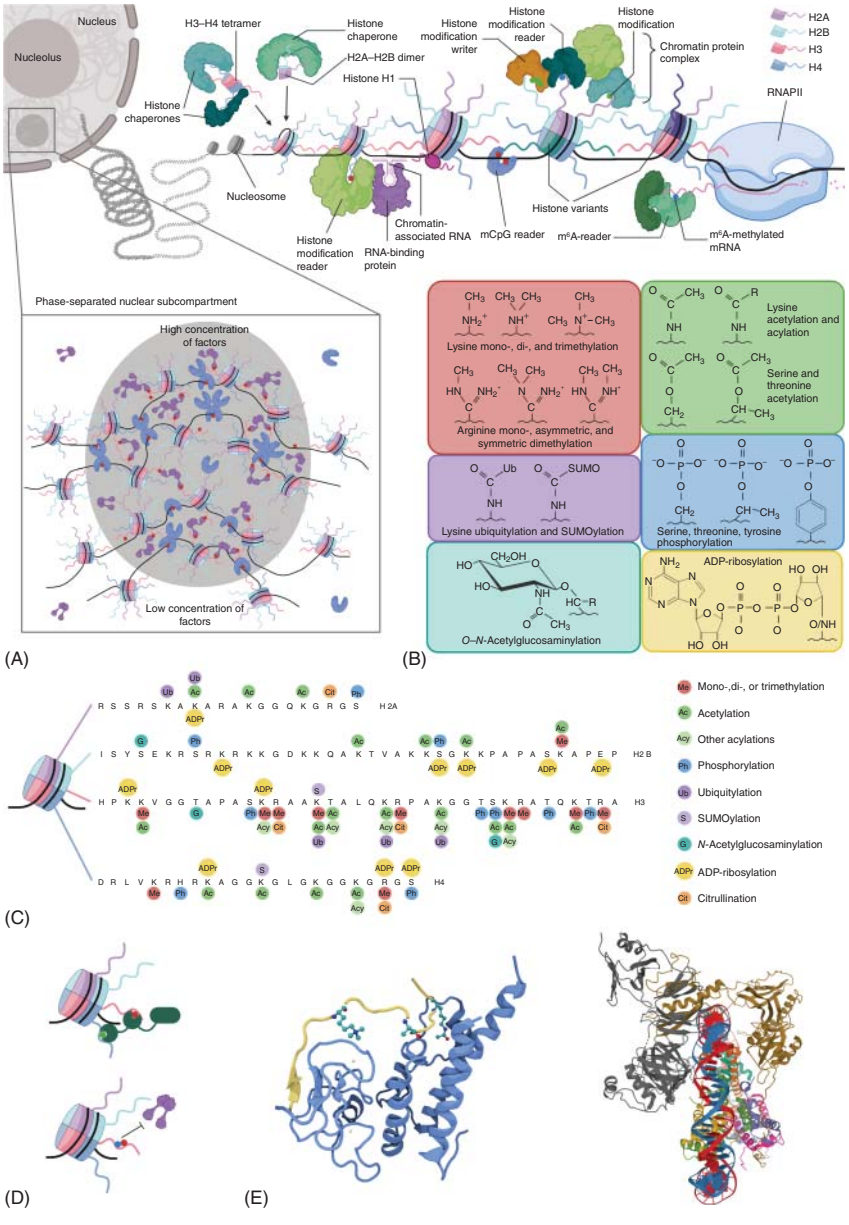


Figure 11.1 ***Introduction to the chromatin interaction system***. (**A**) Schematic representation of the complex chromatin interaction platform and its components. The system comprises multiple protein–protein and protein–nucleic acid interactions. Many of these interactions are controlled by posttranslational modifications (PTMs) of the histone proteins and chemical modifications of RNA and DNA. The concept of protein- and/or nucleic acid-driven phase separation can explain the existence of subnuclear chromatin compartments containing specific proteins and exerting specialized functions (e.g. splicing, transcription, etc.). (**B**) Common PTMs of amino acids in histone proteins. (**C**) Complexity of PTMs on the N-terminal tails of histones. The panel is not exhaustive and new marks are still being discovered. (**D**) The combination of chromatin modifications forms a code that is instructive for protein interactions and thus functional outputs. Positive crosstalk (*top*): Proteins with multiple reader domains recognize combinations of modifications. Negative crosstalk (*bottom*): Inhibition of interaction of a protein with a chromatin modification by a neighboring modification. (**E**) Structural studies facilitate the understanding of protein–histone tail and nucleosome recognition. Crystal structure of TRIM33 PHD and Bromo domains binding H3K9me3K18ac in a coordinated manner (PDB ID: 3U5O) (*left*). Cryo-EM structure of the FACT chaperone interacting with a partial nucleosome (PDB ID: 6UPK) (*right*). This figure was generated with Biorender.com.

Liquid–liquid phase separation (LLPS) is defined as the spontaneous, reversible unmixing of a solution into two distinct liquid phases: one concentrated, condensed and one dilute phase. Proteins and RNAs with low-complexity regions that display multivalent interactions with each other can form supramolecular clusters and induce LLPS [1]. Interaction of proteins with distinct modifications on chromatin can stimulate this process by increasing the local concentration of phase-separating proteins. Polymer–polymer phase separation (PPPS) occurs when proteins bridge different chromatin fibers, which become a densely packed globule surrounded by more solvent exposed regions (Figure 11.1A). While it is not yet fully resolved which processes, LLPS or PPPS, are most relevant for chromatin biology, both scenarios result in certain molecules becoming enriched in or depleted from one of the two phases [2].

11.2 Interaction of Proteins with Histone Posttranslational Modifications

11.2.1 The History of Histone Posttranslational Modifications and the Histone Code

Initially, histones were thought to be mere accessory proteins that facilitate the packaging and compaction of DNA inside the cell nucleus by providing a static scaffold. Now, it is known that histones and histone posttranslational modifications (PTMs) play important regulatory roles in all cellular processes that depend on DNA. A large number of PTMs – and certainly the best studied ones – are found in the exposed N- and C-terminal histone tails (Figure 11.1B,C). Based on their topology, it is easy to envision these serving as docking sites for multiple proteins. PTMs are also found in the histone globular domains, where they regulate nucleosome assembly, stability, and neighboring internucleosomal contacts [3–6].

Acetylation of lysine residues in the N-terminal tails of histones was first described in 1963 [7]. A year later, histone acetylation was shown to stimulate transcription in cell-free calf thymus extracts, suggesting that this PTM might directly influence histone-DNA interactions [8]. In 1964, lysine methylation of histones was first detected [9], and in 1966 the incorporation of ^{32}P into histones was discovered in rats, implying protein phosphorylation [10]. In recent years, the types of histone PTMs identified have exploded to include ADP-ribosylation, ubiquitylation, sumoylation, and various types of acylation, glycosylation, and serotonylation (see Table 11.1) (reviewed in [11]).

The vast complexity of the histone PTM system originates not only from the many types of chemical modifications: (i) Certain sites in the various histones are targeted by different modifications (for example, the lysine 9 residue of histone H3 can be methylated, acetylated, acylated, or ubiquitylated). (ii) The same modifications can occur on different sites of the various histones (for example, histone H3 can be acetylated on residues lysine 4, lysine 9, lysine 14, etc.). (iii) Methylations occur in different stages (i.e. lysines can be mono-, di-, or tri-methylated; arginines can be monomethylated as well as symmetrically or asymmetrically dimethylated). (iv) Lastly, complex patterns of modifications with multiple sites in individual histones or different histones within a nucleosome or a stretch of the genome likely exist (Figure 11.1B,C) [12]. Over 500 unique histone marks (different site-specific modifications) have been described to date (reviewed in [11]). Uniformly, histone modifications are annotated by (i) type of histone, (ii) site of modification, (iii) type of modification, and (iv) modification level; for example, H3K9me3 refers to histone H3 modified on the lysine 9 residue by trimethylation.

As electrostatic forces between positively charged histones and negatively charged DNA are a main component of histone-DNA binding, it was proposed early that PTMs could modulate nucleosome stability and internucleosomal contacts. However, several types of modifications do not have charge changing properties. Their functional mechanisms were revealed when the first proteins that specifically recognize histone modifications were identified on the basis of rational guesses derived from cellular and genetic experiments [13–15]. Such factors are commonly referred to as reader proteins. These contain specialized domains that recognize histone residues carrying PTMs in sequence-specific contexts, generally within short linear peptide motifs (5–10 amino acids) (Figure 11.1D,E and refer to Table 11.2) [15–24]. Reader proteins may contain multiple reader domains or may be part of multiprotein complexes, including other reader proteins. Often, readers recruit, or are part of, chromatin-modifying enzymes that write or erase certain modifications, or that remodel chromatin (Figure 11.1).

With the discovery of many histone PTMs and the elucidation of their molecular working mechanisms, it was proposed that distinct histone PTMs can act in combination or sequentially to function as a "histone code" that controls fundamental chromatin-mediated processes [25, 26]. In agreement with this hypothesis, we now know that different regions of the genome (e.g. enhancers, promoters, and repetitive elements) are decorated by specific patterns of histone PTMs. These appear to be instructive for the fate of such regions (see Figure 11.1D).

Table 11.1 Types and examples of chromatin modifications.

Modified residue	Modification	Effector function	Examples	Effect on chromatin processes
Histones				
Lysine	Monomethylation	Docking sites for chromatin binding domains	H3K4me1	Marks active and inactive enhancers
			H4K20me1	DNA repair, chromatin condensation, gene expression
	Dimethylation		H3K27me2	Enhancer silencing
			H3K36me2	Preventing the spread of repressive H3K27me3, modulating DNA methylation
	Trimethylation		H3K4me3	Activation of gene expression
			H3K9me3	Repression of gene expression, heterochromatin formation
			H3K36me3	Repression of transcription initiation, splicing regulation
	Acetylation	Charge neutralization (hydrophobic group), docking sites for chromatin binding domains	H2BK5ac	Active promoters
			H3K27ac	Active enhancers and promoters
			H4K5acK8ac	Chromatin remodeling, activating gene expression
	Formylation		?	?
	Propionylation		H4K16pr	?
	Butyrylation		H4K5acK8bu	Chromatin remodeling, activating gene expression
	Crotonylation		H3K9cr	Gene expression
			H3K14cr	Spreading of acylation, open chromatin
	2-Hydroxyisobutyrylation	Charge neutralization (polar group), docking sites for chromatin binding domains	H4K8hib	Associated with highly transcribed genes
	β-Hydroxybutyrylation		H3K9bhb	Activation of gene expression

(*continued*)

Table 11.1 (Continued)

Modified residue	Modification	Effector function	Examples	Effect on chromatin processes
	Malonylation	Charge reversal (acidic group), docking sites for chromatin binding domains	H2AK119ma	Inhibition of H2AS121 phosphorylation, regulation of chromosome segregation?
	Succinylation		H3K122succ	Activation of transcription?
	Glutarylation		H4K91glu	Destabilization of nucleosome, activation of transcription, DNA damage repair
	Ubiquitylation	Sterical interference	H2BK34ub	Nucleosome destabilization, transcriptional elongation
		Docking sites for chromatin binding domains	H2AK119ub	Transcriptional repression
			H3K18ubK23ub	DNA maintenance methylation
	SUMOylation	Sterical interference, docking site for chromatin binding domains	H4K12su	Inhibition of chromatin compaction, gene repression
	Poly-ADP-ribosylation	Charge reversal, sterical interference, docking sites for chromatin binding domains?	H3K27par, H3K37par, H4K16par	Chromatin decompaction
	5-Hydroxylation	?	?	Inhibition of acetylation, gene repression?
Arginine	Monomethylation	Docking sites for chromatin binding domains	H3R2me1	Enriched on active promoters
	Symmetric dimethylation		H3R8me2s	Transcriptional repression or activation
			H3R2me2sR8me2s	Transcriptional activation
			H4R3me2s	Transcriptional repression
	Asymmetric dimethylation		H3R2me2a	Transcriptional repression
			H3R8me2a	Inhibition of heterochromatin formation
	Mono-ADP-ribosylation	?	H3R117mar	Activation of gene expression
	Poly-ADP-ribosylation	?	?	?
	Citrullination	Charge neutralization, inhibition of protein binding, inhibition of methylation	H3R8cit, H3R26cit	Activation of gene expression
			H3R17cit	Inhibition of gene expression

Serine	Acetylation	?	H3S10ac	?
	Phosphorylation	Docking sites for chromatin binding domains, addition of negative charge	H2A.XS139ph	DNA damage repair, chromatin decompaction
			H3K9acS10ph	Activation of transcription
			H3S10ph	Mitotic chromatin compaction
	N-Acetylglucosaminylation	?	H2AS40gc	DNA damage repair
			H2BS112gc	Enriched near active genes
		Competition with other modifications	H3S10gc	Co-occurrence with active and repressive marks
	Poly-ADP-ribosylation	?	H3S10par	DNA damage repair
Threonine	Acetylation	?	H3T22ac	?
	Phosphorylation	Docking site for chromatin binding domains, addition of negative charge	H3T3ph	Chromosome segregation, transcriptional repression, heterochromatin formation
			H3T6ph	Activation of gene expression
	N-Acetylglucosaminylation	?	H3T32gc	Inhibition of mitotic entry
Tyrosine	Acetylation	?	?	?
	Phosphorylation	Docking site for chromatin binding domains, addition of negative charge	H3Y41	Heterochromatin decompaction, activation of transcription
			H4Y51	DNA damage repair
	Hydroxylation	Affect internucleosomal interactions?	H3BY83oh, H4Y88oh	Alteration of chromatin structure?
Histidine	Phosphorylation	?	H4H18ph	?

(*continued*)

Table 11.1 (Continued)

Modified residue	Modification	Effector function	Examples	Effect on chromatin processes
Glutamate	Mono-ADP-ribosylation	?	H2BE18marE19mar	DNA damage repair
	Poly-ADP-ribosylation	Docking site for chromatin binding domains	H2AXE141par	DNA damage repair
		Sterical interference, docking site for chromatin binding domains?	H2BE2par	Chromatin decompaction
Glutamine	Methylation	Inhibition of protein binding	H2AQ104me	Inhibition of nucleosome deposition, RNAPI transcription
	Serotonylation	Docking site for chromatin binding domains	H3K4me3H3Q5ser	Activation of transcription
DNA				
Cytosine	5-Methylation	Nucleosome compaction, docking site for chromatin-binding domains, inhibition of protein binding	mCpG	Repression of transcription, heterochromatin formation
	5-Hydroxymethylation	Docking site for chromatin-binding domains	hmCpG	Regulation of transcription, DNA repair
	5-Formylation	Docking site for chromatin-binding domains, alteration of DNA structure	fCpG	Regulation of transcription, DNA repair, repression of transcription elongation
	5-Carboxylation	Docking site for chromatin-binding domains	caCpG	Regulation of transcription, DNA repair, repression of transcription elongation
RNA				
Adenosine	N6-methylation	Docking site for protein domains, structural changes	RRm^{6}ACH (R = G/A and H = A/C/U)	Enhancement of translation, regulation of RNA stability

Table 11.2 Chromatin modifications and their protein binding domains.

Modification recognized	Reader domain	Example (reader – modification)
Histones		
Lysine methylation	Chromo	HP1 – H3K9me3
	PHD	KDM5B – H3K4me3
	Tudor	PHF1 – H3K36me3, UHRF1 – H3K9me2/3
	MBT	L3MBTL – H4K20me1
	ZF-CW	ZCWPW1 – H3K4me3, ASHH2 – H3K4me1
	PWWP	NSD2 – H3K36me2, BRPF1 – H3K36me3
	ADD	ATRX – H3K9me3
	Ankyrin repeats	G9A, GLP – H3K9me1/2
	WD40	EED – H3K27me3
	BAH	BAHD1 – H3K27me3
Lysine acetylation and acylation	Bromo	BPTF – H4K5acK8ac
	Double PHD	MOZ – H3K14cr
	YEATS	AF9 – H3K9cr
Lysine ubiquitylation	UBDs	DNMT1 – H3K18ubK23ub, RAD18
Lysine SUMOylation	SIMs	CoREST – H4K12su
Arginine methylation	Tudor	TDR3 – H3R2me2a
	WD40	WDR5 – H3R2me2s
Serine phosphorylation	14-3-3	14-3-3ζ – H3K9acS10ph or H3S10phK14ac
	BRCT	MDC1 – H2AXS139ph
Threonine phosphorylation	14-3-3	?
	BIR	Survivin – H3T3ph
Tyrosine phosphorylation	SH2	ABL1 – H4Y51ph
Glutamate poly-ADP-ribosylation	zf-GRF	NEIL3 – H2AXE141par
Glutamine serotonylation	PHD	TAF3 – H3K4me3Q5ser
DNA		
5-Methylcytosine	MBD	MBD1 – symmetric mCpG
	SRA	UHRF1 – hemimethylated mCpG
	ZF	ZFP57 – TGCmCGC

(*continued*)

Table 11.2 (Continued)

Modification recognized	Reader domain	Example (reader – modification)
5-Hydroxymethylcytosine	MBD	MeCP2 – symmetric and asymmetric hmCpG
	SRA	SUVH5 – symmetric hmCpG, UHRF2 – hmCpG
5-Formylcytosine	MBD	MBD4 – symmetric or asymmetric 5fCpG
5-Carboxylcytosine	MBD	MBD3 – symmetric caCpG
	SRA	UHRF1 – symmetric caCpG
	ZF	WT1 – GmCGTGGGGcaCG
RNA		
N6-methyladenosine	YTH	YTHDF2 – Gm^6ACU/A
	KH	IGFBPs – $UGGm^6AC$
	RGG	HNRNPG – $AGGm^6AC$

11.2.2 Peptides and Nucleosomal Templates for Studying Histone PTMs

As the majority of histone PTMs work via reader domains and as the interaction motifs are short and linear, peptides have become central for studying the readout of histone marks. Nowadays, solid phase synthesis methods provide straight-forward access to histone peptides of various modification types. The peptides can be functionalized, for example, via biotinylation for immobilization on solid support or fluorescently tagged to measure binding affinities of reader domains [27]. Peptides can also be derivatized to display reactive C-termini for protein engineering [28], or modified to contain unnatural amino acids or photo-reactive cross-linkers such as benzophenone and diazirine (Figure 11.2A) [29, 30].

For capturing the complexities of the histone modification system, more elaborate substrates are required. This includes the study of crosstalk of distant histone modifications and, in particular, in different histone proteins, the interplay of histone and DNA modifications, as well as the analysis of downstream effects (deposition of new marks, nucleosome remodeling, etc.) (Figure 11.1E). To this end, several methods for incorporating PTMs directly into histone proteins have been put forward. These include derivatization of non-native cysteine residues incorporated at sites of modification via mutagenesis [31] or of chemical precursors introduced into the histone sequences via genetic code expansion [32, 33]. In a semisynthetic manner, modified histone tail peptides are conjugated with recombinant histone core proteins through enzymatic conjugation [34], native chemical ligation [35, 36], protein trans-splicing [37], or chemoselective ligation [38]. To obtain "designer chromatin" for various experimental studies, the obtained modified and purified histones are incorporated into mononucleosomes or nucleosomal arrays using

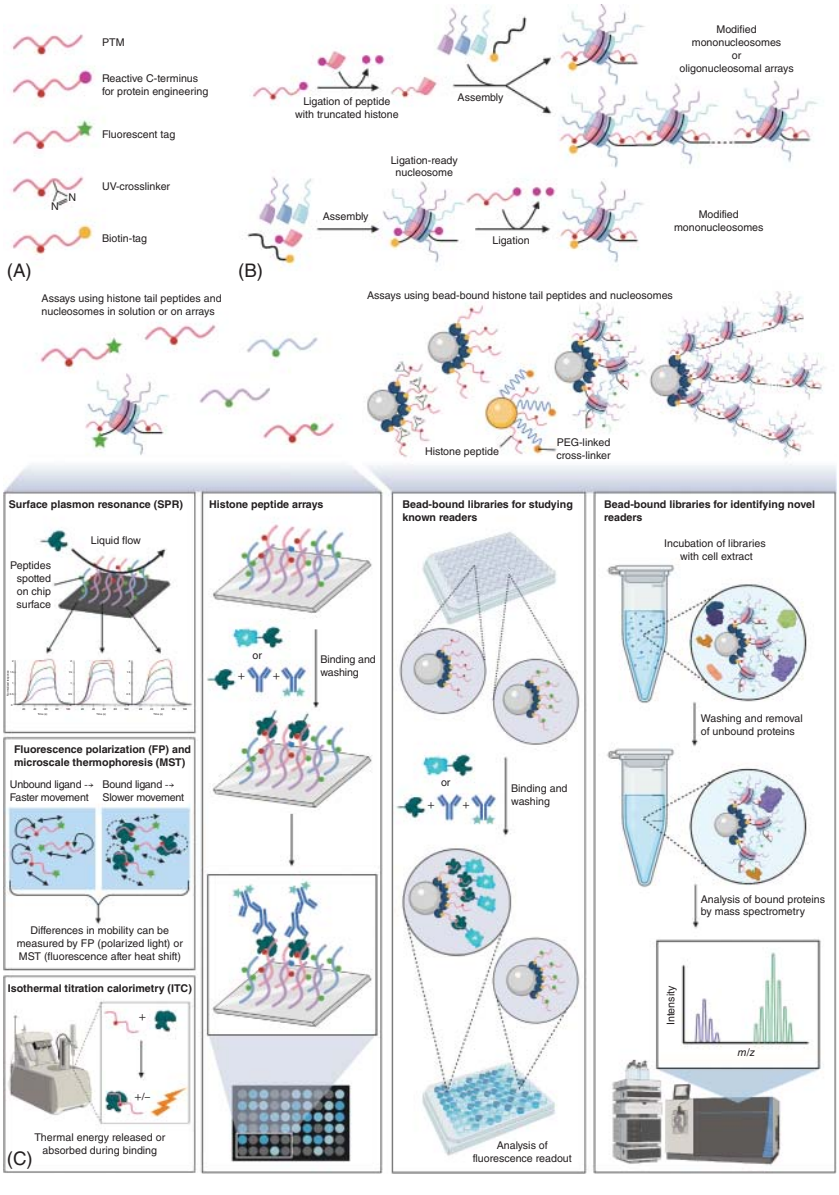


Figure 11.2 ***Analysis of histone PTM readout using synthetic peptide and nucleosome templates.*** (**A**) Examples of functionalized histone tail peptides used in various applications. (**B**) To obtain "designer chromatin", modified histone tails are conjugated with recombinant tailless histones either before or after nucleosome assembly. (**C**) Histone peptides and "designer chromatin" are used to quantitatively or qualitatively characterize binding specificities of known reader proteins in solution (FP, MST, ITC) or on solid support (SPR, peptide arrays) (*left*). Bead-bound chromatin ligands enable the study of known readers in plate format or the identification of novel readers from cell extract using mass spectrometry (*right*). This figure was generated with Biorender.com.

salt-dialysis methods in conjunction with DNA templates that facilitate nucleosome positioning [39, 40]. Alternatively, modified tail peptides can be directly ligated to tailless histones post nucleosome assembly (Figure 11.2B) [38].

11.2.3 Qualitative Analysis of Histone PTM Readout

Affinity enrichment experiments capture interactions between ligands and target proteins. In these schemes, the unmodified and modified ligands (histone peptides, mononucleosomes, or nucleosome arrays) are immobilized on solid support (beads, membranes, and glass surfaces) and used as baits to recover recombinant or native reader proteins. Affinity purification experiments are mostly used as preliminary screening tools, as these only offer qualitative information of the interaction between a reader and its target binding site.

11.2.3.1 Characterizing Binding Specificities of Known Readers

Classical pull-down experiments are employed to routinely screen for binding preferences of readers to several targets in parallel. In these experiments, immobilized histone peptides or nucleosomal baits are incubated with the recombinant or native reader protein. Recovery of the reader on the bait is usually detected using SDS-PAGE followed by staining or western blotting [41].

The main drawback of this approach is the limited number of interactions that can be assessed at a time. Therefore, more comprehensive high-throughput screening alternatives using a large number of baits have been developed. These include arrays of many histone peptides and nucleosomes. The multiple binding targets are spotted onto membranes, coupled to multi-well plate surfaces or libraries that are generated by combining large numbers of individually bait-coupled beads. Binding of the reader is mostly detected using primary and secondary antibodies, and fluorescence or chemiluminescence readout (Figure 11.2C) [38, 42]. In the opposite scheme, readers are immobilized and exposed to libraries of soluble peptides or mononucleosomes. Interactions are monitored, for example, using fluorescently labeled peptides [43], or via barcoded nucleosomes and DNA sequencing [44]. The library-based assays can be used to explore binding preferences (i.e. modification specificity and dependence on sequence context) of known chromatin readers to histone PTMs [38, 45–49], to characterize readers found by sequence homology [50, 51] or to validate the specificity of antibodies raised against defined histone PTMs [48, 50–52].

Histone peptide arrays have been described that contain hundreds [53, 54], thousands [47], or close to ten thousand synthetic histone peptides [45]. Increasing the dimension of libraries up to one million unique peptides has been achieved by using spectrally encoded beads that are linked to specific peptides. These beads are generated by incorporating fluorescent lanthanide nanophosphors into the bead structure. Varying the ratios of different fluorophores added to individual beads can theoretically produce 10^6 distinguishable codes [55–58]. In contrast to peptide libraries, the nucleosome libraries described to date only contain 100–300 members. However, with the recent generation of ligation-ready nucleosomes,

it has become possible to generate larger numbers of modified nucleosomes at once [38].

11.2.3.2 Identification of New Reader Proteins

Affinity purification using baits of histone peptides, mononucleosomes, or nucleosomal arrays can be used to identify new histone PTM readers from cell extracts using mass spectrometry (MS) (Figure 11.2C). Such experiments are now routinely performed, thanks to advances in separation of complex protein mixtures by affinity purification, gel electrophoresis, reverse phase HPLC and ion mobility, and due to improvements in the sensitivity and resolution of mass spectrometers. Comparing unmodified and modified baits is instrumental in this undertaking to filter true positive from false positive interactions, which can result from highly abundant and/or sticky proteins [59]. Typical histone peptide or chromatin affinity purification experiments result in the identification of around 2000 proteins, of which approximately one to two percent are ultimately found to be significantly enriched on the modified substrate compared to the control substrate [59–61]. To quantify enrichment, the differences between a protein's interaction with modified and unmodified baits need to be determined. This is done using SILAC labeling schemes or isobaric tags post purification [62–64]. In addition, improved data analysis tools have made it possible to carry out label-free quantification of enrichment ratios [65, 66].

Many experiments have now been described that use peptides to study several single or combinatorial marks on the same histone tail [60, 61, 67–73]. To investigate combinatorial chromatin modifications (on histones and DNA or different histones), nucleosome baits are used. These, in contrast to peptide baits, facilitate recruitment of multisubunit complexes and DNA-binding factors [34, 61, 74–76].

An extension to analyzing histone PTMs as the basis for protein interactions is the study of nucleosome surfaces that interact with chromatin proteins. Mutagenesis has been employed to disrupt charges of the solvent-exposed protein surface on recombinant nucleosomes, which were then used for pulldowns with cell extracts [77].

Most histone PTM readers bind their preferred modification in the mid-micromolar range [75]. Weaker interaction partners (dissociation constants up to 100 mM) can be recovered in the histone peptide pull-downs and on histone peptide arrays by adjusting the washing conditions that remove unspecifically bound proteins at the cost of increasing the chance of false positive identifications [42, 75]. To avoid the problem of false discovery in the study of weak or transient interactions, as well as to identify enzymatic activities that remove the modification of interest, histone peptides containing photo cross-linkable residues are used. To this end, samples are exposed to UV light for cross-linking after only a very brief incubation with the cell extract. As before, retained proteins can then be identified by MS [29, 30, 78]. Another strategy makes use of the self-assembled multi-valent photo-cross-linking technique, in which a histone peptide and a PEG-linked photo cross-linker are assembled on the same nanoparticle. In this approach, binding proteins are photo-cross-linked outside their binding region,

thereby avoiding any interference of the cross-linkable moiety with the binding reaction (Figure 11.2C) [79].

11.2.4 Molecular Parameters of Histone PTM–Reader Interaction

For determining binding parameters of histone–PTM interaction pairs, multiple methods have been described. Their applicability depends on the nature of the interaction (weak vs. strong, fast vs. slow), size difference of the interaction partners (peptide vs. protein), availability and accessibility of the reaction partners, possibility of introducing labels such as fluorophores, as well as other parameters. The quantification schemes generally rely on different readout methods of physical parameters that change with the titration of the interaction partners, that is from the unbound to the bound state.

Semiquantitative information with respect to histone PTM reader–ligand interactions can be obtained from several experiments such as electrophoretic mobility shift assays (EMSAs), analytical gel filtration, or analytical ultracentrifugation as long as the molecular volume or the hydrodynamic radius of the resulting reader–ligand complex can be discriminated from that of the individual components [80].

Surface plasmon resonance (SPR) enables the analysis of the binding kinetics (on- and off-rates) of histone peptides and nucleosomes with reader proteins [81]. When doing titrations, affinity parameters are also accessible. Traditionally, SPR measurements were performed with one histone peptide bound to the surface of the sensor chip via an affinity tag and one target reader domain at a time [82, 83]. More recently, modifications to the detection method have allowed the simultaneous imaging of different areas on the surface of the sensor chip [84, 85]. This way, SPR imaging can be coupled with peptide spotting onto the chip to interrogate reader interactions with over 100 different peptide baits (Figure 11.2C) [86].

Techniques with immobilized ligands can suffer from under- or over-estimation of binding strengths between readers and ligands due to probe orientation bias and background binding associated with the solid support. Thus, in-solution methods are most widely used to quantify thermodynamic parameters (dissociation constants) of the reader–ligand (peptides, nucleosomes) interaction. Tryptophan fluorescence spectroscopy measures changes in intrinsic fluorescence of tryptophan residues in the reader as a function of increasing ligand concentration [87]. The method can detect small conformational changes in the reader in the nanomolar regime and requires neither the ligand nor the reader to be labeled with fluorophores. However, the reader domain must contain a tryptophan residue directly involved in the binding event. This limits the application of the method [88].

Fluorescence polarization (FP) assays require the introduction of an external fluorescent tag to the smaller interaction partner, usually the ligand [27]. The method has been applied to investigate several interactions between histone peptides and reader proteins [3, 89–92]. Technological advances allow for high-throughput anisotropy imaging where more than 2000 different histone peptide–reader pairs can be screened in parallel [91]. FP relies on the size differences of ligand and reader

and is limited to smaller interaction pairs (i.e. peptides and protein domains). Microscale thermophoresis (MST), in contrast, can be used for studying interactions between readers and full-length histones or nucleosomes, either of which needs to be fluorescently labeled for detection (Figure 11.2C) [93–96]. High-throughput readout is possible when automated capillary handling robots are used.

Affinity tags and fluorescent labels may interfere with ligand–reader interactions [97, 98]. Isothermal titration calorimetry (ITC) is a label-free method that directly measures the enthalpy changes in the course of forming a ligand–reader complex. ITC has been used to quantify the binding events between modified histone peptides, mononucleosomes and nucleosomal arrays with isolated reader domains, and reader protein complexes [41, 81, 83, 99–101]. Automated systems can be used for high-throughput analyses.

With the introduction of highly sensitive MS instruments, a method for the quantification of histone PTM–reader interactions that does not require purified reaction components has recently been put forward. On the basis of affinity purification schemes, apparent binding affinities can be calculated from quantitative MS experiments, where increasing concentrations of immobilized peptides or nucleosomes are titrated against a constant amount of cell extract [63].

11.2.5 Cellular Assays to Characterize Histone PTM–Reader Interactions

While informative, the *in vitro* experimentation on isolated components and cellular extracts falls short in recapitulating the complex cellular chromatin-signaling-pathways of histone PTMs. Different approaches have been taken to study histone PTM–reader interactions in cellular context.

11.2.5.1 Visualizing Histone–Reader Interactions

Immunofluorescence (IF) allows analyzing the co-localization of histone PTMs and reader proteins in cellular context. In classical fluorescence microscopy, the limit of spatial resolution is about 200 nm. Various methods for super-resolution imaging have been introduced during the last years that overcome this diffraction-based barrier [102]. While IF cannot provide direct evidence for physical interaction between histone PTMs and readers, the close cellular co-localization is, nonetheless, very informative (Figure 11.3A). Also, IF based approaches such as fluorescence recovery after photobleaching (FRAP), fluorescence correlation spectroscopy (FCS), and Förster resonance energy transfer (FRET) can provide information about the dynamics and kinetics of cellular association [103–105]. High-throughput analyses of histone modification readout in live cells have been accomplished using automated single-cell imaging. This technique has been applied, for example, to visualize the distribution and abundance of histone variants and histone PTMs in isolated nuclei and in fixed cells [106–108]. While these reports were not directly aimed at visualizing reader-modification interactions, the high-throughput methodologies pave the road for such analyses in the future.

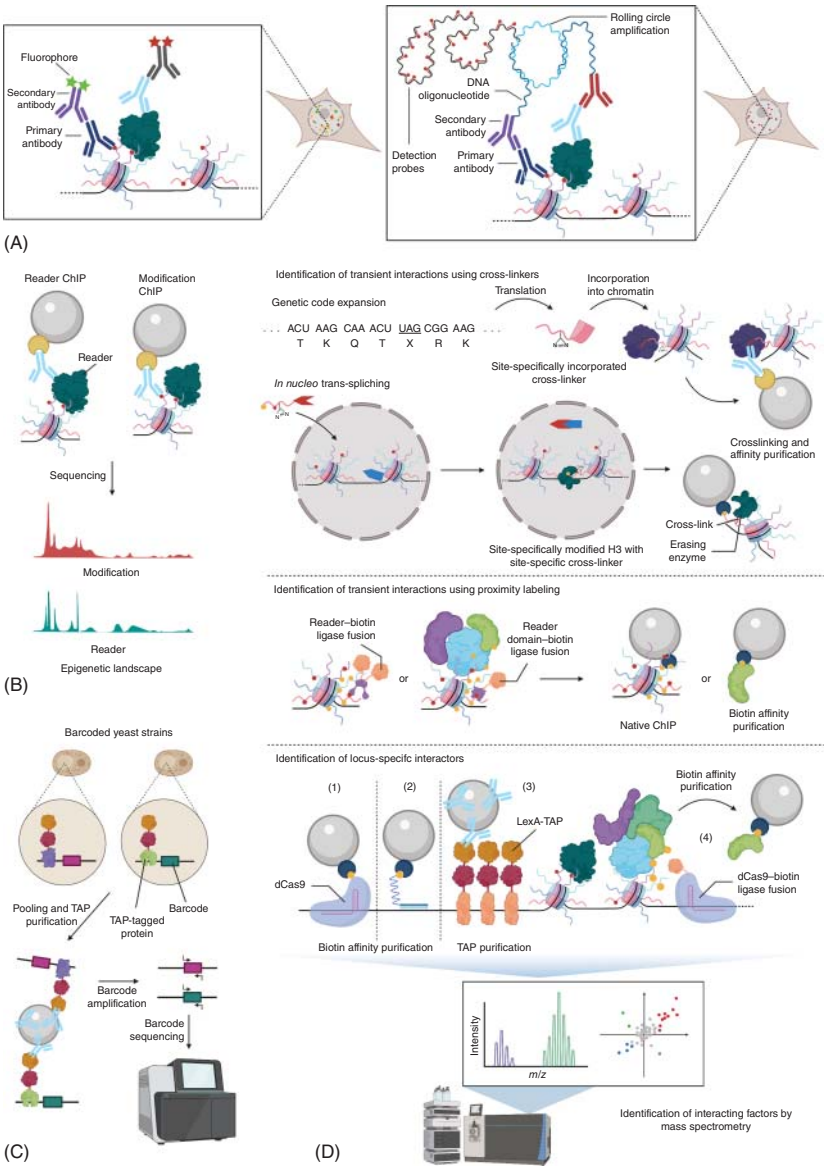


Figure 11.3 ***Cellular assays to identify and characterize histone PTM readout***.
(**A**) Co-immunofluorescence using two or more distinct fluorescence signals establishes proximity of the labeled targets (*left*). In proximity ligation assays (PLA), a fluorescence signal will only be generated if the detected components are less than 40 nm apart, as this enables rolling circle amplification of DNA for labeling (*right*). (**B**) Chromatin immunoprecipitation (ChIP) followed by sequencing (ChIP-seq) deduces the genome-wide distributions of chromatin modifications and chromatin interacting proteins.
(**C**) A TAP-tagged protein library is expressed in yeast strains carrying distinct barcodes at a specific locus. TAP purification of pooled samples, barcode amplification and sequencing inform about the pool of proteins interacting with the locus of interest. (**D**) Protein factors associated with a specific feature of interest are traditionally identified by mass-spectrometry-based methods. To capture native and transient interactions, crosslinkers are introduced into histone proteins using *in situ* ligation methods (*top*). Another approach to identify transient interactions or interaction partners of proteins, for which antibodies are not available, is proximity labeling using biotin ligase fusion constructs (*middle*). Factors interacting with a specific genomic region can be detected after purification of the locus using unique guide RNA and tagged dCas9 (1), tagged locked nucleic acids (2), or by engineering the locus with exogenous binding sites for factors like LexA (3). Purification of the locus can be circumvented by employing biotin ligase-fused dCas9, which biotinylates the local interactome (4) (*bottom*). This figure was generated with Biorender.com.

Proximity ligation assays (PLA) capture interactions between epitopes that are less than 40 nm apart. This is roughly the distance separating four consecutive nucleosomes. In PLA, specific primary and secondary antibodies are used to target a distinct reader protein and a particular histone modification. The secondary antibodies are linked to distinct DNA oligonucleotides, which anneal to connector oligonucleotides. If the two different secondary antibodies are in close proximity, these connector oligonucleotides can be ligated, amplified, and labeled with fluorescent probes, resulting in a hundred-fold increase in the interaction signal to visualize single-molecule interactions (Figure 11.3A) [109, 110]. Recently, high-throughput proximity ligation assay (hiPLA) was used to study interactions of (modified) histones with the nuclear lamina [111].

11.2.5.2 Chromatin Immunoprecipitation

Chromatin immunoprecipitation (ChIP) is a straightforward method to assess the local or global distribution of histone PTMs and reader proteins. In this approach, specific antibodies are used in affinity purification schemes to recover genomic material associated with a particular chromatin feature (Figure 11.3B) [112–115]. Since antibodies need to be of a specific quality for working in ChIP experiments, various consortia have established criteria and comprehensive lists of ChIP-verified reagents [116, 117]. Besides specific antibodies against the protein of interest, tagging in combination with well-characterized anti-tag antibodies can be used in the case of reader proteins. Examples include EGFP [118], the tandem affinity purification (TAP) tag [119], the localization and affinity purification tag [120], the hexahistidine and biotinylation signal tag [121], or the triple FLAG tag [122]. High-throughput recombination in yeast [119], or in mammalian cells [120], as well as the application of targeted genome-editing techniques such

as zinc finger nucleases or CRISPR/Cas9 [123–130] have facilitated large-scale tagging and analysis of modification readers and DNA-binding factors. Recently, protein trans-splicing was used to fuse cellularly expressed, truncated histones with modified, HA-tagged histone peptides intracellularly [37]. This approach could theoretically be used to ChIP an engineered histone carrying a defined modification.

When interrogating readers by ChIP, formaldehyde cross-linking is normally used to fix the chromatin association. In contrast, histone PTM ChIPs are often performed under native, non-crosslinked conditions. To obtain locus specific information with high resolution, chromatin is fragmented by sonication or enzymatic digestion (MNase-ChIP, ChIP-exo) before immunoprecipitation [112, 113, 131, 132]. Fragmenting chromatin and recovering genomic elements are major bottlenecks of ChIP experiments. Different approaches where enzymatic activities for cleaving the DNA (CUT&RUN, [133]) or for cleaving and tagging the DNA (CUT&TAG, [134]) are fused to the antibodies used in ChIP have been developed to overcome these hurdles. When testing the association of a chromatin feature with a defined genomic element, the readout of a ChIP experiment is done by directed, quantitative (real-time) PCR (ChIP-PCR). Alternatively, the recovered DNA is globally sequenced by next generation sequencing (NGS) methods (ChIP-seq) [135]. This establishes so called landscapes of epigenetic features (Figure 11.3B).

Specificity in ChIP experiments is probed by using a different antibody against the same target. Further validation can be obtained by using chromatin preparations from cells lacking the reader or the enzyme(s) depositing the histone PTM of interest [117, 122, 136].

ChIP and related schemes provide only indirect information about a reader's association with certain chromatin marks, since the two features are assessed in independent experiments with a relation via the identified, associated DNA sequence. In the classical approach, the local (ChIP-PCR) or global (ChIP-seq) signals of the reader are compared with those for histone modifications. A plethora of histone PTMs and reader proteins have been and are being mapped in different experimental systems (organisms, cell types, and cell lines). The data are available, for example, via the ENCODE and ROADMAP projects, or the Cistrome and NIH databases [137–139]. Sequential ChIP applies IP of one chromatin feature after the other and therefore allows unambiguous analysis of histone PTM reader co-distribution and interaction [140].

ChIP-MS is an extension of the classical ChIP protocol to define – besides the stretches of genomic DNA – proteins and complexes that are associated with a particular histone PTM or reader, either directly via protein–protein interaction or indirectly via co-association with a common chromatin element [67, 141, 142]. In ChIP-MS, the chromatin IPs are analyzed by MS (Figure 11.3B). The defined protein complexes originate from a mixture of genomic loci. To identify protein complexes that are found at a particular genomic locus, different chromatin affinity purification methods were developed: (i) The genomic locus of interest can be edited to contain protein-binding sites, for example, a LexA operator cluster. This enables affinity purification of the proteome associated with the edited locus via a protein that binds to this site (e.g. LexA-TAP, [143]). (ii) Nucleic acids complementary

to a locus of interest, for example desthiobiotin-immobilized locked nucleic acid oligonucleotide probes, can be used to enrich the proteome of a target element [144]. (iii) Specific guide RNAs of the CRISPR/Cas9 system can be used to target a tagged (e.g. Protein A, biotin or FLAG) inactive Cas9 (dCas9) to a locus of interest, which is then affinity purified (Figure 11.3F) [145–147].

To identify the interactome of certain genomic features without the need for MS, a yeast library, in which each yeast strain contains a single TAP-tagged protein, as well as a synthetic genetic array integrated into a specific locus and labeled with unique barcodes, has been used to study the general and specific chromatin interactome of the promoter and terminator region. After pooling the different yeast strains, TAP-ChIP is performed and the barcodes are sequenced to identify which proteins bind to the locus (Figure 11.3C) [148].

11.2.5.3 Cellular Labeling and Affinity Enrichment

Weak, transient, or erasing interactions between histone PTMs and their readers can be studied by introducing unnatural amino acids with photo-cross-linking potential into cells. For example, cells grown in medium containing photo-lysine as the sole source of lysine have been used to identify histone-interacting proteins after UV cross-linking followed by histone isolation and MS [149]. *In cellulo* genetic code expansion can be used to site-specifically incorporate modified, photo-cross-linkable amino acids into a histone, thereby enabling the identification of more localized interaction partners of either the soluble or chromatin-bound pool [150]. Both methods are not specific for the interactome of distinct PTMs, as these are established post cross-linker incorporation. To overcome this issue, synthetic histone tails harboring photo-cross-linkers next to defined histone modifications have been used in protein trans-splicing in isolated nuclei. This allowed trapping of transient reader-modification interactions that could not be recovered in native affinity purification schemes (Figure 11.3D) [78].

Another approach to enrich for transient interactions is proximity labeling. In this scheme, a protein of interest is fused to a biotin ligase that biotinylates proteins within a 10–20 nm radius. The tagged interactome of the target factor can then be isolated by affinity purification and analyzed by MS (BioID, APEX). When fused to a reader protein of interest, the histone PTM target pattern of the bait and/or factors associated with the alleged main target PTM of the factor can be identified (ChromID) (Figure 11.3E) [151, 152]. The scheme has also been used to identify proteins interacting with particular genomic loci, for example, repeat regions or specific promoters by fusing the biotin ligase to dCas9 that is recruited to the target regions by specific guide RNAs (Figure 11.3F) [153, 154].

11.3 Interaction of Proteins with Modified Nucleic Acids

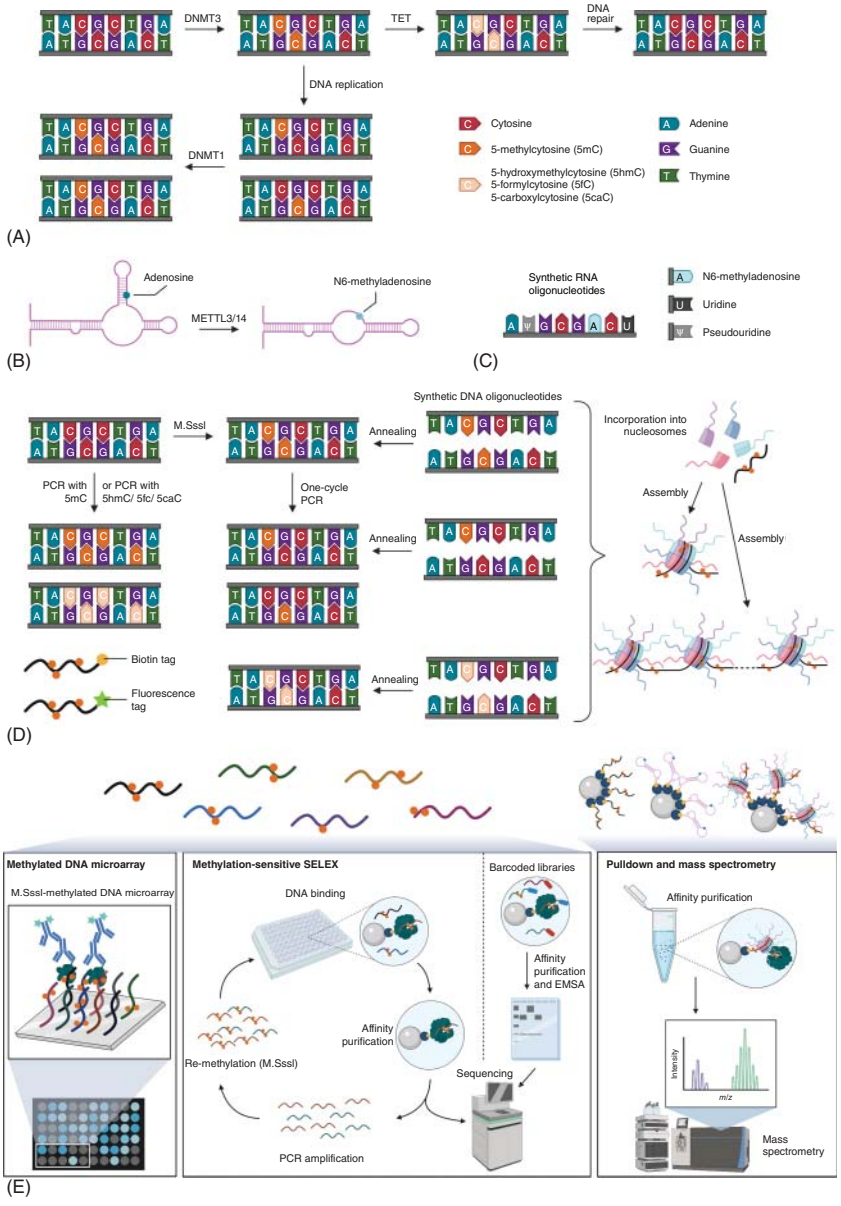
11.3.1 Discovery of DNA Methylation and the First Reader Proteins

Mammalian DNA is preferentially methylated on the 5 position of cytosine residues within CpG dinucleotides (5mCpG). Generally, the modification is found

symmetric on both DNA strands, and after DNA replication the originating hemimethylated state (only one strand of DNA in each daughter DNA molecule methylated) is converted to the fully methylated state due to the palindromic nature of the CpG sequence. In mammals, CpG dinucleotide sequences are generally underrepresented in the genome, but cluster (i.e. appear with higher than random frequency) in so called CpG islands of few hundred base pairs upstream of many housekeeping genes [155, 156]. Non CpG-directed methylation of cytosines also exists with various abundance in different mammalian cell types [157, 158], and in particular in plants [159]. While DNA methylation is abundant in some organisms (e.g. mammals, plants), it is of low frequency in other species.

The presence of 5mC in DNA was described as early as in 1925 in *Mycobacterium tuberculosis* [160] and 1948 in mammals [161]. In mammals, DNA methylation is established by the DNMT3A and B and maintained by the DNMT1 enzymes. The modification is not directly reversible but is removed via the oxidizing activities of TET enzymes (intermediates are 5-hydroxymethyl C [5hmC], 5-formyl C [5fC], and 5-carboxyl C [5caC]), deglycosylation, and base excision repair (Figure 11.4A) [162]. Methods to detect the different forms of DNA methylation with single base-pair resolution (bisulfite sequencing, BS-seq, and related approaches) have enabled the precise mapping of whole-genome DNA methylation patterns. The effects of DNA methylation, which include transcriptional repression, regulation of splicing, alterations in chromatin structure, and the activation of DNA repair, are mediated through effects on nucleosomal DNA wrapping [163] and on interacting proteins, highlighting the importance for studying and understanding the mCpG proteome [164].

Figure 11.4 ***Studying interaction partners of modified DNA and RNA.*** (**A**) In mammalian cells, cytosine in the palindromic CpG dinucleotides is symmetrically 5-methylated by DNMT3. Active demethylation is oxidatively performed by TET enzymes via 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine followed by DNA damage repair. DNA replication results in the generation of hemimethylated DNA (methylated parent, unmethylated daughter strand). DNMT1 targets hemimethylated DNA, converting it to fully methylated DNA. (**B**) Specific adenine residues in RNAs are targeted by the METTL3/14 complex. Methylation at N6 can result in changes of the RNA secondary structures. (**C**) RNA modifications can be introduced during chemical solid-phase oligonucleotide synthesis. (**D**) Differently methylated DNA is generated by several methods that result in site-specific incorporation (annealing of synthetic ssDNA oligonucleotides), methylation of all CpG sites (enzymatic methylation by M.SssI), or all Cs in the sequence (PCR with modified dCTP). Hemimethylated DNA is either generated by one-cycle PCR, or through annealing of a methylated and a non-methylated DNA oligonucleotide. Functionalization (e.g. biotin or fluorescent tags) is accomplished via DNA oligonucleotide synthesis (used for PCR or in annealing reactions). Modified DNA can be incorporated into "designer chromatin" analogous to unmodified DNA. (**E**) DNA oligonucleotides are used to determine the binding specificities of known readers on (methylated) DNA microarrays or in methylation-sensitive SELEX approaches in solution. Bead-bound DNA, RNA, and nucleosomes are used as baits for identifying novel modification interactors using mass spectrometry. This figure was generated with Biorender.com.



Initial evidence for specific mCpG-binding proteins came from nuclease protection assays. Methylation-insensitive restriction enzymes did not cleave most mCpG-containing motifs in intact nuclei, but behaved like methylation-sensitive nucleases, indicating that protein factors protect mCpG sites [165]. The first reader protein for mCpG, MeCP1, was then described on the basis of EMSA and competition experiments using methylated and unmethylated oligonucleotide probes [166]. It later became apparent that MeCP1 is a nine-subunit protein complex (MBD2-NuRD) [167–170]. MeCP2 was incidentally discovered after performing southwestern blotting (probing a protein blot membrane with methylated and unmethylated DNA oligonucleotides) in an attempt to further characterize MeCP1 [171]. Other mCpG-binding domain (MBD)-containing proteins were identified by homology searches [172, 173]. Several proteins have now been described that can be specifically recruited to, or are repelled by, symmetrically methylated, hemimethylated, or unmethylated CpG, as well as by the products of 5mC oxidation, 5hmC, 5fC, and 5caC [174–177].

11.3.2 RNA Modifications

The first modified RNA base, pseudouridine, was detected in 1957 and characterized in 1959 [178–180]. Until today, more than 150 RNA base modifications in mRNAs and in ncRNAs have been described [181] (see for example MODOMICS database [182]).

For mRNAs and transcriptional regulation (epitranscriptomics), the most extensively studied modification is N6-methyladenosine (m^6A) [183]. The METTL3/14 complex has been identified as the methyltransferase mediating m^6A in RNA [184]. Similar to mCpG, the m^6A modification is not directly reversible but is removed via stepwise oxidation to hm^6A and f^6A [185]. m^6A and other RNA modifications elicit their functions through the modulation of protein–RNA interactions. Besides this, modified RNA bases might influence the relative abundance of RNA secondary structures by affecting base pairing energies [186]. This may in turn impact on protein–RNA interactions and RNA functions (Figure 11.4B).

11.3.3 Modified DNA and RNA Templates

Methylated or otherwise modified libraries of random DNA or RNA sequences are generated by using modifying enzymes (for example by M.SssI for mCpG) or via incorporation of modified dCTP in PCR amplification. These schemes result in modification of all CpG sites or all Cs, respectively. Site-specific methylation is accomplished via incorporation of modified nucleotides during solid-phase oligonucleotide synthesis [187, 188]. Oligonucleotides can further be specifically labeled with affinity or fluorescent tags for pulldown, thermodynamic, or kinetic studies. Hemimethylated DNA is generated either by annealing-based techniques of complementary methylated and unmethylated single-stranded DNA or by one-cycle PCR of fully methylated double-stranded templates [188, 189]. Modified DNA can be assembled into designer chromatin (mononucleosomes, nucleosomal arrays) (Figure 11.4C,D).

11.3.4 *In Vitro* Assays for Identifying Readers of Nucleic Acid Methylation

11.3.4.1 Affinity Purification to Identify Novel Modification Readers

Synthetic DNA or RNA oligonucleotides, mononucleosomes, and nucleosomal arrays carrying distinct modifications (e.g. CpG symmetrically methylated or hemimethylated, templates containing oxidation products of 5mC or m^6A) have been used as baits to pull down proteins from cell or tissue extracts to identify novel interactors [74, 175, 187, 190–195]. After purification and for identification, proteins are subjected to LC-MS/MS (Figure 11.4E) [196]. In a wider approach, libraries of baits have been screened in plate format to determine the effect of sequence variation on DNA–protein interaction [197]. Also, an MS-based method for determining the apparent Kds for the interaction of hundreds of proteins with DNA on a proteome-wide scale has been developed [63]. For these experiments, nucleotide sequences are chosen randomly or can be modeled after cellular sequences (e.g. specific CpG islands, promoters).

Affinity purification studies generally identify different types of reader proteins: (i) proteins that specifically associate with one or more of the modifications, (ii) proteins that are repelled by the modification(s), and (iii) proteins that are recruited to the modification in specific contexts (e.g. specific sequences, symmetrically methylated, or hemimethylated DNA) [196].

11.3.4.2 Characterizing Binding Specificities of Known Readers

To define methylation-dependent binding specificities and to identify methylation-sensitive sequence motifs of known (DNA- and RNA-binding) proteins, *in vitro* selection of modified or unmodified oligonucleotide libraries or purified nucleic acids has been employed. Single to hundreds of individual recombinant proteins, either on microarrays or in solution, are incubated with DNA or RNA libraries. Protein-associated oligonucleotides are isolated and processed for identification by sequencing or MS (Figure 11.4E) [198–200]. Alternatively, DNA microarrays are methylated and bound recombinant (tagged) proteins can be detected using immunostaining (Figure 11.4E) [177].

Methylation-sensitive SELEX (systematic evolution of ligands by exponential enrichment) has been applied to enrich for high-affinity binding sequences containing meCpG [198]. Proteins are incubated with DNA libraries as described above, but instead of sequencing the DNA directly, it is amplified and re-methylated. This enables repeating the cycle of protein binding, purification, and PCR amplification to sequentially enrich high affinity ligands.

Using SELEX to analyze methylated and unmethylated sequences as separate pools can mask the effect of methylation on binding if a factor absolutely requires a CpG in its binding motif. Different techniques have been developed to overcome this limitation: (i) In bisulfite SELEX, the selected libraries are pooled, subjected to one more round of selection, and sequencing of the initial pool as well as the selected pool informs about a factor's methylation specificity [198]. (ii) EpiSELEX-seq and (iii) Methyl-Spec-seq make use of two distinctly barcoded libraries for methylated

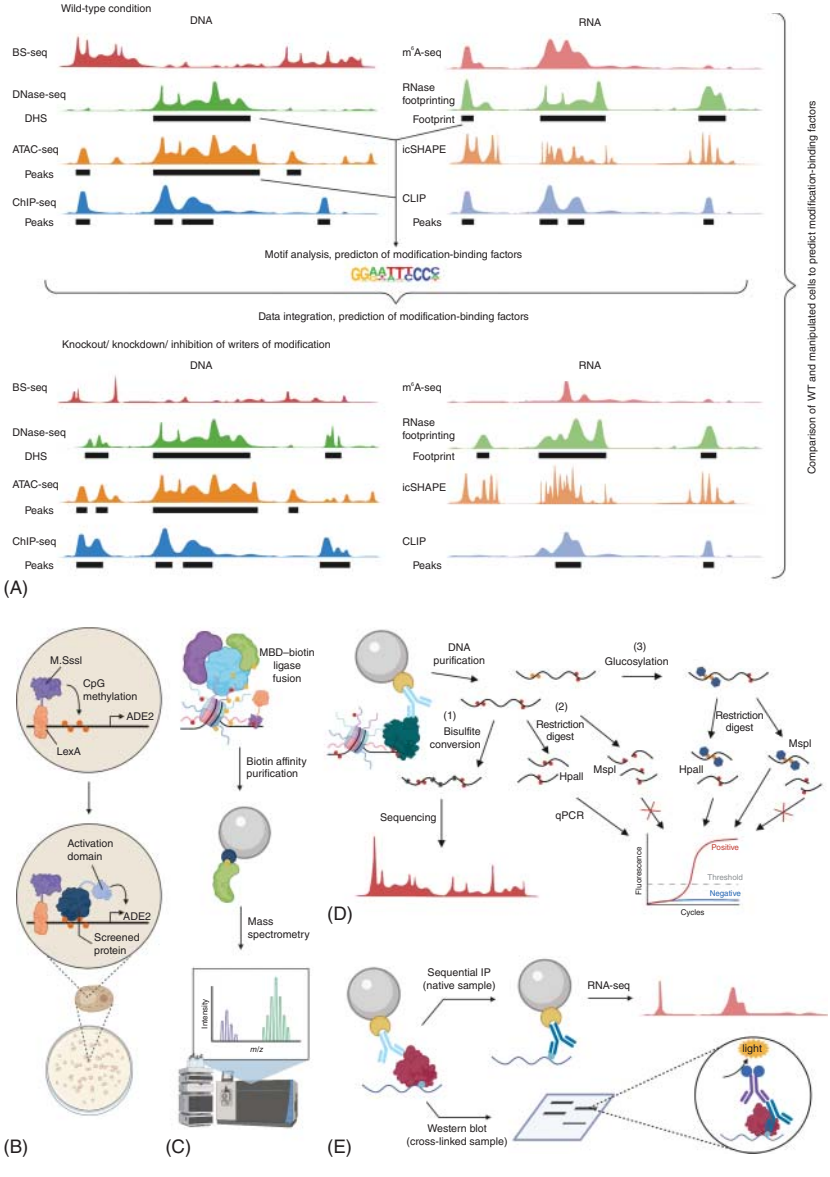
and unmethylated sequences. These are pooled, incubated with the protein of interest, and EMSAs are used to isolate bound DNA for sequencing (Figure 11.4E, bottom panel 2). The effect of methylation on binding is then derived from the relative fold-enrichments between input, or unbound, and bound fractions [201, 202].

11.3.5 Cellular Assays for Identifying Readers of Nucleic Acid Modifications

Analogous to comparing ChIP-seq tracks of histone marks and of chromatin-binding proteins for deducing associations of proteins with specific marks, results from different NGS-based methods can be aligned to predict DNA and RNA modification-binding proteins [177, 194, 203–206]. DNA and RNA modifications are mapped genome-wide using BS-seq or antibody-dependent DNA and RNA modification sequencing methods [207]. Protein binding sites are identified through ChIP-seq, RNA immunoprecipitation-based methods (RIP), and DNA or RNA accessibility (DNase hypersensitive sites [DHS], ATAC-seq, and RNase footprinting) (Figure 11.5A) [208–211]. Publicly available datasets for these kinds of studies can be downloaded from the ENCODE or ROADMAP project databases.

More direct evidence of modification-sensitive binding can be obtained when such assays are combined with disruption of the modification by pharmacological

Figure 11.5 ***Identifying interaction partners of modified nucleic acids in vivo***. (**A**) Different methods that determine the genome-wide distribution of factors and modifications in wild type (*top*) or mutated/depleted (*bottom*) conditions are compared and integrated to predict factors that may interact with, or be repelled by modified DNA or RNA bases. BS-seq: bisulfite sequencing (distribution of 5mC); DNase-seq: DNase hypersensitive sites (chromatin accessibility); ATAC-seq: assay for transposase-accessible chromatin using sequencing (chromatin accessibility); ChIP-seq: chromatin immunoprecipitation with sequencing (distribution of chromatin modifications and factors); m^6A-seq (distribution of m^6A); RNase footprinting (RNA accessibility); icSHAPE: *in vivo* click selective 2-hydroxyl acylation and profiling experiment (RNA secondary structure); CLIP: cross-linking immunoprecipitation (distribution of RNA-binding factor). (**B**) The methylation-sensitive yeast one-hybrid system makes use of the absence of DNA methylation in *S. cerevisiae* and allows expression library screening of CpG methylation binding factors via a gene expression and colony growth readout. (**C**) A construct of mCpG-interacting MBD domains fused to biotin ligase enables proximity labeling and subsequent identification of factors bound to mCpG-associated loci. (**D**) Different ChIP-based approaches allow deducing whether a protein interacts with methylated or unmethylated loci. (1) Bisulfite sequencing determines if the DNA sequences ChIPed with the protein of interest are methylated or unmethylated. (2) Locus-specific methylation at CCGG sites can be assessed via PCR after differential restriction enzyme digest using HpaII (methylation-sensitive) and MspI (methylation-insensitive). (3) hmCpG in CCGG sites can be detected after specific enzymatic glucosylation of hmC that inhibits MspI-digestion. (**E**) To determine whether a factor directly interacts with modified RNA, sequential immunoprecipitations (IPs) for the protein and the modification of interest are performed, followed by RNA sequencing. Alternatively, protein IP can be done on cross-linked samples and the RNA modification is detected in western blot. This figure was generated with Biorender.com.



inhibitors (for example 5-aza-deoxycytidine for CpG methylation), knockdown, knockout, or overexpression of modifying enzymes (e.g. DNMTs or regulatory factors) (Figure 11.5A) [212–216]. Using DNA methylation-deficient TKO cells (i.e. devoid of DNMT1, DNMT3A, and DNMT3B), sequence motifs and their corresponding methylation-sensitive transcription factors were deduced by analyzing DHSs unique to these cells as compared to wild type [217]. In a similar manner, cellular differentiation can inform about methylation-sensitive binding factors of the genome. During differentiation, DNA methylation changes at specific loci and absence or presence of candidate factors at these sites can indicate methylation-sensitive recruitment [218, 219].

Another direct approach for screening for factors that bind chromatin in a DNA methylation-sensitive manner is the use of a yeast one-hybrid system. Since *S. cerevisiae* is deficient in DNA methylation, targeting of the methyltransferase M.SssI to a bait sequence leads to specific CpG methylation of this region (Figure 11.5B). Whole cDNA libraries can then be screened for binding to the methylated bait sequence [220].

As described for histone PTMs, proximity labeling (ChromID) can be used to identify the interactome of meCpG containing regions of the genome. In this case, a protein containing an MBD domain (e.g. MBD1) is fused to the biotin ligase for recovering interacting proteins (Figure 11.5C) [151]. Also, chromatin enrichment for proteomics (ChEP) that monitors the global chromatin proteome has been put forward [221]. The scheme can be combined with perturbations of the DNA methylome to identify factors that potentially interact with chromatin in a methylation-dependent manner.

To determine the level of DNA methylation associated with a given factor, DNA recovered by ChIP can be subjected to bisulfite conversion before sequencing [222–225]. Another approach utilizes methylation-sensitive restriction enzymes to digest DNA after ChIP. The isoschizomeric enzymes MspI (not methylation-sensitive) and HpaII (methylation-sensitive) are widely used. If a candidate factor coprecipitates unmethylated DNA, it will be digested in both reactions. In consequence, and different from a factor binding methylated DNA, no PCR amplification or sequencing will be possible [226]. Using the same assay, hmC can be detected after selective enzymatic glucosylation, which makes the hmCpG site resistant to MspI digestion. hmC-containing DNA can thus be specifically amplified after enzymatic treatment (Figure 11.5D) [227].

For RNA modification, sequential IPs for the protein of interest and the nucleic acid modification followed by sequencing have been performed (PAR-CLIP-MeRIP, [216]).

RNA crosslinking and immunoprecipitation of the protein of interest followed by detection of the RNA modification in western blot can be applied to investigate the interaction of a protein with an RNA modification independent of identifying the specific RNA it interacts with (Figure 11.5E) [204]. The reverse approach, RNA pulldown to identify interacting proteins (eRIC [enhanced RNA interactome capture]), has been used to characterize the poly-A RNA-binding proteome [228]. Indirect effects of modifications mediated via RNA structural changes or a newly

available motif are assessed using icSHAPE (*in vivo* click selective 2-hydroxyl acylation and profiling experiment) and can be integrated with cross-linking IP (CLIP) data [229]. By combining these methods with perturbations in m^6A levels (e.g. via knockout of METTL3/14 or oxidase inhibitors) and m^6A-seq, the m^6A-dependent RNA-interactome was probed [228].

11.4 UHRF1 as an Example of a Multidomain Reader/Writer Protein of Histone and DNA Modifications

There is ample interplay within and between the different histone PTM and nucleic acid modification systems [71, 230–232]. These do not only interface functionally (i.e. mediating similar or opposing effects) but also biochemically. Many chromatin modification enzymes are highly sensitive to the preexisting modification status (i.e. stimulating or repressive effects). Also, multivalent binding proteins and multiprotein complexes exist that have the ability to recognize different chromatin modifications simultaneously (i.e. patterns of modification). We use the example of the ubiquitin-like containing PHD and RING finger domains 1 protein (UHRF1) to illustrate the intricacies of chromatin modification crosstalk with particular focus on the experimental approaches that have uncovered the working mode of this epigenetic regulator.

UHRF1 is an essential protein, which is required for DNA maintenance methylation and embryonic development. The factor consists of five distinct domains: a ubiquitin-like (UBL) domain, a tandem tudor domain (TTD), a plant homeo domain (PHD), a SET and RING-associated (SRA) domain, and a really interesting new gene (RING) domain with E3 ubiquitin ligase activity (Figure 11.6A). Initially, cellular fractionation, co-IP, and far western blotting using purified histones and polynucleosomes identified UHRF1 as a chromatin-associated protein with binding preference for, and ubiquitination activity toward H3 (Figure 11.6B) [233]. Later, different experimental approaches characterized its SRA domain as a hmCpG-binding domain with additional affinity toward symmetric caCpG: (i) pulldowns with a PCR-amplified and M.SssI-methylated endogenous target, (ii) EMSAs, including EMSA-western analysis, EMSAs with radioactively labelled DNA, and competition EMSAs with fluorescently tagged oligonucleotides, (iii) colocalization studies by IF, (iv) crystallization of the SRA domain-DNA complex, (v) molecular modeling and dynamics simulations, and (vi) biochemical assays, including MST with oligonucleotides (Figure 11.6C) [174, 234–242].

In an unbiased pulldown screen using H3K9me0/2 peptides immobilized on beads, nuclear extracts and MS, UHRF1 was identified as an H3K9 methylation reader [243]. Several follow-up studies established that the protein's TTD domain recognizes H3K9me, whereas the PHD domain interacts with the very N-terminus of unmodified H3. The two domains can synergize in binding the H3 tail in a manner that is regulated by UHRF1 posttranslational modification, allosteric ligands, and alternative RNA splicing. The insights into the chromatin binding modes were derived from: (i) structural studies, including crystallization of the domains in

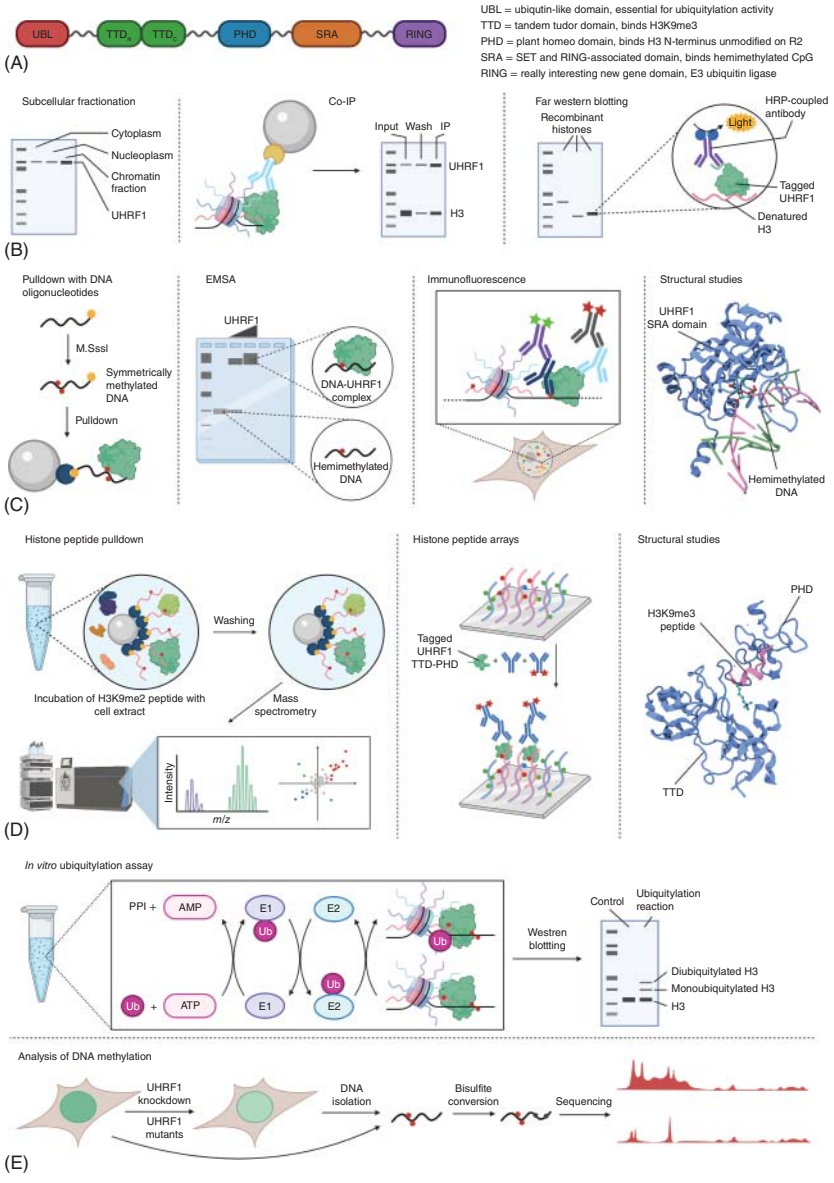


Figure 11.6 ***Summary of methods used to characterize the multidomain chromatin reader protein UHRF1***. (**A**) Schematic representation of the domain structure of UHRF1 and known functions of the individual domains. (**B**) Methods used to identify UHRF1 as a chromatin- and H3-binding protein. (**C**) Methods used to identify UHRF1 as an mCpG-binding protein with preference for the hemimethylated state. Structure of the SRA domain in complex with hemimethylated CpG sites (PDB ID: 3F8I). (**D**) Methods used to identify UHRF1 as an H3K9me binding factor. Structural studies showed how the combined TTD-PHD module interacts with the H3K9 methylated tail peptide (PDB ID: 4GY5). (**E**) To understand UHRF1 functions, its enzymatic activity was probed in ubiquitylation assays with a variety of substrates. Maintenance of DNA methylation is the main function of UHRF1 described in the literature. Loss of UHRF1 or mutation of its domains are associated with a loss in DNA methylation as assessed by bisulfite sequencing or microarray. This figure was generated with Biorender.com.

complex with H3K9me3 peptides, molecular modeling, NMR, and SAXS studies, (ii) peptide pulldowns and peptide arrays with hundreds to thousands of peptides carrying single and combinatorial modifications using full-length protein, isolated domains, or deletion constructs, and (iii) biochemical assays, including native gel electrophoresis, FP, ITC, co-IP, and MST (Figure 11.6D) [23, 97, 244–255].

To understand UHRF1 functions in the cell, interaction partners were identified in high-throughput assays, including yeast two-hybrid screens [256] and UHRF1 pulldown experiments [257]. Further, a plethora of functional assays using wildtype and mutant proteins or protein domains were performed: (i) co-localization with specific chromatin features by IF, (ii) association with chromatin by fractionation or FRAP experiments, (iii) altering the availability of UHRF1 binding partners by overexpression of methyltransferases, demethylases, by introducing blocking peptides or by knockdown, (iv) analysis of DNA methylation by dot blot, ELISA, or BS methods, and (v) *in vitro* ubiquitylation assays (Figure 11.6E) [97, 98, 243, 244, 257–266].

The overall picture emerging from the different lines of investigation is that in DNA maintenance methylation (i.e. after DNA replication), UHRF1 senses hemimethylated DNA via its SRA. Via unknown mechanisms, this activates the protein's H3 ubiquitin ligase activity [98]. H3ub is a signal for recruitment and activation of DNMT1 [267]. To what degree H3K9me recognition is interfacing with this pathway is not fully clear. Also, the role of UHRF1's chromatin modification sensing and modifying activities in DNA damage repair and transcriptional regulation are still intensively investigated.

11.5 Histone Chaperones and Chromatin Remodeling Complexes

11.5.1 Chromatin Assembly and Remodeling

Histone chaperones and ATP-dependent chromatin remodeling complexes work together to assemble, organize, and position nucleosomes along the DNA sequence. Nucleosomes are assembled in a step-wise manner, with the H3-H4 tetramer associating with DNA before addition of the H2A-H2B dimers, which can thus be more easily removed from the nucleosome (Figure 11.1A).

Chaperones are proteins that associate with free histones to neutralize their positive charge in order to prevent nonspecific interactions (Figure 11.1A). *In vitro*, all histone chaperones stimulate ATP-independent nucleosome assembly without being part of the final complex; *in vivo*, some chaperones only facilitate histone storage or transport [268].

Chromatin remodelers are enzymes that modulate the position of nucleosomes (sliding), change the composition of nucleosomal histones (i.e. incorporating histone variants), or orchestrate disassembly and reassembly of nucleosomes at nonadjacent positions (transfer). All known chromatin remodeling complexes contain an ATP-dependent helicase/translocase of the Snf2 family to facilitate translocation

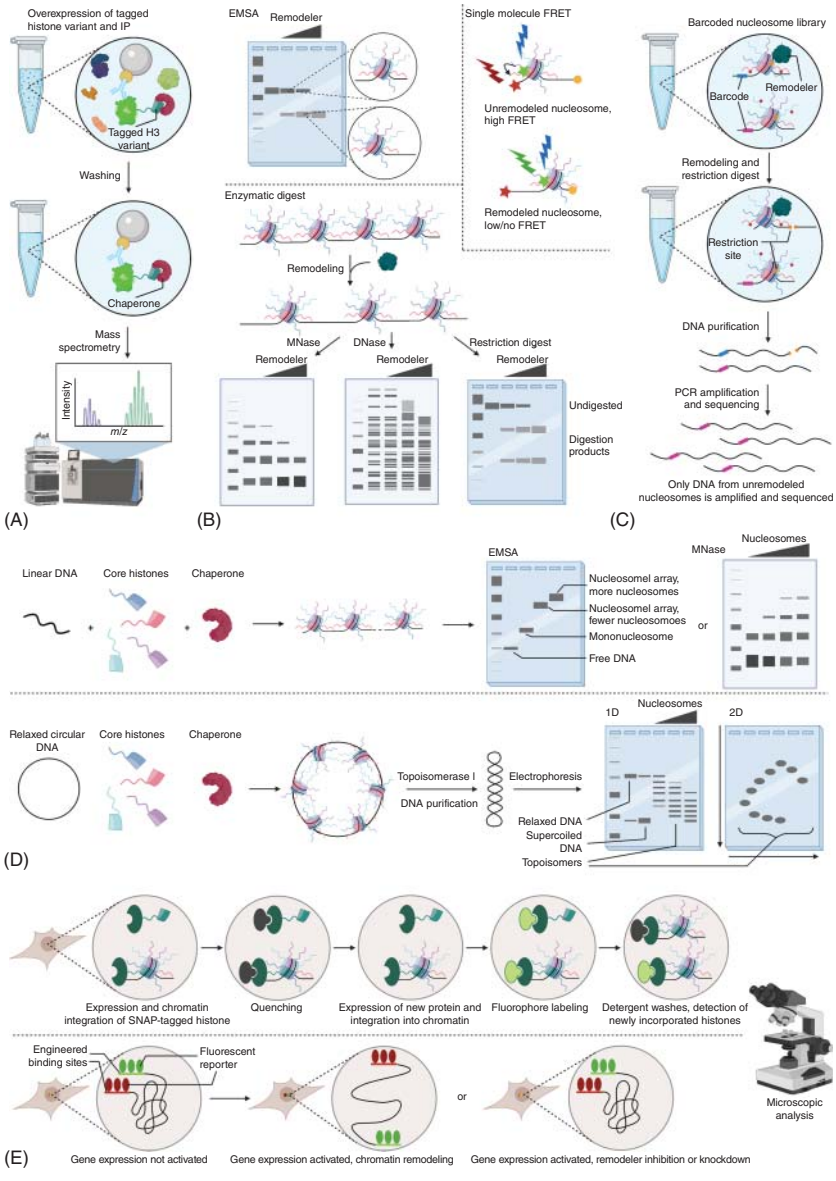
of DNA around the histone core. Additional domains or proteins regulate the ATPase and mediate specific targeting, for example, via recognition of chromatin modifications [269, 270].

11.5.2 Discovery of Histone Chaperones and Chromatin Remodelers

Analysis of supercoiling of circular DNA is the most widely used assay to study nucleosome assembly and thus chaperone and chromatin remodeler activity. Circular, supercoiled DNA is relaxed by the addition of topoisomerase I. When nucleosomes are assembled on the DNA, enzyme-mediated relaxation is incomplete. The number of superhelical turns remaining in the purified DNA is indicative of the number of nucleosomes assembled. Differently supercoiled species (topoisomers) are separated by 1D or 2D gel electrophoresis [271, 272]. Specific MNase, DNase, and restriction enzyme digestion patterns can also be used to probe nucleosome assemblies (Figure 11.7D).

When purified DNA and histones are directly mixed at physiological ionic strength, they form insoluble precipitates. Assembly into nucleosomes is, however, possible by mixing DNA and histones at 2 M salt and slowly dialyzing to physiological salt concentrations [273] or when negatively charged molecules like pectin, polyglutamic acid, or RNA are added [274, 275]. Also, it was found that addition of cell-free *Xenopus laevis* egg extract to mixtures of DNA and histones facilitates rapid *de novo* nucleosome assembly [276–278]. Nucleoplasmin was the first component identified in this experimental system that is required for chromatin assembly [279]. This was followed by the characterization of several other chaperones in *Xenopus* egg extracts [280–283] and from other sources that promote nucleosome assembly *in vitro* in a constitutive or replication-dependent manner [284–288].

Figure 11.7 ***Studying histone chaperones and chromatin remodeling enzymes.*** (**A**) Overexpression of tagged histone (variants) is used to identify chaperones by coprecipitation and mass spectrometry. (**B**) Chromatin remodeling can be studied *in vitro* using recombinant mono- or oligonucleosomes. Nucleosome sliding and histone eviction can be measured by electrophoretic mobility shift assays (EMSA), by single-molecule Förster resonance energy transfer (FRET), and by MNase, DNase, or restriction digest. (**C**) To study effects of histone modifications or mutations on remodeler activity, barcoded nucleosome libraries containing restriction sites that only become accessible after remodeling are used. (**D**) Chromatin assembly by chaperones is studied on linear or circular DNA templates in combination with gel electrophoresis. Native gel electrophoresis distinguishes different levels of chromatin assembly directly (*top*), while supercoiling assays provide an indirect readout via topoisomerase activity (*bottom*). (**E**) Chaperone function in cells can be studied using pulse-chase experiments of SNAP-tagged histones that are incorporated into chromatin and labeled with fluorescent ligands (*top*). Remodeler activity is assessed in cells engineered to contain two loci with DNA binding sites for fluorescent reporters (e.g. LacI, TetR) on either side of a region of interest, for example a promoter. The distance of the reporters informs about chromatin decompaction and remodeling activity (wild type compared to inhibited or knock down (kd) conditions). This figure was generated with Biorender.com.



The first chromatin remodeler was initially described in yeast as a gene important for mating-type switching (SWI) and sucrose fermentation (SNF) with general activating function on transcription [289]. Later, a multi-subunit SWI/SNF complex was chromatographically purified from cell nuclear extract and characterized as an ATP-dependent chromatin remodeling complex. Nucleosome-disrupting activity was demonstrated in various assays. (i) Reconstituted nucleosomes were incubated with the purified SWI/SNF complex and treated with DNase I; in the presence of ATP, the DNase I digestion pattern changed. (ii) Supercoiling assays followed by 2D gel electrophoresis indicated a SWI/SNF- and ATP-dependent loss of nucleosomes from circular DNA. (iii) DNase foot-printing showed 10- to 100-fold enhancement of transcription factor binding to nucleosomal DNA in the presence of SWI/SNF and ATP [290].

In the following years, other chromatin remodeling enzymes were described based on homology to the Snf2 helicase domain [291–296]. Interestingly, some of these factors were independently identified in genetic screens for factors involved in the regulation of specific phenotypes implying that these factors are important regulators of genome readout [297–300].

11.5.3 Methods for Identifying Histone Chaperones and Remodeling Factors

11.5.3.1 Immunoprecipitation Assays

To identify histone chaperones and remodeling factors, including factors specific for distinct histone variants, endogenous histones have been singly or dually tagged, for example with EGFP, FLAG-HA, or the TAP tag, followed by single or sequential affinity purification under native or crosslinking conditions (Figure 11.7A) [301–307]. Also, histone complexes, for example H3-H4, have been used as affinity baits for enrichment of candidate factors [308, 309]. To identify chaperones and remodelers that interact with different histone variants in a chromatin environment, chromatin is fragmented (native or crosslinked) and tagged histones are pulled down [305]. In the reverse approach, factors of interest are immunoprecipitated from cell extract after chromatin fragmentation. Associated histones (and other interacting proteins) can be detected by MS or western blot [310, 311].

11.5.3.2 Computational Methods

Chaperones may be predicted based on functional requirements: DNA replication, repair, and transcription require nucleosome eviction and assembly, and thus chaperones. This assumption resulted, for example, in the identification of MCM2-FACT as a histone chaperone complex in the eukaryotic replisome [312] and the histone fold-containing subunits of POLε and RPA as replication-dependent H3-H4 chaperones [313–315].

Since all known remodeling factors contain an enzymatic ATPase subunit of the Snf2 family within the helicase superfamily 2, homology searches have been the main tool to identify new proteins with this chromatin modulating activity. Whereas more than 1300 Snf2 family proteins, divided into 24 subfamilies, were identified

from protein sequence data (UniProt, UniRef100) of eukaryotes, eubacteria, and archaea [316], not all Snf2-type ATPases are functional in nucleosome remodeling [317].

11.5.4 Assays to Study Chaperone and Remodeler Activities

In vitro remodeling assays with ATP-dependent complexes and/or chaperones can be performed on nucleosome or chromatin array substrates: (i) the distinct electrophoretic mobility of differently positioned nucleosomes can be analyzed by EMSA, (ii) the accessibility of restriction sites or general DNase accessibility of nucleosomal DNA can be assessed, (iii) MNase digest reveals nucleosome positioning on chromatin arrays, and (iv) single molecule FRET assays, with one fluorophore of the FRET pair attached to the DNA and the other to one of the histones, and (v) FP assays with the fluorophore attached to the end of the DNA, can inform about nucleosome sliding and histone eviction (Figure 11.7B) [318–328]. These kinds of assays can also be used for high-throughput screening.

Barcoded nucleosome libraries containing differently modified or mutated nucleosomes, as well as nucleosomes containing histone variants, can be designed with a restriction site in the nucleosomal DNA that becomes only accessible upon nucleosome remodeling. The library is simultaneously incubated with chromatin remodelers and a restriction enzyme, and DNA is isolated and sequenced at different time points to inform about remodeling kinetics and remodeler preferences for certain nucleosome modifications (Figure 11.7C) [329].

To confirm nucleosome assembly, and thus chaperone activity *in vitro*, purified chaperones, histones, and DNA, or intermediate complexes of nucleosome assembly are incubated together. DNA-histone complexes can be identified by gel filtration, EMSA-type experiments, the analysis of MNase digest, or the plasmid supercoiling assay described above (Figure 11.7D) [309, 313, 315, 330, 331].

11.5.5 Cellular Assays

Chaperones are defined as having chromatin assembly activity *in vitro*. As this does not necessarily translate into *in vivo* function, it is important to study their effects in cellular systems. Histone ChIP and ChIP-seq studies have been combined with chaperone knockout to assess the effect of chaperones on preventing or facilitating histone-DNA interactions/incorporation into nucleosomes [304, 307, 311, 332–334]. ChIP can also inform about genomic targets of specific chaperones [307, 311, 335, 336]. To assess chromatin remodeler and chaperone function, mutations, knockdown, knockout, or specific protein degradation are combined with studies of chromatin accessibility (MNase-seq, ATAC-seq) and ChIP(-seq) for the proteins of interest [337–339].

To assess histone deposition *in vivo*, affinity- or fluorescence-tagged histones can be transiently expressed, or SNAP- or CLIP-tagged histones can be used for fluorescent pulse-labeling and chasing [340, 341]. Global histone incorporation into DNA is measured by western blot or by fluorescence microscopy after detergent

washes to remove free histones (Figure 11.7E). Histone mobility can be measured by FRAP. Combining these assays with the knockdown of chaperones or remodelers provides information about the role of individual proteins/complexes for histone incorporation [342]. The histone residues required for specific histone-chaperone interaction have been probed by expression of histone mutants and assessment of their interaction with histone chaperones by pulldown and western blot [305, 343].

In yeast cells, MNase-ChIP-seq revealed nucleosome-specific positions of remodelers genome-wide, highlighting that some complexes associate preferentially with the 5′ or 3′ ends of genes, with specific genic nucleosomes (−2, +1, +2 to +4 positions, etc.), and with specific genomic regions [344]. Combining this assay with knockout studies of individual remodeler subunits can be used to assess the effects of specific remodelers on the positioning of individual nucleosomes. ChIP experiments in combination with exonuclease digest (ChIP-exo) have been used to precisely map remodeler-DNA contacts at base-pair resolution. Together with molecular modeling, this approach predicted how the ISW2 ATPase is oriented on the nucleosome relative to the nucleosome-free region [344, 345]. Integrating remodeler ChIP-seq with histone modification and chromatin factor ChIP-seq, BS-seq, DNase-seq, as well as genome organization data have enabled prediction of binding preferences and effects of binding *in vivo* [337, 346, 347]. Combining such studies with factor knockout provides more direct insights into remodeler functions [348].

Locus-specific effects of chromatin remodelers can be studied microscopically using a reporter system consisting of LacO and TetO repeats, which is visualized by fluorescently tagged LacR and TetR, on either side of an inducible locus in yeast. The distance between the fluorescent signals is indicative of the chromatin compaction state and loss of either Snf2 or FACT prevented chromatin decompaction upon induction (Figure 11.7E) [349]. In related approaches, mammalian cells have been engineered to carry an inducible transgene array with repetitive LacO and TRE sequences to allow for visualization and transcriptional activation of the array. In this controlled system, chromatin remodeling processes can be studied before or after activation of the array [350]. It is also possible to target chaperones to such repetitive arrays by LacI fusion, and their colocalization with histone variants can be assessed microscopically [351].

Studying the exact mechanism of chromatin remodeling *in vivo* is challenging. Initially, remodeling was studied in yeast strains harboring recombinase sites upstream and downstream of a specific, inducible promoter, so that recombination results in circularization of this element. As outlined previously, the amount of supercoils in such DNA informs about the number of nucleosomes in the promoter region before and after induction [352]. However, remodelers can generate accessible chromatin either by sliding or by disassembly of nucleosomes. An evolved system is able to distinguish these two possibilities. If circularization is induced before activation of the promoter, this results in either retention of all nucleosomes on the plasmid (sliding) or loss of nucleosomes (disassembly). Using such assay in combination with remodeler knockout strains, the responsible remodeling complexes have been identified [353].

11.6 Challenges in Chromatin Interactomics

The study of chromatin biology and of interactions within the chromatin regulatory network has come a long way. Initially, the main focus was on defining different modifications of chromatin and with particular focus on histone proteins. Several types of protein modifications were first described on histones and new ones are still being uncovered. The number of modifications that single sites, individual histones, distinct nucleosomes, and stretches of chromatin putatively can display is staggering. A second phase of chromatin study was devoted to defining the modification enzyme systems, the readout and translation mechanisms of the modifications, and the protein complexes that govern chromatin assembly and remodeling. Over the last two decades, the field has seen major progress in understanding the working mode of these factors and, in particular, their functional domains in isolation and on the molecular level.

While advances have been made in comprehending the interplay of few chromatin modifications, it is not fully clear which patterns of marks really exist. Only for few combination pairs, coexistence has been verified on a molecular level. MS-based approaches seem capable to resolve the issue. However, the hurdle of fragmentation and separation of chromatin components for analysis needs to be overcome. While epigenetic landscapes established by ChIP approaches suggest coexistence of marks and chromatin factors, this is not necessarily the case in molecular terms as the experiments average over large cell numbers (i.e. one cell might have one modification at a certain position of the genome and another cell might have another mark there; then the ChIP composite would list both modifications present in the population at the given region). While single-cell experiments address the ensemble issue, only one chromatin feature can be detected in a single cell at a time. New methodologies need to be developed that allow mapping of the full complexity of the chromatin interactome at a given site of the genome and at a defined cellular state.

As illustrated on the example of UHRF1, a single reader can have multiple domains that recognize distinct chromatin modifications. Similarly, protein complexes contain multiple reader domains that putatively can interact with combinatorially modified chromatin regions. Chromatin remodelers work on nontrivial substrates and chaperones function in complex chromatin assembly. The biochemical analysis of these systems requires efficient and widely accessible methods for making compound substrates (going from peptides to histones and nucleosomes/nucleosomal arrays containing different histone variants, chromatin modifications, and nucleosome status) in high throughput. Further, methodologies need to be developed that enable the simultaneous study of multiple properties, interactions, and functions of complex "designer chromatin" with multiple factors and with high sensitivity.

On a reader level, it is still unclear how multivalent or synergistic interactions with multiple chromatin marks affect protein/chromatin interaction and function. Is simultaneous interaction with all marks required for chromatin targeting? Are specific marks regulating enzymatic activities rather than localization? Are interactions constitutive or regulated and how are these controlled on a molecular level?

Another level of complexity is brought about by the nature and dimension of the complex cellular chromatin environment, in which different chromatin fibers influence each other through sub-compartment formation and phase separation. As these events are still incompletely understood and difficult to manipulate, it is challenging to study the effects of 3D chromatin organization on protein–chromatin interactions.

Cellular chromatin does not merely contain nucleosomes and interacting proteins but also a plethora of RNAs and small molecules. More elaborate *in vitro* systems and *in vivo* approaches are needed to study the combined effects of all aspects of chromatin on the interactome.

Both chromatin-interacting protein complexes and chromatin states are dynamically regulated *in vivo* in developmental, cell type, or cell state-specific contexts. The chromatin interactome has traditionally been characterized in few transformed/cancer cell lines. However, it is becoming increasingly important to study it in different physiological contexts. Reader complexes, chaperones, and chromatin remodelers can exist in a number of variant assemblies depending on the developmental stage or the cell type analyzed. For the most part it is not clear how these are regulated or what the functional differences are. Also, multiple dynamic interactions of modifying enzymes, modification readers, chaperones, and remodeling complexes exist within the chromatin interactome. These are clearly essential for the sequential generation and removal of chromatin marks and all chromatin-mediated processes.

There is clearly plenty of work left to do before the chromatin interaction and regulatory system is comprehended. It is expected that the chromatin biology field will carry on to draw from many other research areas of interactomics, but at the same time will continue to stimulate research in related fields with the richness of first-time discoveries of basic phenomena made in its study.

References

1 Hyman, A.A., Weber, C.A., and Julicher, F. (2014). Liquid–liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* 30: 39–58.

2 Erdel, F. and Rippe, K. (2018). Formation of chromatin subcompartments by phase separation. *Biophys. J.* 114 (10): 2262–2270.

3 Fischle, W., Wang, Y., Jacobs, S.A. et al. (2003). Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev.* 17 (15): 1870–1881.

4 Huang, H., Lin, S., Garcia, B.A., and Zhao, Y. (2015). Quantitative proteomic analysis of histone modifications. *Chem. Rev.* 115 (6): 2376–2418.

5 Rothbart, S.B. and Strahl, B.D. (2014). Interpreting the language of histone and DNA modifications. *Biochim. Biophys. Acta* 1839 (8): 627–643.

6 Müller, M.M. and Muir, T.W. (2015). Histones: at the crossroads of peptide and protein chemistry. *Chem. Rev.* 115 (6): 2296–2349.

7 Phillips, D.M. (1963). The presence of acetyl groups of histones. *Biochem. J.* 87: 258–263.

8 Allfrey, V.G., Faulkner, R., and Mirsky, A.E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc. Natl. Acad. Sci. U.S.A.* 51: 786–794.

9 Murray, K. (1964). The occurrence of ε-*N*-methyl lysine in histones. *Biochemistry* 3: 10–15.

10 Ord, M.G. and Stocken, L.A. (1966). Metabolic properties of histones from rat liver and thymus gland. *Biochem. J.* 98 (3): 888–897.

11 Andrews, F.H., Strahl, B.D., and Kutateladze, T.G. (2016). Insights into newly discovered marks and readers of epigenetic information. *Nat. Chem. Biol.* 12 (9): 662–668.

12 Suganuma, T. and Workman, J.L. (2011). Signals and combinatorial functions of histone modifications. *Annu. Rev. Biochem.* 80: 473–499.

13 Bannister, A.J., Zegerman, P., Partridge, J.F. et al. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* 410 (6824): 120–124.

14 Lachner, M., O'Carroll, D., Rea, S. et al. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* 410 (6824): 116–120.

15 Dhalluin, C., Carlson, J.E., Zeng, L. et al. (1999). Structure and ligand of a histone acetyltransferase bromodomain. *Nature* 399 (6735): 491–496.

16 Owen, D.J., Ornaghi, P., Yang, J.C. et al. (2000). The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase gcn5p. *EMBO J.* 19 (22): 6141–6149.

17 Nielsen, P.R., Nietlispach, D., Mott, H.R. et al. (2002). Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature* 416 (6876): 103–107.

18 Jacobs, S.A. and Khorasanizadeh, S. (2002). Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* 295 (5562): 2080–2083.

19 Huyen, Y., Zgheib, O., Ditullio, R.A. Jr., et al. (2004). Methylated lysine 79 of histone H3 targets 53BP1 to DNA double-strand breaks. *Nature* 432 (7015): 406–411.

20 Huang, Y., Fang, J., Bedford, M.T. et al. (2006). Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A. *Science* 312 (5774): 748–751.

21 Zhao, K., Chai, X., and Marmorstein, R. (2003). Structure of the yeast Hst2 protein deacetylase in ternary complex with 2′-*O*-acetyl ADP ribose and histone peptide. *Structure* 11 (11): 1403–1411.

22 Macdonald, N., Welburn, J.P., Noble, M.E. et al. (2005). Molecular basis for the recognition of phosphorylated and phosphoacetylated histone h3 by 14-3-3. *Mol. Cell* 20 (2): 199–211.

23 Rajakumara, E., Wang, Z., Ma, H. et al. (2011). PHD finger recognition of unmodified histone H3R2 links UHRF1 to regulation of euchromatic gene expression. *Mol. Cell* 43 (2): 275–284.

24 Li, Y., Sabari, B.R., Panchenko, T. et al. (2016). Molecular coupling of histone crotonylation and active transcription by AF9 YEATS domain. *Mol. Cell* 62 (2): 181–193.

25 Strahl, B.D. and Allis, C.D. (2000). The language of covalent histone modifications. *Nature* 403 (6765): 41–45.

26 Jenuwein, T. and Allis, C.D. (2001). Translating the histone code. *Science* 293 (5532): 1074–1080.

27 Jacobs, S.A., Fischle, W., and Khorasanizadeh, S. (2003). Assays for the determination of structure and dynamics of the interaction of the chromodomain with histone peptides. In: *Methods in Enzymology*, vol. 376 (ed. C. Wu and C.D. Allis), 131–148. Academic Press.

28 Shimko, J.C., Howard, C.J., Poirier, M.G., and Ottesen, J.J. (2013). Preparing semisynthetic and fully synthetic histones h3 and h4 to modify the nucleosome core. *Methods Mol. Biol.* 981: 177–192.

29 Li, X., Foley, E.A., Molloy, K.R. et al. (2012). Quantitative chemical proteomics approach to identify post-translational modification-mediated protein–protein interactions. *J. Am. Chem. Soc.* 134 (4): 1982–1985.

30 Yang, T., Liu, Z., and Li, X.D. (2015). Developing diazirine-based chemical probes to identify histone modification 'readers' and 'erasers'. *Chem. Sci.* 6 (2): 1011–1017.

31 Simon, M.D., Chu, F., Racki, L.R. et al. (2007). The site-specific installation of methyl-lysine analogs into recombinant histones. *Cell* 128 (5): 1003–1012.

32 Neumann, H., Hancock, S.M., Buning, R. et al. (2009). A method for genetically installing site-specific acetylation in recombinant histones defines the effects of H3 K56 acetylation. *Mol. Cell* 36 (1): 153–163.

33 Wilkins, B.J., Rall, N.A., Ostwal, Y. et al. (2014). A cascade of histone modifications induces chromatin condensation in mitosis. *Science* 343 (6166): 77–80.

34 Kalb, R., Latwiel, S., Baymaz, H.I. et al. (2014). Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. *Nat. Struct. Mol. Biol.* 21 (6): 569–571.

35 McGinty, R.K., Kim, J., Chatterjee, C. et al. (2008). Chemically ubiquitylated histone H2B stimulates hDot1L-mediated intranucleosomal methylation. *Nature* 453 (7196): 812–816.

36 Ajish Kumar, K.S., Haj-Yahya, M., Olschewski, D. et al. (2009). Highly efficient and chemoselective peptide ubiquitylation. *Angew. Chem. Int. Ed.* 48 (43): 8090–8094.

37 David, Y., Vila-Perelló, M., Verma, S., and Muir, T.W. (2015). Chemical tagging and customizing of cellular chromatin states using ultrafast trans-splicing inteins. *Nat. Chem.* 7 (5): 394–402.

38 Aparicio Pelaz, D., Yerkesh, Z., Kirchgäßner, S. et al. (2020). Examining histone modification crosstalk using immobilized libraries established from ligation-ready nucleosomes. *Chem. Sci.* 11 (34): 9218–9225.

39 Lowary, P.T. and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* 276 (1): 19–42.

40 Dyer, P.N., Edayathumangalam, R.S., White, C.L. et al. (2004). Reconstitution of nucleosome core particles from recombinant histones and DNA. *Methods Enzymol.* 375: 23–44.

41 Malecek, K. and Ruthenburg, A. (2012). Validation of histone-binding partners by peptide pull-downs and isothermal titration calorimetry. *Methods Enzymol.* 512: 187–220.

42 Petell, C.J., Pham, A.T., Skela, J., and Strahl, B.D. (2019). Improved methods for the detection of histone interactions with peptide microarrays. *Sci. Rep.* 9 (1): 6265.

43 Kim, J., Daniel, J., Espejo, A. et al. (2006). Tudor, MBT and chromo domains gauge the degree of lysine methylation. *EMBO Rep.* 7 (4): 397–403.

44 Nguyen, U.T.T., Bittova, L., Müller, M.M. et al. (2014). Accelerated chromatin biochemistry using DNA-barcoded nucleosome libraries. *Nat. Methods* 11 (8): 834–840.

45 Price, J.V., Tangsombatvisit, S., Xu, G. et al. (2012). On silico peptide microarrays for high-resolution mapping of antibody epitopes and diverse protein–protein interactions. *Nat. Med.* 18 (9): 1434–1440.

46 Winkler, D.F., Hilpert, K., Brandt, O., and Hancock, R.E. (2009). Synthesis of peptide arrays using SPOT-technology and the CelluSpots-method. *Methods Mol. Biol.* 570: 157–174.

47 Fuchs, S.M., Krajewski, K., Baker, R.W. et al. (2011). Influence of combinatorial histone modifications on antibody and effector protein recognition. *Curr. Biol.* 21 (1): 53–58.

48 Garske, A.L., Craciun, G., and Denu, J.M. (2008). A combinatorial H4 tail library for exploring the histone code. *Biochemistry* 47 (31): 8094–8102.

49 Dhayalan, A., Rajavelu, A., Rathert, P. et al. (2010). The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *J. Biol. Chem.* 285 (34): 26114–26120.

50 Nady, N., Min, J., Kareta, M.S. et al. (2008). A SPOT on the chromatin landscape? Histone peptide arrays as a tool for epigenetic research. *Trends Biochem. Sci* 33 (7): 305–313.

51 Kungulovski, G., Kycia, I., Tamas, R. et al. (2014). Application of histone modification-specific interaction domains as an alternative to antibodies. *Genome Res.* 24 (11): 1842–1853.

52 Garske, A.L., Oliver, S.S., Wagner, E.K. et al. (2010). Combinatorial profiling of chromatin binding modules reveals multisite discrimination. *Nat. Chem. Biol.* 6 (4): 283–290.

53 Frank, R. (2002). The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports--principles and applications. *J. Immunol. Methods* 267 (1): 13–26.

54 Bock, I., Kudithipudi, S., Tamas, R. et al. (2011). Application of Celluspots peptide arrays for the analysis of the binding specificity of epigenetic reading domains to modified histone tails. *BMC Biochem.* 12 (1): 48.

55 Gerver, R.E., Gómez-Sjöberg, R., Baxter, B.C. et al. (2012). Programmable microfluidic synthesis of spectrally encoded microspheres. *Lab Chip* 12 (22): 4716–4723.

56 Nguyen, H.Q., Baxter, B.C., Brower, K. et al. (2017). Programmable microfluidic synthesis of over one thousand uniquely identifiable spectral codes. *Adv. Opt. Mater.* 5 (3): 1600548. https://doi.org/10.1002/adom.201600548. Epub 2016 Oct 18. PMID: 28936383; PMCID: PMC5604317.

57 Nguyen, H.Q., Roy, J., Harink, B. et al. (2019). Quantitative mapping of protein-peptide affinity landscapes using spectrally encoded beads. *eLife* 8: e40499.

58 Feng, Y., White, A.K., Hein, J.B. et al. (2020). MRBLES 2.0: high-throughput generation of chemically functionalized spectrally and magnetically encoded hydrogel beads using a simple single-layer microfluidic device. *Microsyst. Nanoeng.* 6: 109.

59 Wysocka, J. (2006). Identifying novel proteins recognizing histone modifications using peptide pull-down assay. *Methods* 40 (4): 339–343.

60 Vermeulen, M., Mulder, K.W., Denissov, S. et al. (2007). Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* 131 (1): 58–69.

61 Nikolov, M., Stützer, A., Mosch, K. et al. (2011). Chromatin affinity purification and quantitative mass spectrometry defining the interactome of histone modification patterns. *Mol. Cell. Proteomics* 10 (11): M110.005371.

62 Ong, S.E., Mittler, G., and Mann, M. (2004). Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nat. Methods* 1 (2): 119–126.

63 Makowski, M.M., Grawe, C., Foster, B.M. et al. (2018). Global profiling of protein-DNA and protein-nucleosome binding affinities using quantitative mass spectrometry. *Nat. Commun.* 9 (1): 1653.

64 Nahnsen, S., Bielow, C., Reinert, K., and Kohlbacher, O. (2013). Tools for label-free peptide quantification. *Mol. Cell. Proteomics* 12 (3): 549–556.

65 Cox, J., Hein, M.Y., Luber, C.A. et al. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* 13 (9): 2513–2526.

66 Chen, P., Guo, Z., Chen, C. et al. (2020). Identification of dual histone modification-binding protein interaction by combining mass spectrometry and isothermal titration calorimetric analysis. *J. Adv. Res.* 22: 35–46.

67 Engelen, E., Brandsma, J.H., Moen, M.J. et al. (2015). Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry. *Nat. Commun.* 6 (1): 7155.

68 Bluhm, A., Casas-Vila, N., Scheibe, M., and Butter, F. (2016). Reader interactome of epigenetic histone marks in birds. *Proteomics* 16 (3): 427–436.

69 Vezzoli, A., Bonadies, N., Allen, M.D. et al. (2010). Molecular basis of histone H3K36me3 recognition by the PWWP domain of Brpf1. *Nat. Struct. Mol. Biol.* 17 (5): 617–619.

70 Hoeijmakers, W.A.M., Miao, J., Schmidt, S. et al. (2019). Epigenetic reader complexes of the human malaria parasite, *Plasmodium falciparum*. *Nucleic Acids Res.* 47 (22): 11574–11588.

71 Winter, S., Simboeck, E., Fischle, W. et al. (2008). 14-3-3 proteins recognize a histone code at histone H3 and are required for transcriptional activation. *EMBO J.* 27 (1): 88–99.

72 Vermeulen, M., Eberl, H.C., Matarese, F. et al. (2010). Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* 142 (6): 967–980.

73 Kunowska, N., Rotival, M., Yu, L. et al. (2015). Identification of protein complexes that bind to histone H3 combinatorial modifications using super-SILAC and weighted correlation network analysis. *Nucleic Acids Res.* 43 (3): 1418–1432.

74 Bartke, T., Vermeulen, M., Xhemalce, B. et al. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* 143 (3): 470–484.

75 Ruthenburg, A.J., Li, H., Milne, T.A. et al. (2011). Recognition of a mononucleosomal histone modification pattern by BPTF via multivalent interactions. *Cell* 145 (5): 692–706.

76 Shema-Yaacoby, E., Nikolov, M., Haj-Yahya, M. et al. (2013). Systematic identification of proteins binding to chromatin-embedded ubiquitylated H2B reveals recruitment of SWI/SNF to regulate transcription. *Cell Rep.* 4 (3): 601–608.

77 Skrajna, A., Goldfarb, D., Kedziora, K.M. et al. (2020). Comprehensive nucleosome interactome screen establishes fundamental principles of nucleosome binding. *Nucleic Acids Res.* 48 (17): 9415–9432.

78 Burton, A.J., Haugbro, M., Gates, L.A. et al. (2020). In situ chromatin interactomics using a chemical bait and trap approach. *Nat. Chem.* 12 (6): 520–527.

79 Zhai, G., Dong, H., Guo, Z. et al. (2018). An efficient approach for selective enrichment of histone modification readers using self-assembled multivalent photoaffinity peptide probes. *Anal. Chem.* 90 (19): 11385–11392.

80 Muthurajan, U., Mattiroli, F., Bergeron, S. et al. (2016). In vitro chromatin assembly: strategies and quality control. *Methods Enzymol.* 573: 3–41.

81 Hiragami-Hamada, K., Soeroes, S., Nikolov, M. et al. (2016). Dynamic and flexible H3K9me3 bridging via HP1beta dimerization establishes a plastic state of condensed chromatin. *Nat. Commun.* 7: 11310.

82 Ruthenburg, A.J., Wang, W., Graybosch, D.M. et al. (2006). Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. *Nat. Struct. Mol. Biol.* 13 (8): 704–712.

83 Munari, F., Soeroes, S., Zenn, H.M. et al. (2012). Methylation of lysine 9 in histone H3 directs alternative modes of highly dynamic interaction of heterochromatin protein hHP1β with the nucleosome. *J Biol Chem.* 287 (40): 33756–33765.

84 Villiers, M.-B., Cortès, S., Brakha, C. et al. (2009). Polypyrrole–peptide microarray for biomolecular interaction analysis by SPR imaging. In: *Peptide Microarrays: Methods and Protocols* (ed. M. Cretich and M. Chiari), 317–328. Totowa, NJ: Humana Press.

85 Scarano, S., Scuffi, C., Mascini, M., and Minunni, M. (2010). Surface plasmon resonance imaging (SPRi)-based sensing: a new approach in signal sampling and management. *Biosens. Bioelectron.* 26 (4): 1380–1385.

86 Zhao, S., Yang, M., Zhou, W. et al. (2017). Kinetic and high-throughput profiling of epigenetic interactions by 3D-carbene chip-based surface plasmon resonance imaging technology. *Proc. Natl. Acad. Sci. U.S.A.* 114 (35): E7245–E7254.

87 Chen, Y. and Barkley, M.D. (1998). Toward understanding tryptophan fluorescence in proteins. *Biochemistry* 37 (28): 9976–9982.

88 Gatchalian, J., Wang, X., Ikebe, J. et al. (2017). Accessibility of the histone H3 tail in the nucleosome for binding of paired readers. *Nat. Commun.* 8 (1): 1489.

89 Jacobs, S.A., Taverna, S.D., Zhang, Y. et al. (2001). Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. *EMBO J.* 20 (18): 5232–5241.

90 Liu, F., Chen, X., Allali-Hassani, A. et al. (2009). Discovery of a 2,4-diamino-7-aminoalkoxyquinazoline as a potent and selective inhibitor of histone lysine methyltransferase G9a. *J. Med. Chem.* 52 (24): 7950–7953.

91 Cheow, L.F., Viswanathan, R., Chin, C.S. et al. (2014). Multiplexed analysis of protein-ligand interactions by fluorescence anisotropy in a microfluidic platform. *Anal. Chem.* 86 (19): 9901–9908.

92 Hard, R., Li, N., He, W. et al. (2018). Deciphering and engineering chromodomain-methyllysine peptide recognition. *Sci. Adv.* 4 (11): eaau1447.

93 Seidel, S.A.I., Dijkman, P.M., Lea, W.A. et al. (2013). Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. *Methods* 59 (3): 301–315.

94 Schubert, T., Pusch Miriam, C., Diermeier, S. et al. (2012). Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. *Mol. Cell* 48 (3): 434–444.

95 Willhoft, O., McCormack, E.A., Aramayo, R.J. et al. (2017). Crosstalk within a functional INO80 complex dimer regulates nucleosome sliding. *eLife* 6: e25782. https://doi.org/10.7554/eLife.25782. PMID: 28585918; PMCID: PMC5472440.

96 Schubert, T. and Langst, G. (2015). Studying epigenetic interactions using MicroScale Thermophoresis (MST). *AIMS Biophys.* 2 (3): 370–380.

97 Gelato, K.A., Tauber, M., Ong, M.S. et al. (2014). Accessibility of different histone H3-binding domains of UHRF1 is allosterically regulated by phosphatidylinositol 5-phosphate. *Mol. Cell* 54 (6): 905–919.

98 Harrison, J.S., Cornett, E.M., Goldfarb, D. et al. (2016). Hemi-methylated DNA regulates DNA methylation inheritance through allosteric activation of H3 ubiquitylation by UHRF1. *eLife* 5: e17101.

99 Zhao, D., Guan, H., Zhao, S. et al. (2016). YEATS2 is a selective histone crotonylation reader. *Cell Res.* 26 (5): 629–632.

100 Kato, H., Jiang, J., Zhou, B.R. et al. (2013). A conserved mechanism for centromeric nucleosome recognition by centromere protein CENP-C. *Science* 340 (6136): 1110–1113.

101 Yang, D., Fang, Q., Wang, M. et al. (2013). Nalpha-acetylated Sir3 stabilizes the conformation of a nucleosome-binding loop in the BAH domain. *Nat. Struct. Mol. Biol.* 20 (9): 1116–1118.

102 Schermelleh, L., Ferrand, A., Huser, T. et al. (2019). Super-resolution microscopy demystified. *Nat. Cell Biol.* 21 (1): 72–84.

103 Schmiedeberg, L., Weisshart, K., Diekmann, S. et al. (2004). High- and low-mobility populations of HP1 in heterochromatin of mammalian cells. *Mol. Biol. Cell* 15 (6): 2819–2833.

104 Mazza, D., Stasevich, T.J., Karpova, T.S., and McNally, J.G. (2012). Monitoring dynamic binding of chromatin proteins in vivo by fluorescence correlation spectroscopy and temporal image correlation spectroscopy. *Methods Mol. Biol. (Clifton, NJ)* 833: 177–200.

105 Deindl, S., Hwang, W.L., Hota, S.K. et al. (2013). ISWI remodelers slide nucleosomes with coordinated multi-base-pair entry steps and single-base-pair exit steps. *Cell* 152 (3): 442–452.

106 Zane, L., Chapus, F., Pegoraro, G., and Misteli, T. (2017). HiHiMap: single-cell quantitation of histones and histone posttranslational modifications across the cell cycle by high-throughput imaging. *Mol. Biol. Cell* 28 (17): 2290–2302.

107 Hayashi-Takanaka, Y., Kina, Y., Nakamura, F. et al. (2020). Histone modification dynamics as revealed by multicolor immunofluorescence-based single-cell analysis. *J. Cell Sci.* 133 (14): jcs243444.

108 Sardo, L., Lin, A., Khakhina, S. et al. (2017). Real-time visualization of chromatin modification in isolated nuclei. *J. Cell Sci.* 130 (17): 2926–2940.

109 Petruk, S., Sedkov, Y., Johnston, D.M. et al. (2012). TrxG and PcG proteins but not methylated histones remain associated with DNA through replication. *Cell* 150 (5): 922–933.

110 Lazarchuk, P., Roy, S., Schlacher, K., and Sidorova, J. (2019). Detection and quantitation of acetylated histones on replicating DNA using in situ proximity ligation assay and click-it chemistry. *Methods Mol. Biol.* 1983: 29–45.

111 Serebryannyy, L.A. and Misteli, T. (2019). HiPLA: high-throughput imaging proximity ligation assay. *Methods* 157: 80–87.

112 Ren, B., Robert, F., Wyrick, J.J. et al. (2000). Genome-wide location and function of DNA binding proteins. *Science* 290 (5500): 2306–2309.

113 Iyer, V.R., Horak, C.E., Scafe, C.S. et al. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409 (6819): 533–538.

114 Braunstein, M., Rose, A.B., Holmes, S.G. et al. (1993). Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. *Genes Dev.* 7 (4): 592–604.

115 Orlando, V. (2000). Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem. Sci* 25 (3): 99–104.

116 Venkataraman, A., Yang, K., Irizarry, J. et al. (2018). A toolbox of immunoprecipitation-grade monoclonal antibodies to human transcription factors. *Nat. Methods* 15 (5): 330–338.

117 Landt, S.G., Marinov, G.K., Kundaje, A. et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22 (9): 1813–1831.

118 Sanchez, P., Daniels, K.J., Park, Y.N., and Soll, D.R. (2014). Generating a battery of monoclonal antibodies against native green fluorescent protein for immunostaining, FACS, IP, and ChIP using a unique adjuvant. *Monoclonal Antibodies Immunodiagn. Immunother.* 33 (2): 80–88.

119 Lambert, J.P., Fillingham, J., Siahbazi, M. et al. (2010). Defining the budding yeast chromatin-associated interactome. *Mol. Syst. Biol.* 6: 448.

120 Poser, I., Sarov, M., Hutchins, J.R. et al. (2008). BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* 5 (5): 409–415.

121 Wang, C.I., Alekseyenko, A.A., LeRoy, G. et al. (2013). Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in *Drosophila*. *Nat. Struct. Mol. Biol.* 20 (2): 202–209.

122 Savic, D., Partridge, E.C., Newberry, K.M. et al. (2015). CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. *Genome Res.* 25 (10): 1581–1589.

123 Bibikova, M., Beumer, K., Trautman, J.K., and Carroll, D. (2003). Enhancing gene targeting with designed zinc finger nucleases. *Science* 300 (5620): 764.

124 Doudna, J.A. and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346 (6213): 1258096.

125 Wright, D.A., Li, T., Yang, B., and Spalding, M.H. (2014). TALEN-mediated genome editing: prospects and perspectives. *Biochem. J* 462 (1): 15–24.

126 Bukhari, H. and Müller, T. (2019). Endogenous fluorescence tagging by CRISPR. *Trends Cell Biol.* 29 (11): 912–928.

127 Jinek, M., Chylinski, K., Fonfara, I. et al. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337 (6096): 816–821.

128 Cong, L., Ran, F.A., Cox, D. et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* 339 (6121): 819–823.

129 Mali, P., Yang, L., Esvelt, K.M. et al. (2013). RNA-guided human genome engineering via Cas9. *Science* 339 (6121): 823–826.

130 Kungulovski, G. and Jeltsch, A. (2016). Epigenome editing: state of the art, concepts, and perspectives. *Trends Genet.* 32 (2): 101–113.

131 O'Neill, L.P. and Turner, B.M. (1995). Histone H4 acetylation distinguishes coding regions of the human genome from heterochromatin in a differentiation-dependent but transcription-independent manner. *EMBO J.* 14 (16): 3946–3957.

132 O'Neill, L.P. and Turner, B.M. (2003). Immunoprecipitation of native chromatin: NChIP. *Methods* 31 (1): 76–82.

133 Skene, P.J. and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* 6: e21856. https://doi.org/10.7554/eLife.21856. PMID: 28079019; PMCID: PMC5310842.

134 Kaya-Okur, H.S., Wu, S.J., Codomo, C.A. et al. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* 10 (1): 1930.

135 Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316 (5830): 1497–1502.

136 Egelhofer, T.A., Minoda, A., Klugman, S. et al. (2011). An assessment of histone-modification antibody quality. *Nat. Struct. Mol. Biol.* 18 (1): 91–93.

137 Consortium EP, Moore, J.E., Purcaro, M.J., Pratt, H.E. et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583 (7818): 699–710.

138 Mei, S., Qin, Q., Wu, Q. et al. (2017). Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.* 45 (D1): D658–D662.

139 Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J. et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518 (7539): 317–330.

140 Desvoyes, B., Sequeira-Mendes, J., Vergara, Z. et al. (2018). Sequential ChIP protocol for profiling bivalent epigenetic modifications (ReChIP). *Methods Mol. Biol. (Clifton, NJ)* 1675: 83–97.

141 Soldi, M. and Bonaldi, T. (2014). The ChroP approach combines ChIP and mass spectrometry to dissect locus-specific proteomic landscapes of chromatin. *J. Visualized Exp.* 11 (86): 51220. https://doi.org/10.3791/51220. PMID: 24747196; PMCID: PMC4166860.

142 Papachristou, E.K., Kishore, K., Holding, A.N. et al. (2018). A quantitative mass spectrometry-based approach to monitor the dynamics of endogenous chromatin-associated protein complexes. *Nat. Commun.* 9 (1): 2311.

143 Byrum, S.D., Raman, A., Taverna, S.D., and Tackett, A.J. (2012). ChAP-MS: a method for identification of proteins and histone posttranslational modifications at a single genomic locus. *Cell Rep.* 2 (1): 198–205.

144 Déjardin, J. and Kingston, R.E. (2009). Purification of proteins associated with specific genomic loci. *Cell* 136 (1): 175–186.

145 Waldrip, Z.J., Byrum, S.D., Storey, A.J. et al. (2014). A CRISPR-based approach for proteomic analysis of a single genomic locus. *Epigenetics* 9 (9): 1207–1211.

146 Liu, X., Zhang, Y., Chen, Y. et al. (2017). In situ capture of chromatin interactions by biotinylated dCas9. *Cell* 170 (5): 1028–1043 e19.

147 Fujita, T., Yuno, M., and Fujii, H. (2018). An enChIP system for the analysis of bacterial genome functions. *BMC Res. Notes* 11 (1): 387.

148 Korthout, T., Poramba-Liyanage, D.W., van Kruijsbergen, I. et al. (2018). Decoding the chromatin proteome of a single genomic locus by DNA sequencing. *PLoS Biol.* 16 (7): e2005542.

149 Yang, T., Li, X.M., Bao, X. et al. (2016). Photo-lysine captures proteins that bind lysine post-translational modifications. *Nat. Chem. Biol.* 12 (2): 70–72.

150 Kleiner, R.E., Hang, L.E., Molloy, K.R. et al. (2018). A chemical proteomics approach to reveal direct protein–protein interactions in living cells. *Cell Chem. Biol.* 25 (1): 110–120 e3.

151 Villaseñor, R., Pfaendler, R., Ambrosi, C. et al. (2020). ChromID identifies the protein interactome at chromatin marks. *Nat. Biotechnol.* 38 (6): 728–736.

152 Ummethum, H. and Hamperl, S. (2020). Proximity labeling techniques to study chromatin. *Front. Genet.* 11: 450.

153 Gao, X.D., Tu, L.C., Mir, A. et al. (2018). C-BERST: defining subnuclear proteomic landscapes at genomic elements with dCas9-APEX2. *Nat. Methods* 15 (6): 433–436.

154 Myers, S.A., Wright, J., Peckner, R. et al. (2018). Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling. *Nat. Methods* 15 (6): 437–439.

155 Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* 321 (6067): 209–213.

156 Illingworth, R.S. and Bird, A.P. (2009). CpG islands – 'a rough guide'. *FEBS Lett.* 583 (11): 1713–1720.

157 de Mendoza, A., Poppe, D., Buckberry, S. et al. (2021). The emergence of the brain non-CpG methylation system in vertebrates. *Nat. Ecol. Evol.* 5 (3): 369–378.

158 Patil, V., Ward, R.L., and Hesson, L.B. (2014). The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics* 9 (6): 823–828.

159 Zhang, H., Lang, Z., and Zhu, J.K. (2018). Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* 19 (8): 489–506.

160 Johnson, T.B. and Coghill, R.D. (1925). Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus1. *J. Am. Chem. Soc.* 47 (11): 2838–2844.

161 Hotchkiss, R.D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J. Biol. Chem.* 175 (1): 315–332.

162 Greenberg, M.V.C. and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* 20 (10): 590–607.

163 Lee, J.Y. and Lee, T.H. (2012). Effects of histone acetylation and CpG methylation on the structure of nucleosomes. *Biochim. Biophys. Acta* 1824 (8): 974–982.

164 Sriraman, A., Debnath, T.K., Xhemalce, B., and Miller, K.M. (2020). Making it or breaking it: DNA methylation and genome integrity. *Essays Biochem.* 64 (5): 687–703.

165 Antequera, F., Macleod, D., and Bird, A.P. (1989). Specific protection of methylated CpGs in mammalian nuclei. *Cell* 58 (3): 509–517.

166 Meehan, R.R., Lewis, J.D., McKay, S. et al. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* 58 (3): 499–507.

167 Ng, H.H., Zhang, Y., Hendrich, B. et al. (1999). MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat. Genet.* 23 (1): 58–61.

168 Feng, Q. and Zhang, Y. (2001). The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. *Genes Dev.* 15 (7): 827–832.

169 Feng, Q., Cao, R., Xia, L. et al. (2002). Identification and functional characterization of the p66/p68 components of the MeCP1 complex. *Mol. Cell. Biol.* 22 (2): 536–546.

170 Brackertz, M., Boeke, J., Zhang, R., and Renkawitz, R. (2002). Two highly related p66 proteins comprise a new family of potent transcriptional repressors interacting with MBD2 and MBD3. *J. Biol. Chem.* 277 (43): 40958–40966.

171 Lewis, J.D., Meehan, R.R., Henzel, W.J. et al. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* 69 (6): 905–914.

172 Cross, S.H., Meehan, R.R., Nan, X., and Bird, A. (1997). A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. *Nat. Genet.* 16 (3): 256–259.

173 Hendrich, B. and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* 18 (11): 6538–6547.

174 Arita, K., Ariyoshi, M., Tochio, H. et al. (2008). Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* 455 (7214): 818–821.

175 Iurlaro, M., Ficz, G., Oxley, D. et al. (2013). A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* 14 (10): R119.

176 Karemaker, I.D. and Vermeulen, M. (2018). ZBTB2 reads unmethylated CpG island promoters and regulates embryonic stem cell differentiation. *EMBO Rep.* 19 (4): e44993. https://doi.org/10.15252/embr.201744993. Epub 2018 Feb 1. PMID: 29437775; PMCID: PMC5891434.

177 Mann, I.K., Chatterjee, R., Zhao, J. et al. (2013). CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. *Genome Res.* 23 (6): 988–997.

178 Davis, F.F. and Allen, F.W. (1957). Ribonucleic acids from yeast which contain a fifth nucleotide. *J. Biol. Chem.* 227 (2): 907–915.

179 Yu, C.T. and Allen, F.W. (1959). Studies on an isomer of uridine isolated from ribonucleic acids. *Biochim. Biophys. Acta* 32: 393–406.

180 Cohn, W.E. (1959). 5-Ribosyl uracil, a carbon–carbon ribofuranosyl nucleoside in ribonucleic acids. *Biochim. Biophys. Acta* 32: 569–571.

181 Mathlin, J., Le Pera, L., and Colombo, T. (2020). A census and categorization method of epitranscriptomic marks. *Int. J. Mol. Sci.* 21 (13): 4684. https://doi.org/10.3390/ijms21134684. PMID: 32630140; PMCID: PMC7370119.

182 Boccaletto, P., Machnicka, M.A., Purta, E. et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46 (D1): D303–D307.

183 Wei, C.M., Gershowitz, A., and Moss, B. (1975). Methylated nucleotides block 5′ terminus of HeLa cell messenger RNA. *Cell* 4 (4): 379–386.

184 Liu, J., Yue, Y., Han, D. et al. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.* 10 (2): 93–95.

185 Fu, Y., Jia, G., Pang, X. et al. (2013). FTO-mediated formation of N6-hydroxymethyladenosine and N6-formyladenosine in mammalian RNA. *Nat. Commun.* 4: 1798.

186 Zhao, B.S., Roundtree, I.A., and He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell Biol.* 18 (1): 31–42.

187 Bartels, S.J., Spruijt, C.G., Brinkman, A.B. et al. (2011). A SILAC-based screen for methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein. *PLoS One* 6 (10): e25884.

188 Patino-Parrado, I., Gomez-Jimenez, A., Lopez-Sanchez, N., and Frade, J.M. (2017). Strand-specific CpG hemimethylation, a novel epigenetic modification functional for genomic imprinting. *Nucleic Acids Res.* 45 (15): 8822–8834.

189 Schrader, A., Gross, T., Thalhammer, V., and Langst, G. (2015). Characterization of Dnmt1 binding and DNA methylation on nucleosomes and nucleosomal arrays. *PLoS One* 10 (10): e0140076.

190 Mittler, G., Butter, F., and Mann, M. (2009). A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res.* 19 (2): 284–293.

191 Spruijt, C.G., Gnerlich, F., Smits, A.H. et al. (2013). Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* 152 (5): 1146–1159.

192 Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S. et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485 (7397): 201–206.

193 Edupuganti, R.R., Geiger, S., Lindeboom, R.G.H. et al. (2017). N(6)-methyladenosine (m(6)A) recruits and repels proteins to regulate mRNA homeostasis. *Nat. Struct. Mol. Biol.* 24 (10): 870–878.

194 Huang, H., Weng, H., Sun, W. et al. (2018). Recognition of RNA N(6)-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nat. Cell Biol.* 20 (3): 285–295.

195 Bogdanovic, O., Smits, A.H., de la Calle Mustienes, E. et al. (2016). Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat. Genet.* 48 (4): 417–426.

196 Sequeira, V.M. and Vermeulen, M. (2019). Identifying readers for (hydroxy)methylated DNA using quantitative interaction proteomics: advances and challenges ahead. *J. Mol. Biol.* S0022–2836 (19): 30714–30714. https://doi.org/10.1016/j.jmb.2019.12.014. Epub ahead of print. PMID: 31866296.

197 Hubner, N.C., Nguyen, L.N., Hornig, N.C., and Stunnenberg, H.G. (2015). A quantitative proteomics tool to identify DNA–protein interactions in primary cells or blood. *J. Proteome Res.* 14 (2): 1315–1329.

198 Yin, Y., Morgunova, E., Jolma, A. et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* 356 (6337): eaaj2239.

199 Arguello, A.E., Leach, R.W., and Kleiner, R.E. (2019). In vitro selection with a site-specifically modified RNA library reveals the binding preferences of N(6)-methyladenosine reader proteins. *Biochemistry* 58 (31): 3386–3395.

200 Hu, S., Wan, J., Su, Y. et al. (2013). DNA methylation presents distinct binding sites for human transcription factors. *eLife* 2: e00726.

201 Kribelbauer, J.F., Laptenko, O., Chen, S. et al. (2017). Quantitative analysis of the DNA methylation sensitivity of transcription factor complexes. *Cell Rep.* 19 (11): 2383–2395.

202 Zuo, Z., Roy, B., Chang, Y.K. et al. (2017). Measuring quantitative effects of methylation on transcription factor-DNA binding affinity. *Sci. Adv.* 3 (11): eaao1799.

203 Karemaker, I.D. and Vermeulen, M. (2018). Single-cell DNA methylation profiling: technologies and biological applications. *Trends Biotechnol.* 36 (9): 952–965.

204 Alarcón, C.R., Goodarzi, H., Lee, H. et al. (2015). HNRNPA2B1 is a mediator of m(6)A-dependent nuclear RNA processing events. *Cell* 162 (6): 1299–1308.

205 Zhu, H., Wang, G., and Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* 17 (9): 551–565.

206 Lövkvist, C., Sneppen, K., and Haerter, J.O. (2017). Exploring the link between nucleosome occupancy and DNA methylation. *Front. Genet.* 8: 232.

207 Feng, L. and Lou, J. (2019). DNA methylation analysis. *Methods Mol. Biol. (Clifton, NJ)* 1894: 181–227.

208 Lee, F.C.Y. and Ule, J. (2018). Advances in CLIP technologies for studies of protein–RNA interactions. *Mol. Cell* 69 (3): 354–369.

209 Lin, C. and Miles, W.O. (2019). Beyond CLIP: advances and opportunities to measure RBP–RNA and RNA–RNA interactions. *Nucleic Acids Res.* 47 (11): 5490–5501.

210 Ji, Z., Song, R., Huang, H. et al. (2016). Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat. Biotechnol.* 34 (4): 410–413.

211 Silverman, I.M., Li, F., Alexander, A. et al. (2014). RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biol.* 15 (1): R3.

212 Yoon, H.G., Chan, D.W., Reynolds, A.B. et al. (2003). N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. *Mol. Cell* 12 (3): 723–734.

213 Maurano, M.T., Wang, H., John, S. et al. (2015). Role of DNA methylation in modulating transcription factor occupancy. *Cell Rep.* 12 (7): 1184–1195.

214 Yang, F., Jin, H., Que, B. et al. (2019). Dynamic m(6)A mRNA methylation reveals the role of METTL3-m(6)A-CDCP1 signaling axis in chemical carcinogenesis. *Oncogene* 38 (24): 4755–4772.

215 Wu, Z., Shi, Y., Lu, M. et al. (2020). METTL3 counteracts premature aging via m6A-dependent stabilization of MIS12 mRNA. *Nucleic Acids Res.* 48 (19): 11083–11096.

216 Liu, N., Dai, Q., Zheng, G. et al. (2015). N(6)-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* 518 (7540): 560–564.

217 Domcke, S., Bardet, A.F., Adrian Ginno, P. et al. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* 528 (7583): 575–579.

218 Dekkers, K.F., Neele, A.E., Jukema, J.W. et al. (2019). Human monocyte-to-macrophage differentiation involves highly localized gain and loss of DNA methylation at transcription factor binding sites. *Epigenet. Chromatin* 12 (1): 34.

219 Tsankov, A.M., Gu, H., Akopian, V. et al. (2015). Transcription factor binding dynamics during human ES cell differentiation. *Nature* 518 (7539): 344–349.

220 Feng, S.Y., Ota, K., and Ito, T. (2010). A yeast one-hybrid system to screen for methylated DNA-binding proteins. *Nucleic Acids Res.* 38 (20): e189.

221 Kustatscher, G., Wills, K.L., Furlan, C., and Rappsilber, J. (2014). Chromatin enrichment for proteomics. *Nat. Protoc.* 9 (9): 2090–2099.

222 Brinkman, A.B., Gu, H., Bartels, S.J. et al. (2012). Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* 22 (6): 1128–1138.

223 Gao, F., Ji, G., Gao, Z. et al. (2014). Direct ChIP-bisulfite sequencing reveals a role of H3K27me3 mediating aberrant hypermethylation of promoter CpG islands in cancer cells. *Genomics* 103 (2–3): 204–210.

224 D'Anna, F., Van Dyck, L., Xiong, J. et al. (2020). DNA methylation repels binding of hypoxia-inducible transcription factors to maintain tumor immunotolerance. *Genome Biol.* 21 (1): 182.

225 Statham, A.L., Robinson, M.D., Song, J.Z. et al. (2012). Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res.* 22 (6): 1120–1127.

226 Lopes, E.C., Valls, E., Figueroa, M.E. et al. (2008). Kaiso contributes to DNA methylation-dependent silencing of tumor suppressor genes in colon cancer cell lines. *Cancer Res.* 68 (18): 7258–7263.

227 Qin, S., Zhang, B., Tian, W. et al. (2015). Kaiso mainly locates in the nucleus in vivo and binds to methylated, but not hydroxymethylated DNA. *Chin. J. Cancer Res.* 27 (2): 148–155.

228 Perez-Perri, J.I., Rogell, B., Schwarzl, T. et al. (2018). Discovery of RNA-binding proteins and characterization of their dynamic responses by enhanced RNA interactome capture. *Nat. Commun.* 9 (1): 4408.

229 Sun, L., Fazal, F.M., Li, P. et al. (2019). RNA structure maps across mammalian cellular compartments. *Nat. Struct. Mol. Biol.* 26 (4): 322–330.

230 Winter, S. and Fischle, W. (2010). Epigenetic markers and their cross-talk. *Essays Biochem.* 48 (1): 45–61.

231 Fischle, W., Wang, Y., and Allis, C.D. (2003). Histone and chromatin cross-talk. *Curr. Opin. Cell Biol.* 15 (2): 172–183.

232 Du, J., Johnson, L.M., Jacobsen, S.E., and Patel, D.J. (2015). DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* 16 (9): 519–532.

233 Citterio, E., Papait, R., Nicassio, F. et al. (2004). Np95 is a histone-binding protein endowed with ubiquitin ligase activity. *Mol. Cell. Biol.* 24 (6): 2526–2535.

234 Unoki, M., Nishidate, T., and Nakamura, Y. (2004). ICBP90, an E2F-1 target, recruits HDAC1 and binds to methyl-CpG through its SRA domain. *Oncogene* 23 (46): 7601–7610.

235 Bostick, M., Kim, J.K., Estève, P.-O. et al. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science (New York, NY)* 317 (5845): 1760–1764.

236 Sharif, J., Muto, M., Takebayashi, S. et al. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* 450 (7171): 908–912.

237 Hashimoto, H., Horton, J.R., Zhang, X. et al. (2008). The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature* 455 (7214): 826–829.

238 Avvakumov, G.V., Walker, J.R., Xue, S. et al. (2008). Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature* 455 (7214): 822–825.

239 Qian, C., Li, S., Jakoncic, J. et al. (2008). Structure and hemimethylated CpG binding of the SRA domain from human UHRF1. *J. Biol. Chem.* 283 (50): 34490–34494.

240 Frauer, C., Hoffmann, T., Bultmann, S. et al. (2011). Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain. *PLoS One* 6 (6): e21306.

241 Schneider, M., Trummer, C., Stengl, A. et al. (2020). Systematic analysis of the binding behaviour of UHRF1 towards different methyl- and carboxylcytosine modification patterns at CpG dyads. *PLoS One* 15 (2): e0229144.

242 Bianchi, C. and Zangi, R. (2013). UHRF1 discriminates against binding to fully-methylated CpG-Sites by steric repulsion. *Biophys. Chem.* 171: 38–45.

243 Karagianni, P., Amazit, L., Qin, J., and Wong, J. (2008). ICBP90, a novel methyl K9 H3 binding protein linking protein ubiquitination with heterochromatin formation. *Mol. Cell. Biol.* 28 (2): 705–717.

244 Rottach, A., Frauer, C., Pichler, G. et al. (2010). The multi-domain protein Np95 connects DNA methylation and histone modification. *Nucleic Acids Res.* 38 (6): 1796–1804.

245 Nady, N., Lemak, A., Walker, J.R. et al. (2011). Recognition of multivalent histone states associated with heterochromatin by UHRF1 protein. *J. Biol. Chem.* 286 (27): 24300–24311.

246 Lallous, N., Legrand, P., McEwen, A.G. et al. (2011). The PHD finger of human UHRF1 reveals a new subgroup of unmethylated histone H3 tail readers. *PLoS One* 6 (11): e27599.

247 Hu, L., Li, Z., Wang, P. et al. (2011). Crystal structure of PHD domain of UHRF1 and insights into recognition of unmodified histone H3 arginine residue 2. *Cell Res.* 21: 1374–1378.

248 Wang, C., Shen, J., Yang, Z. et al. (2011). Structural basis for site-specific reading of unmodified R2 of histone H3 tail by UHRF1 PHD finger. *Cell Res.* 21 (9): 1379–1382.

249 Xie, S., Jakoncic, J., and Qian, C. (2012). UHRF1 double tudor domain and the adjacent PHD finger act together to recognize K9me3-containing histone H3 tail. *J. Mol. Biol.* 415 (2): 318–328.

250 Rothbart, S.B., Dickson, B.M., Ong, M.S. et al. (2013). Multivalent histone engagement by the linked tandem Tudor and PHD domains of UHRF1 is required for the epigenetic inheritance of DNA methylation. *Genes Dev.* 27 (11): 1288–1298.

251 Fang, J., Cheng, J., Wang, J. et al. (2016). Hemi-methylated DNA opens a closed conformation of UHRF1 to facilitate its histone recognition. *Nat. Commun.* 7: 11197.

252 Houliston, R.S., Lemak, A., Iqbal, A. et al. (2017). Conformational dynamics of the TTD-PHD histone reader module of the UHRF1 epigenetic regulator reveals multiple histone-binding states, allosteric regulation, and druggability. *J. Biol. Chem.* 292 (51): 20947–20959.

253 Gao, L., Tan, X.-F., Zhang, S. et al. (2018). An intramolecular interaction of UHRF1 reveals dual control for its histone association. *Structure (London, England: 1993)* 26 (2): 304–311.e3.

254 Kori, S., Jimenji, T., Ekimoto, T. et al. (2020). Serine 298 phosphorylation in linker 2 of UHRF1 regulates ligand-binding property of its tandem tudor domain. *J. Mol. Biol.* 432 (14): 4061–4075. https://doi.org/10.1016/j.jmb.2020.05.006. Epub 2020 May 16. PMID: 32428527.

255 Tauber, M., Kreuz, S., Lemak, A. et al. (2020). Alternative splicing and allosteric regulation modulate the chromatin binding of UHRF1. *Nucleic Acids Res.* 48 (14): 7728–7747.

256 Achour, M., Jacq, X., Ronde, P. et al. (2008). The interaction of the SRA domain of ICBP90 with a novel domain of DNMT1 is involved in the regulation of VEGF gene expression. *Oncogene* 27 (15): 2187–2197.

257 Ferry, L., Fournier, A., Tsusaka, T. et al. (2017). Methylation of DNA ligase 1 by G9a/GLP recruits UHRF1 to replicating DNA and regulates DNA methylation. *Mol. Cell* 67 (4): 550–565.e5.

258 Papait, R., Pistore, C., Grazini, U. et al. (2008). The PHD domain of Np95 (mUHRF1) is involved in large-scale reorganization of pericentromeric heterochromatin. *Mol. Biol. Cell* 19 (8): 3554–3563.

259 Rothbart, S.B., Krajewski, K., Nady, N. et al. (2012). Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat. Struct. Mol. Biol.* 19 (11): 1155–1160.

260 Liu, X., Gao, Q., Li, P. et al. (2013). UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. *Nat. Commun.* 4: 1563. https://doi.org/10.1038/ncomms2562. PMID: 23463006.

261 Cheng, J., Yang, Y., Fang, J. et al. (2013). Structural insight into coordinated recognition of trimethylated histone H3 lysine 9 (H3K9me3) by the plant homeodomain (PHD) and tandem tudor domain (TTD) of UHRF1 (ubiquitin-like, containing PHD and RING finger domains, 1) protein. *J. Biol. Chem.* 288 (2): 1329–1339.

262 Nishiyama, A., Yamaguchi, L., Sharif, J. et al. (2013). Uhrf1-dependent H3K23 ubiquitylation couples maintenance DNA methylation and replication. *Nature* 502 (7470): 249–253.

263 Zhao, Q., Zhang, J., Chen, R. et al. (2016). Dissecting the precise role of H3K9 methylation in crosstalk with DNA maintenance methylation in mammals. *Nat. Commun.* 7: 12464. https://doi.org/10.1038/ncomms12464. PMID: 27554592; PMCID: PMC5426519.

264 Vaughan, R.M., Dickson, B.M., Whelihan, M.F. et al. (2018). Chromatin structure and its chemical modifications regulate the ubiquitin ligase substrate selectivity of UHRF1. *Proc. Natl. Acad. Sci. U.S.A.* 115 (35): 8775–8780.

265 Vaughan, R.M., Kupai, A., Foley, C.A. et al. (2020). The histone and non-histone methyllysine reader activities of the UHRF1 tandem Tudor domain are dispensable for the propagation of aberrant DNA methylation patterning in cancer cells. *Epigenet. Chromatin* 13 (1): 44.

266 Kong, X., Chen, J., Xie, W. et al. (2019). Defining UHRF1 domains that support maintenance of human colon cancer DNA methylation and oncogenic properties. *Cancer Cell* 35 (4): 633–648.e7.

267 Ishiyama, S., Nishiyama, A., Saeki, Y. et al. (2017). Structure of the Dnmt1 reader module complexed with a unique two-mono-ubiquitin mark on histone H3 reveals the basis for DNA methylation maintenance. *Mol. Cell* 68 (2): 350–360.e7.

268 Gurard-Levin, Z.A., Quivy, J.P., and Almouzni, G. (2014). Histone chaperones: assisting histone traffic and nucleosome dynamics. *Annu. Rev. Biochem.* 83: 487–517.

269 Clapier, C.R., Iwasa, J., Cairns, B.R., and Peterson, C.L. (2017). Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nat. Rev. Mol. Cell Biol.* 18 (7): 407–422.

270 Yan, L. and Chen, Z. (2020). A unifying mechanism of DNA translocation underlying chromatin remodeling. *Trends Biochem. Sci* 45 (3): 217–227.

271 Germond, J.E., Hirt, B., Oudet, P. et al. (1975). Folding of the DNA double helix in chromatin-like structures from simian virus 40. *Proc. Natl. Acad. Sci. U.S.A.* 72 (5): 1843–1847.

272 Cebrián, J., Kadomatsu-Hermosa, M.J., Castán, A. et al. (2015). Electrophoretic mobility of supercoiled, catenated and knotted DNA molecules. *Nucleic Acids Res.* 43 (4): e24.

273 Tatchell, K. and Van Holde, K.E. (1977). Reconstitution of chromatin core particles. *Biochemistry* 16 (24): 5295–5303.

274 Stein, A., Whitlock, J.P. Jr., and Bina, M. (1979). Acidic polypeptides can assemble both histones and chromatin in vitro at physiological ionic strength. *Proc. Natl. Acad. Sci. U.S.A.* 76 (10): 5000–5004.

275 Sobolewski, C.H., Klump, H.H., and Lindsey, G.G. (1993). A novel nucleosome assembly procedure (with a little help from pectin). *FEBS Lett.* 318 (1): 27–29.

276 Laskey, R.A., Honda, B.M., Mills, A.D., and Finch, J.T. (1978). Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. *Nature* 275 (5679): 416–420.

277 Laskey, R.A., Honda, B.M., Mills, A.D. et al. (1978). Chromatin assembly and transcription in eggs and oocytes of Xenopus laevis. *Cold Spring Harbor Symp. Quant. Biol.* 42 (Pt 1): 171–178.

278 Laskey, R.A., Mills, A.D., and Morris, N.R. (1977). Assembly of SV40 chromatin in a cell-free system from *Xenopus* eggs. *Cell* 10 (2): 237–243.

279 Earnshaw, W.C., Honda, B.M., Laskey, R.A., and Thomas, J.O. (1980). Assembly of nucleosomes: the reaction involving *X. laevis* nucleoplasmin. *Cell* 21 (2): 373–383.

280 Dilworth, S.M., Black, S.J., and Laskey, R.A. (1987). Two complexes that contain histones are required for nucleosome assembly in vitro: role of nucleoplasmin and N1 in *Xenopus* egg extracts. *Cell* 51 (6): 1009–1018.

281 Kleinschmidt, J.A., Fortkamp, E., Krohne, G. et al. (1985). Co-existence of two different types of soluble histone complexes in nuclei of *Xenopus laevis* oocytes. *J. Biol. Chem.* 260 (2): 1166–1176.

282 Kleinschmidt, J.A. and Franke, W.W. (1982). Soluble acidic complexes containing histones H3 and H4 in nuclei of *Xenopus laevis* oocytes. *Cell* 29 (3): 799–809.

283 Kleinschmidt, J.A., Seiter, A., and Zentgraf, H. (1990). Nucleosome assembly in vitro: separate histone transfer and synergistic interaction of native histone complexes purified from nuclei of *Xenopus laevis* oocytes. *EMBO J.* 9 (4): 1309–1318.

284 Ishimi, Y. and Kikuchi, A. (1991). Identification and molecular cloning of yeast homolog of nucleosome assembly protein I which facilitates nucleosome assembly in vitro. *J. Biol. Chem.* 266 (11): 7025–7029.

285 Ishimi, Y., Hirosumi, J., Sato, W. et al. (1984). Purification and initial characterization of a protein which facilitates assembly of nucleosome-like structure from mammalian cells. *Eur. J. Biochem.* 142 (3): 431–439.

286 Ishimi, Y., Yasuda, H., Hirosumi, J. et al. (1983). A protein which facilitates assembly of nucleosome-like structures in vitro in mammalian cells. *J. Biochem.* 94 (3): 735–744.

287 Stillman, B. (1986). Chromatin assembly during SV40 DNA replication in vitro. *Cell* 45 (4): 555–565.

288 Smith, S. and Stillman, B. (1989). Purification and characterization of CAF-I, a human cell factor required for chromatin assembly during DNA replication in vitro. *Cell* 58 (1): 15–25.

289 Stern, M., Jensen, R., and Herskowitz, I. (1984). Five SWI genes are required for expression of the HO gene in yeast. *J. Mol. Biol.* 178 (4): 853–868.

290 Kwon, H., Imbalzano, A.N., Khavari, P.A. et al. (1994). Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. *Nature* 370 (6489): 477–481.

291 Laurent, B.C., Yang, X., and Carlson, M. (1992). An essential *Saccharomyces cerevisiae* gene homologous to SNF2 encodes a helicase-related protein in a new family. *Mol. Cell. Biol.* 12 (4): 1893–1902.

292 Okabe, I., Bailey, L.C., Attree, O. et al. (1992). Cloning of human and bovine homologs of SNF2/SWI2: a global activator of transcription in yeast *S. cerevisiae*. *Nucleic Acids Res.* 20 (17): 4649–4655.

293 Elfring, L.K., Deuring, R., McCallum, C.M. et al. (1994). Identification and characterization of *Drosophila* relatives of the yeast transcriptional activator SNF2/SWI2. *Mol. Cell. Biol.* 14 (4): 2225–2234.

294 Khavari, P.A., Peterson, C.L., Tamkun, J.W. et al. (1993). BRG1 contains a conserved domain of the SWI2/SNF2 family necessary for normal mitotic growth and transcription. *Nature* 366 (6451): 170–174.

295 Muchardt, C. and Yaniv, M. (1993). A human homologue of *Saccharomyces cerevisiae* SNF_2/SWI_2 and *Drosophila* brm genes potentiates transcriptional activation by the glucocorticoid receptor. *EMBO J.* 12 (11): 4279–4290.

296 Chiba, H., Muramatsu, M., Nomoto, A., and Kato, H. (1994). Two human homologues of *Saccharomyces cerevisiae* SWI_2/SNF_2 and Drosophila brahma are transcriptional coactivators cooperating with the estrogen receptor and the retinoic acid receptor. *Nucleic Acids Res.* 22 (10): 1815–1820.

297 Tamkun, J.W., Deuring, R., Scott, M.P. et al. (1992). Brahma: a regulator of *Drosophila* homeotic genes structurally related to the yeast transcriptional activator SNF_2/SWI_2. *Cell* 68 (3): 561–572.

298 Davis, J.L., Kunisawa, R., and Thorner, J. (1992). A presumptive helicase (MOT1 gene product) affects gene expression and is required for viability in the yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 12 (4): 1879–1892.

299 Delmas, V., Stokes, D.G., and Perry, R.P. (1993). A mammalian DNA-binding protein that contains a chromodomain and an SNF2/SWI2-like helicase domain. *Proc. Natl. Acad. Sci. U.S.A.* 90 (6): 2414–2418.

300 Ebbert, R., Birkmann, A., and Schuller, H.J. (1999). The product of the SNF_2/SWI_2 paralogue INO80 of *Saccharomyces cerevisiae* required for efficient expression of various yeast structural genes is part of a high-molecular-weight protein complex. *Mol. Microbiol.* 32 (4): 741–751.

301 Tagami, H. (2018). Purification of histone variant-interacting chaperone complexes. *Methods Mol. Biol. (Clifton, NJ)* 1832: 51–60.

302 Tagami, H., Ray-Gallet, D., Almouzni, G., and Nakatani, Y. (2004). Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. *Cell* 116 (1): 51–61.

303 Latreille, D., Bluy, L., Benkirane, M., and Kiernan, R.E. (2014). Identification of histone 3 variant 2 interacting factors. *Nucleic Acids Res.* 42 (6): 3542–3550.

304 Obri, A., Ouararhni, K., Papin, C. et al. (2014). ANP32E is a histone chaperone that removes H2A.Z from chromatin. *Nature* 505 (7485): 648–653.

305 Zink, L.M., Delbarre, E., Eberl, H.C. et al. (2017). H3.Y discriminates between HIRA and DAXX chaperone complexes and reveals unexpected insights into human DAXX-H3.3-H4 binding and deposition requirements. *Nucleic Acids Res.* 45 (10): 5691–5706.

306 Valero, M.L., Sendra, R., and Pamblanco, M. (2016). Tandem affinity purification of histones, coupled to mass spectrometry, identifies associated proteins and new sites of post-translational modification in *Saccharomyces cerevisiae*. *J. Proteomics* 136: 183–192.

307 Lewis, P.W., Elsaesser, S.J., Noh, K.M. et al. (2010). Daxx is an H3.3-specific histone chaperone and cooperates with ATRX in replication-independent chromatin assembly at telomeres. *Proc. Natl. Acad. Sci. U.S.A.* 107 (32): 14075–14080.

308 Sato, K., Ishiai, M., Toda, K. et al. (2012). Histone chaperone activity of Fanconi anemia proteins, FANCD2 and FANCI, is required for DNA crosslink repair. *EMBO J.* 31 (17): 3524–3536.

309 Osakabe, A., Tachiwana, H., Takaku, M. et al. (2013). Vertebrate Spt2 is a novel nucleolar histone chaperone that assists in ribosomal DNA transcription. *J. Cell Sci.* 126 (Pt 6): 1323–1332.

310 Mehrotra, P.V., Ahel, D., Ryan, D.P. et al. (2011). DNA repair factor APLF is a histone chaperone. *Mol. Cell* 41 (1): 46–55.

311 Drane, P., Ouararhni, K., Depaux, A. et al. (2010). The death-associated protein DAXX is a novel histone chaperone involved in the replication-independent deposition of H3.3. *Genes Dev.* 24 (12): 1253–1265.

312 Foltman, M., Evrin, C., De Piccoli, G. et al. (2013). Eukaryotic replisome components cooperate to process histones during chromosome replication. *Cell Rep.* 3 (3): 892–904.

313 Bellelli, R., Belan, O., Pye, V.E. et al. (2018). POLE3-POLE4 is a histone H3-H4 chaperone that maintains chromatin integrity during DNA replication. *Mol. Cell* 72 (1): 112–126 e5.

314 He, H., Li, Y., Dong, Q. et al. (2017). Coordinated regulation of heterochromatin inheritance by Dpb3-Dpb4 complex. *Proc. Natl. Acad. Sci. U.S.A.* 114 (47): 12524–12529.

315 Liu, S., Xu, Z., Leng, H. et al. (2017). RPA binds histone H3-H4 and functions in DNA replication-coupled nucleosome assembly. *Science* 355 (6323): 415–420.

316 Flaus, A., Martin, D.M., Barton, G.J., and Owen-Hughes, T. (2006). Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. *Nucleic Acids Res.* 34 (10): 2887–2905.

317 Becker, P.B. and Horz, W. (2002). ATP-dependent nucleosome remodeling. *Annu. Rev. Biochem.* 71: 247–273.

318 Bhardwaj, S.K., Hailu, S.G., Olufemi, L. et al. (2020). Dinucleosome specificity and allosteric switch of the ISW1a ATP-dependent chromatin remodeler in transcription regulation. *Nat. Commun.* 11 (1): 5913.

319 Sabantsev, A., Levendosky, R.F., Zhuang, X. et al. (2019). Direct observation of coordinated DNA movements on the nucleosome during chromatin remodelling. *Nat. Commun.* 10 (1): 1720.

320 Blosser, T.R., Yang, J.G., Stone, M.D. et al. (2009). Dynamics of nucleosome remodelling by individual ACF complexes. *Nature* 462 (7276): 1022–1027.

321 Logie, C. and Peterson, C.L. (1997). Catalytic activity of the yeast SWI/SNF complex on reconstituted nucleosome arrays. *EMBO J.* 16 (22): 6772–6782.

322 Narlikar, G.J., Phelan, M.L., and Kingston, R.E. (2001). Generation and interconversion of multiple distinct nucleosomal states as a mechanism for catalyzing chromatin fluidity. *Mol. Cell* 8 (6): 1219–1230.

323 Maier, V.K., Chioda, M., Rhodes, D., and Becker, P.B. (2008). ACF catalyses chromatosome movements in chromatin fibres. *EMBO J.* 27 (6): 817–826.

324 Willhoft, O., Ghoneim, M., Lin, C.L. et al. (2018). Structure and dynamics of the yeast SWR1-nucleosome complex. *Science* 362 (6411): e17101. https://doi.org/10.7554/eLife.17101. PMID: 27595565; PMCID: PMC5012860.

325 Ludwigsen, J., Hepp, N., Klinker, H. et al. (2018). Remodeling and repositioning of nucleosomes in nucleosomal arrays. *Methods Mol. Biol. (Clifton, NJ)* 1805: 349–370.

326 Manning, B.J. and Yusufzai, T. (2017). The ATP-dependent chromatin remodeling enzymes CHD6, CHD7, and CHD8 exhibit distinct nucleosome binding and remodeling activities. *J. Biol. Chem.* 292 (28): 11927–11936.

327 Briggs, K., Al-Ani, G., Eastlund, A., and Fischer, C.J. (2018). Anisotropy-based nucleosome repositioning assay. *Methods Mol. Biol. (Clifton, NJ)* 1805: 333–347.

328 Xin, H., Takahata, S., Blanksma, M. et al. (2009). yFACT induces global accessibility of nucleosomal DNA without H2A-H2B displacement. *Mol. Cell* 35 (3): 365–376.

329 Dann, G.P., Liszczak, G.P., Bagert, J.D. et al. (2017). ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference. *Nature* 548 (7669): 607–611.

330 Senapati, P., Sudarshan, D., Gadad, S.S. et al. (2015). Methods to study histone chaperone function in nucleosome assembly and chromatin transcription. *Methods Mol. Biol. (Clifton, NJ)* 1288: 375–394.

331 Latrick, C.M., Marek, M., Ouararhni, K. et al. (2016). Molecular basis and specificity of H2A.Z-H2B recognition and deposition by the histone chaperone YL1. *Nat. Struct. Mol. Biol.* 23 (4): 309–316.

332 Andrews, A.J., Chen, X., Zevin, A. et al. (2010). The histone chaperone Nap1 promotes nucleosome assembly by eliminating nonnucleosomal histone DNA interactions. *Mol. Cell* 37 (6): 834–842.

333 Goldberg, A.D., Banaszynski, L.A., Noh, K.M. et al. (2010). Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell* 140 (5): 678–691.

334 Jeronimo, C., Watanabe, S., Kaplan, C.D. et al. (2015). The histone chaperones FACT and Spt6 restrict H2A.Z from intragenic locations. *Mol. Cell* 58 (6): 1113–1123.

335 Pchelintsev, N.A., McBryan, T., Rai, T.S. et al. (2013). Placing the HIRA histone chaperone complex in the chromatin landscape. *Cell Rep.* 3 (4): 1012–1019.

336 Mao, Z., Pan, L., Wang, W. et al. (2014). Anp32e, a higher eukaryotic histone chaperone directs preferential recognition for H2A.Z. *Cell Res.* 24 (4): 389–399.

337 Kwon, S.Y., Grisan, V., Jang, B. et al. (2016). Genome-wide mapping targets of the metazoan chromatin remodeling factor NURF reveals nucleosome remodeling at enhancers, core promoters and gene insulators. *PLos Genet.* 12 (4): e1005969.

338 Klein-Brill, A., Joseph-Strauss, D., Appleboim, A., and Friedman, N. (2019). Dynamics of chromatin and transcription during transient depletion of the RSC chromatin remodeling complex. *Cell Rep.* 26 (1): 279–292 e5.

339 Liscovitch-Brauer, N., Montalbano, A., Deng, J. et al. (2020). Scalable pooled CRISPR screens with single-cell chromatin accessibility profiling. *bioRxiv*. https://doi.org/10.1101/2020.11.20.390971.

340 Gautier, A., Juillerat, A., Heinis, C. et al. (2008). An engineered protein tag for multiprotein labeling in living cells. *Chem. Biol.* 15 (2): 128–136.

341 Keppler, A., Gendreizig, S., Gronemeyer, T. et al. (2003). A general method for the covalent labeling of fusion proteins with small molecules in vivo. *Nat. Biotechnol.* 21 (1): 86–89.

342 Torne, J., Orsi, G.A., Ray-Gallet, D., and Almouzni, G. (2018). Imaging newly synthesized and old histone variant dynamics dependent on chaperones using the SNAP-tag system. *Methods Mol. Biol. (Clifton, NJ)* 1832: 207–221.

343 Mao, P., Kyriss, M.N., Hodges, A.J. et al. (2016). A basic domain in the histone H2B N-terminal tail is important for nucleosome assembly by FACT. *Nucleic Acids Res.* 44 (19): 9142–9152.

344 Yen, K., Vinayachandran, V., Batta, K. et al. (2012). Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell* 149 (7): 1461–1473.

345 de Dieuleveult, M., Yen, K., Hmitou, I. et al. (2016). Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. *Nature* 530 (7588): 113–116.

346 Giles, K.A., Gould, C.M., Du, Q. et al. (2019). Integrated epigenomic analysis stratifies chromatin remodellers into distinct functional groups. *Epigenet. Chromatin* 12 (1): 12.

347 Ye, Z., Chen, Z., Sunkel, B. et al. (2016). Genome-wide analysis reveals positional-nucleosome-oriented binding pattern of pioneer factor FOXA1. *Nucleic Acids Res.* 44 (16): 7540–7554.

348 Goodman, J.V., Yamada, T., Yang, Y. et al. (2020). The chromatin remodeling enzyme Chd4 regulates genome architecture in the mouse brain. *Nat. Commun.* 11 (1): 3419.

349 Dultz, E., Mancini, R., Polles, G. et al. (2018). Quantitative imaging of chromatin decompaction in living cells. *Mol. Biol. Cell* 29 (13): 1763–1777.

350 Shastrula, P.K., Sierra, I., Deng, Z. et al. (2019). PML is recruited to heterochromatin during S phase and represses DAXX-mediated histone H3.3 chromatin assembly. *J. Cell Sci.* 132 (6): jcs220970. https://doi.org/10.1242/jcs.220970. PMID: 30796101; PMCID: PMC6451418.

351 Liu, C.P., Xiong, C., Wang, M. et al. (2012). Structure of the variant histone H3.3-H4 heterodimer in complex with its chaperone DAXX. *Nat. Struct. Mol. Biol.* 19 (12): 1287–1292.

352 Boeger, H., Griesenbeck, J., Strattan, J.S., and Kornberg, R.D. (2003). Nucleosomes unfold completely at a transcriptionally active promoter. *Mol. Cell* 11 (6): 1587–1598.

353 Brown, C.R., Mao, C., Falkovskaia, E. et al. (2011). In vivo role for the chromatin-remodeling enzyme SWI/SNF in the removal of promoter nucleosomes by disassembly rather than sliding. *J. Biol. Chem.* 286 (47): 40556–40565.

12

RNA–Protein Interactomics

Cornelia Kilchert

Justus-Liebig-Universität Gießen, Institut für Biochemie, Heinrich-Buff-Ring 17, 35392 Gießen, Germany

12.1 Introduction

In all living cells, the genetic material is stored in the form of deoxyribonucleic acid (DNA), a relatively static biomolecule that serves a single purpose: acting as a stable repository of hereditary information. That information is made available through the act of transcription, during which the DNA template is copied into a ribonucleic acid (RNA) transcript. Depending on cell type and DNA locus, the amount of transcript can range from <1 copy per cell (with the RNA present in only a few cells within a population) to thousands of copies per cell. The median half-life of mRNAs is roughly proportional to the length of the cell cycle: on the order of minutes in prokaryotes such as *Escherichia coli*, but approaching several hours in cultured mammalian cells [1–3]. Within organisms, RNA half-lives vary by orders of magnitude, ranging from seconds (for ephemeral noncoding RNA species that are only detected in the absence of the nuclear RNA decay machinery) to several days (for ribosomal RNAs) [3–6]. In eukaryotes, in particular, RNAs undergo a host of regulatory interventions [7]: After transcription, RNAs are processed through capping, splicing, and tailing reactions; some are edited, chemically modified, or subjected to partial ribonucleolysis. RNAs are then exported from the nucleus, and may be actively transported to a specific subcellular localization – an early *Drosophila* screen reported transcript-specific localization patterns for 71% of all mRNAs analyzed [8]. The vast majority of cellular RNAs then partake in protein biosynthesis, be it as a template, or as part of the machinery. In the end, all RNAs are degraded, either through constitutive decay pathways or because they trigger one of a variety of specialized quality control mechanisms [9]. All of these events are subject to regulation and critically depend on proteins that interact with RNA. Such RNA-binding proteins (RBPs) either target specific transcripts, groups of transcripts ("regulons"), or affect RNAs globally. Altogether, the number of proteins known to regulate aspects of primary RNA metabolism through direct interaction with RNA is vast. In the last decade, the emerging field of RNA–protein interactomics has expanded the list of known RBPs considerably, with new estimates considering between 1 in 6 and 1 in 12

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

of all proteins capable of interacting with RNA at least transiently [10–12]. Not all of these proteins have an RNA-directed function. Sometimes, it is the RNA that acts as a regulator, a scaffold, a protein sponge, or a guide [13].

This chapter will begin with some initial considerations on the nature of RNA–protein interactions and a short overview of the various functional roles these can play in the cellular context. Next, I will describe two basic experimental building blocks that are very frequently applied in the field and form the basis for many interactomics methods: metabolic RNA labeling and RNA–protein crosslinking. I will then introduce state-of-the-art methods that were developed to assess RNA–protein interactions. The focus will be on exploratory, larger-scale experiments ("interactomics"), which come in two flavors: Copurification or proximity-dependent labeling methods. As is advisable for all high-throughput data, RNA–protein interactions identified in such screens should be verified by direct experimental validation. There is a collection of methods that can be used to assay association of a specific protein with RNA, which include – but are not limited to – electrophoretic mobility shift assays (EMSA), polynucleotide kinase assays ("test-CLIPs"), fluorescence RNA-binding assays, and nuclear magnetic resonance (NMR) titration. Such methods have been described in detail elsewhere [14–18].

12.2 Interactions of Proteins with mRNA and ncRNA

RNA–protein interactions are very abundant. They are required for the RNA-directed functions of proteins that act on RNA – such as translation regulators or RNA-directed enzymes – and for protein-directed functions of RNAs that regulate protein activity, for example, long noncoding (lnc)RNAs that recruit silencing factors to the chromatin. A priori, we are inclined to assume that protein–mRNA interactions (where the "function" of the RNA, which primarily evolved as a template, is clearly defined) regulate translational output of the mRNA, whereas we consider it more plausible that protein–ncRNA interactions coordinate the function of proteins. Most likely, that distinction is arbitrary and the boundaries fluid.

For lncRNA, many different functions have been reported [13]. They can serve as scaffolds that bring proteins together or help to drive the formation of molecular condensates, thereby creating locations with a high local concentration of specific factors. Examples are NEAT1, which drives the assembly of paraspeckles [19, 20], or Xist RNA, which recruits silencing factors to the inactive X chromosome, where they modify histones to silence gene expression [21–24]. LncRNAs can also function as RBP sponges that suppress RBP function, like meiRNA, which sequesters the meiosis inhibitor Mmi1 at the onset of meiosis in fission yeast [25]. In addition, RNA can regulate protein activity either through allostery or by occupying enzymes' active centers in an act of molecular mimicry. Bacterial 6S RNA, for example, mimics B-form promoter DNA and the transcription bubble to block activity of the RNA polymerase holoenzyme [26]; Polycomb repressive complex 2 (PRC2), a histone methyltransferase, is inhibited by RNA through an allosteric mechanism [27]; Rossman fold domains, which are commonly found in metabolic enzymes and

bind nucleotide cofactors such as nicotinamide adenine dinucleotide (NAD), have a propensity to interact with RNA; here, RNA and cofactor binding can be either mutually exclusive or not [28, 29].

When we consider typical RNA–protein interactions *in vivo*, two very different binding modes come to mind: One of them is the stable, highly specific interaction of an RBP with a defined RNA sequence or structure. This is likely to involve a classical RNA-binding domain (RBD) – for example an RNA recognition motif (RRM), a double-stranded RNA-binding domain (dsRBD), or a Pumilio homology domain [30]. The YTH domain-containing protein Mmi1, which recognizes TNAAAC motifs and triggers selective RNA decay, is an example of such an interaction [31, 32]. Structures of stable RNA–RBP complexes abound (e.g. [33]). On the other hand, there are the rather more fluid interactions of RBPs and RNA within biomolecular condensates [34]. RNA has a propensity to trigger disorder-to-order transitions via weak, multivalent interactions; this can result in liquid–liquid phase separation (LLPS) and create subcellular domains with high local concentrations of RBPs and RNA. Many RNA-rich compartments constitute RNA coacervates – viscous, membrane-less organelles that are formed by LLPS and include the nucleolus, nuclear speckles, heterochromatin domains, and stress granules [35–39]. While some RBPs are drivers – or scaffolds – of phase separation, others are recruited as "clients" that partition to existing LLPS granules [34]. Often, intrinsically disordered regions (IDRs) are required for the localization of proteins to RNA granules (such as P-bodies) and can induce LLPS *in vivo* and *in vitro* [40–42]. Notably, IDRs are significantly enriched in RBPs and abundantly crosslink to RNA [10, 11, 29, 43]. For most RNA interactomics approaches described in this chapter, it does not matter whether an RNA–protein interaction occurs in isolation or in a more crowded RNA/RBP-rich environment – they will identify either one. By design, proximity-dependent labeling methods like APEX-Seq, which seeks to identify RNAs that are enriched in a specific subcellular compartment, are particularly useful for studying RBPs that localize to RNA-rich domains (e.g. [44]).

12.3 The Basic Toolbox

12.3.1 Metabolic RNA Labeling with Modified Nucleobases

RNA biologists frequently make use of chemical modifications on nucleobases that confer certain desired characteristics to the labeled RNA, in the best case without interfering with base pairing and normal RNA function. In the field of RNA interactomics, two classes of modifications have proven particularly useful: (i) photoreactive nucleosides, such as 4-thiouridine (4sU) and 6-thioguanosine (6sG), that allow selective RNA–protein crosslinking after irradiation with UV light at 365 nm wavelength [10, 11, 45, 46], and (ii) azide- or alkyne-modified nucleosides, particularly 5-ethynyluridine (5EU), that allow selective isolation of labeled RNAs on a streptavidin resin after linkage to biotin in a bioorthogonal click-chemistry reaction [47, 48].[1]

Cellular RNA polymerases can utilize a range of modified uridines, many of which can be used to label RNA *in vivo* with little toxic side effects. In eukaryotes, incorporation of 4sU is considered to be "nonperturbing": it was shown to lead to only moderate changes in gene expression and – in contrast to 6sG – to have little impact on cell proliferation rates, although it can disrupt rRNA biogenesis at higher concentrations [49, 52–54]. Similarly, few adverse effects have been reported after 5EU labeling; here, longer labeling times can lead to reduced growth rates [6, 48]. Note that the extent to which their incorporation perturbs cellular processes has not been rigorously tested for all modified nucleosides that are currently in use. In general, modified uracils tend to be cheaper than the corresponding modified nucleosides. The choice of uridine over uracil largely depends on the organism, specifically on whether key enzymes of the pyrimidine or nucleoside salvage pathways are present: higher eukaryotes, including insects and mammals, do not have a functional uracil phosphoribosyltransferase (UPRT) [55] and cannot incorporate externally provided labeled uracil into nucleic acids; here, labeled uridines are the only choice. However, uracil incorporation can be enabled by heterologous expression of UPRT from other species, for example, UPRT from the protozoan *Toxoplasma gondii* [50]. UPRT deficiency has even been turned into an advantage: cell-type-restricted expression of UPRT has been exploited to selectively label RNA in a subset of cells within complex tissues of various UPRT-deficient organisms, including *Drosophila*, zebrafish, and mouse [56–59]. Similarly, RNAs produced by virus-infected cells have been selectively labeled using engineered cytomegalovirus that encodes UPRT [60]. Yeasts, plants, and the nematode *Caenorhabditis elegans* have functional UPRT, so do archaea [55]. These organisms will incorporate uracil or uridine into RNA [4, 61–65]. Modified nucleoside uptake can be moderate, but has been improved by expression of the human equilibrative nucleoside transporter (hENT) in both *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* [52, 66]. In *S. cerevisiae*, uptake of 4-thiouracil could be genetically enhanced by overexpression of the uracil permease *FUI1* to enable very short labeling pulses [67]. More evolved targeting strategies have been developed for multicellular organisms, including treatment with a cytochrome P(450) 3A-activated prodrug that can only be converted to 5EU by cells that express CYP3A4, namely hepatocytes [68]. In *Drosophila*, combined expression of UPRT and cytosine deaminase (CD), which converts 5-ethynylcytosine (5EC) to 5EU, has been used to allow cell-type selective 5EU labeling ("EC-tagging") [69].

Commonly used modified nucleosides like 4sU or 5EU have not been used with great success in prokaryotes, where they can be very toxic. Recently, incorporation of clickable 2′-deoxy-2′-azidonucleosides (2AzU) was suggested to constitute a superior alternative in bacteria [70] (Table 12.1).

12.3.2 RNA–Protein Crosslinking

Exposure to 254 nm UV light induces covalent crosslinking of RNA and proteins at a short, "zero-length" distance [79, 80]. The chemistry is complex. Irradiation of nucleobases generates short-lived radicals that can form photoadducts with most amino

Table 12.1 Survey of modified uracils and uridines that have been successfully employed for metabolic RNA labeling in various organisms.

Organism	Modified uracils	Modified uridines
S. cerevisiae	4sU [71, 72], 4sU [FUI1] [67]	4sU [hENT1] [52]
S. pombe	4sU [73, 74]	4sU [75]
Mammals	4sU [UPRT] [58], 5EU [UPRT] [76]	4sU [10, 46], 5EU [47, 48]
C. elegans		4sU [63]
Drosophila	4sU [UPRT] [56], 5EC [UPRT, CD] [69]	4sU [77], 5EU [78]
Plants	4sU [65]	5EU [64]
Archaea	4sU [61]	
Prokaryotes		2AzU [70]

Genetic modifications that are required to ensure efficient labeling are given in square brackets. 4sU: 4-thiouracil/uridine; 5EC: 5-ethynylcytosine; 5EU: 5-ethynyluridine; 2AzU: 2′-deoxy-2′-azidouridine.

acid side chains (with uracil being most reactive), but cysteine, lysine, phenylalanine, tryptophan, and tyrosine are particularly prone to crosslink [81, 82]. *In vivo*, RNA crosslinks are better than DNA by orders of magnitude. However, even for RNA, the efficiency of UV crosslinking is low. Under typical irradiation settings, only a low percentage of RNA–protein interactions will be crosslinked [83]. This low efficiency of UV crosslinking is offset by its high specificity for nucleic acid interactions and the irreversible, covalent nature of the crosslink: because the linkage is stable, nonspecific background can be very effectively removed through stringent washes during the purification of crosslinked RNA–protein complexes [10, 84].

The advent of photoactivatable nucleoside-enhanced (PAR-)crosslinking of metabolically labeled 4sU- or 6sG-containing RNA at 365 nm UV light offered additional advantages: (i) irradiation at 365 nm does not activate DNA bases; therefore, cells are less likely to activate DNA damage pathways; (ii) compared to conventional 254 nm crosslinking, the efficiency of PAR-crosslinking is greatly enhanced (100–1000 fold) [46]; (iii) because metabolic labeling can be spatially and temporally controlled, selective crosslinking of RNA subpopulations is possible. Similar to natural nucleobases, 4sU will crosslink promiscuously upon irradiation, but aromatic amino acids – and particularly histidine – are most reactive [12]. UV crosslinking has the advantage of being very selective for RNA, but the procedure can be cumbersome and lengthy: cell culture media generally absorb UV light and have to be exchanged for an inert buffer prior to crosslinking. For PAR-crosslinking, doses of 3–12 J/cm^2 are not uncommon, and unless your lab has access to a high-density crosslinker, irradiation can take 15 minutes or longer. Because the material heats up during exposure, crosslinking is routinely performed on ice. Especially for dynamic interactomes (for example, to monitor RNA interactome remodeling in response to stress), alternative crosslinking procedures can be desirable that can be applied directly to the growing culture and freeze RNA–protein interactions more rapidly. For this reason, conventional formaldehyde (FA) crosslinking is also

widely applied in the field of RNA interactomics (e.g. [22, 85]). FA forms reactive adducts with nucleophilic amino acids, particularly lysine and arginine, and all four RNA nucleobases, particularly adenine [86, 87]. The reactions of these adducts are complex and can form covalent intramolecular linkages through various chemistries, including methylene bridges. FA crosslinking is reversible at high temperature [86–88]. Because FA readily crosslinks protein–protein interactions as well, FA crosslinking is more likely to stabilize large RNPs and – in contrast to UV crosslinking – will facilitate capture of subunits that are not in direct contact with the RNA. To limit the extent of protein–protein linkages, the samples can be subjected to very mild crosslinking conditions, e.g. 0.05% FA for 10 minutes, but more extensive treatment of up to 3% formaldehyde for 30 minutes has been successfully employed, and conditions should be optimized for each experiment [22, 85, 89]. A variety of alternative chemical crosslinkers can be used to fix RNA–protein interactions, including glutaraldehyde [90, 91], but have not been as popular as FA in the field.[2] As a consequence of the experimental setup, all data derived from crosslinked material are subject to sampling biases because the crosslinkability of RNA–protein interactions can vary with the method.

12.4 RNA–Protein Interactomics

12.4.1 What Proteins Are Bound to my RNA (or RNA in General)?

12.4.1.1 Cataloging the RBPome

Not all RBPs possess an easily recognizable RBD; for example, we now know that IDRs frequently contribute to RNA binding [29, 43]. This made it very challenging to predict RNA binding based on protein sequence alone and fueled the development of experimental approaches that aim to inventory all RNA–protein interactions within the cell.[3] The classical RNA interactome capture (RIC) method relies on UV crosslinking of RNA–protein complexes, followed by enrichment of polyadenylated RNA on oligo-d(T) beads and identification of copurified RBPs by mass spectrometry (MS) [10, 11] (Figure 12.1a). RIC has been applied to many different cell types under many different conditions, and detailed methods protocols have been published for various organisms [73, 97–100]. Adaptations include the inclusion of an additional proteolysis step to co-purify only the RNA-interacting regions (rather than the entire RBP) to more precisely map RNA-binding sites ("RNA-binding domain mapping", RBDmap; "Crosslinked and Adjacent Peptides-based RNA-binding domain Identification", CAPRI)[4] [29, 43] (Figure 12.1b). For some areas of research, a focus on polyadenylated RNA is a severe limitation, since nascent transcripts as well as many noncoding RNAs will not be recovered; it is also not applicable to prokaryotes. Total RNA RIC protocols either employ 5EU labeling and click-chemistry-assisted biotinylation to enrich RNA by streptavidin pull-down ("RNA interactome capture using click chemistry", RICK; "click-chemistry-assisted RIC", CARIC) [47, 93], or aim to selectively enrich crosslinked RNA–protein complexes based on the unique physicochemical attributes of crosslinked material (Figure 12.1c,d). Here, existing

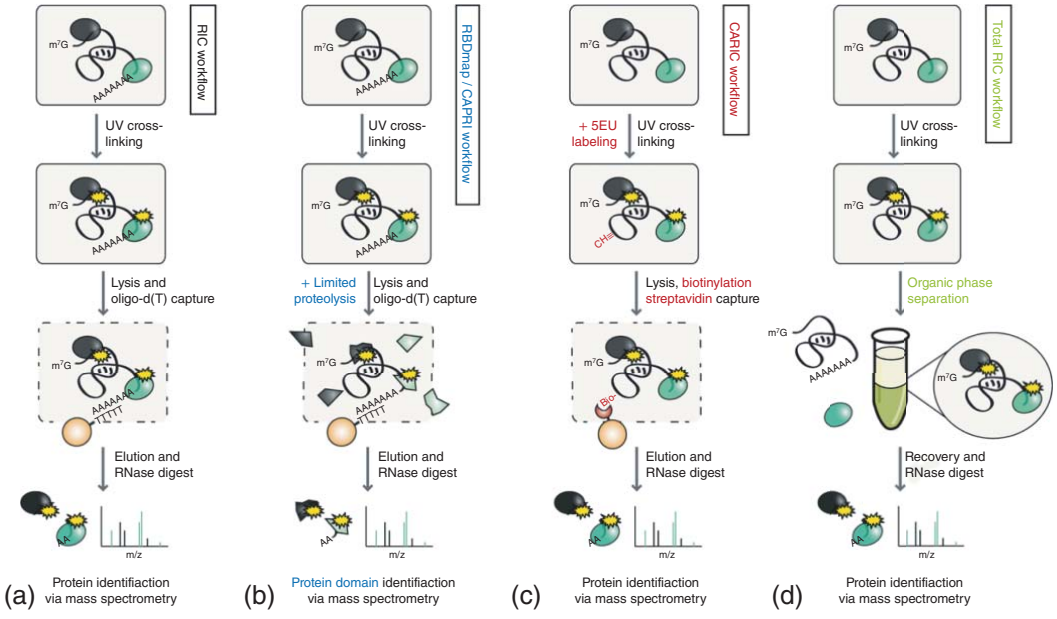


Figure 12.1 Simplified workflows of different global RNA interactome capture methods. (a) In the classical poly(A) + RNA interactome capture (RIC) protocol, RNA–protein complexes are UV crosslinked *in vivo* and polyadenylated RNA–protein complexes are enriched on oligo-d(T) beads after cell lysis. The RNA component of eluted RNPs is removed by RNase digestion and RNA-interacting proteins are identified by mass spectrometry. Source: Adapted from [10, 11]. (b) RBDmap and CAPRI add a limited proteolysis step to the RIC workflow to identify protein domains that mediate RNA binding. In the adapted protocol, those fragments of RNA-interacting proteins that directly crosslink to RNA will be retained during oligo-d(T) selection, while parts of the protein that are not in contact with the RNA will be removed by stringent washes. Source: Adapted from [29, 43]. (c) Click-chemistry-assisted RIC (CARiC) captures protein interactors of RNA that are metabolically labeled with 5-ethynyluridine (5EU). After UV crosslinking and cell lysis, 5EU-labeled RNA–protein complexes are biotinylated via a copper-catalyzed azide–alkyne click chemistry reaction and captured on streptavidin beads. RNA-interacting proteins are eluted by RNase digestion and identified by mass spectrometry. When labeling times are short, CARIC will selectively capture newly synthesized RNA. Source: Adapted from [47, 93]. (d) In total RNA RIC, UV-crosslinked RNA–protein complexes are isolated from the interphase after extraction with phenol-based organic solvents; free RNA will partition to the aqueous phase, while most unbound proteins will be partitioned into the organic phase. Additional extraction steps help to enrich RNA–protein complexes and efficiently remove unbound protein. After RNase digestion, RNA-interacting proteins are identified by mass spectrometry. Source: Adapted from [94–96].

protocols rely either on organic phase separation or on solid-phase extraction ("orthogonal organic phase separation", OOPS; "phenol toluol extraction", PTex; "protein-crosslinked RNA extraction", XRNAX; "total RNA-associated protein purification", TRAPP) [16, 94, 95, 101, 102]. PTex, for example, has been used to determine the RNA interactome of *Salmonella typhimurium* [16]. Because of their sheer abundance, the majority of observed RNA–protein interactions in total RNA RIC will be derived from ribosomes. RIC is semiquantitative and can – depending on the mode of normalization – provide estimates of both the average load of an RNA-interacting protein on RNA and the proportion of an RBP that is bound to RNA; this makes it a suitable method to monitor dynamic changes in the RNA interactome [103, 104]. Among others, comparative RIC has been used to assess RBPome remodeling during embryonic development, and in response to viral infection [105–107].

12.4.1.2 Interactomes of Specific RNAs

Often, researchers want to identify RBPs that are recruited to a specific RNA species – this could, for example, be a long noncoding (lnc)RNA that acts as a scaffold, or an RNA virus that has infected a cell. Biotinylated antisense oligos (ASOs) have been used to selectively capture lncRNAs like Xist, NEAT1, and MALAT1, or viruses like SARS-CoV-2, followed by identification of crosslinked proteins by MS ("capture hybridization analysis of RNA targets," CHART-MS; "comprehensive identification of RNA-binding proteins," ChIRP-MS; "identification of direct RNA-interacting proteins," iDRiP; "RNA antisense purification," RAP) [21–23, 108, 109] (Figure 12.2a). For Northern blots or in situ hybridization, long hybridization probes (several hundred bases) can be used to maximize sensitivity and specificity, but long probes are not compatible with interactomics approaches. Instead, many protocols use multiple ASOs that tile the length of the RNA [21–23, 108, 109]. In addition, the locked nucleic acid (LNA) technology can facilitate the design of probes that hybridize with high affinity [104, 112–114]. LNAs are nucleic acid analogs with a methylene bridge that connects the C-4′ carbon of the ribose with the oxygen atom at C-2′. This modification constrains the conformation of the ribose and significantly stabilizes duplex formation. For all RNA antisense purification experiments, crosslinking of RNA–protein interactions is highly advisable. Because nucleic acids are poly-anions, the stability of a nucleic acid double helix depends on the presence of counterions that shield the negative charges of the phosphates in the backbone. Thus, the specificity with which an ASO hybridizes to its target RNA increases with decreasing salt concentration; this behavior is described by the Schildkraut–Lifson equation [115]. RNA–protein interactions, on the other hand, often involve salt bridges between positively charged amino acids and the phosphate backbone. While higher salt concentrations can serve to screen electrostatic interactions, RBPs tend to be sticky under conditions where ASO hybridization is highly selective. On crosslinked samples, stringent prewashes at higher salt concentrations can be included which will reduce background binding.

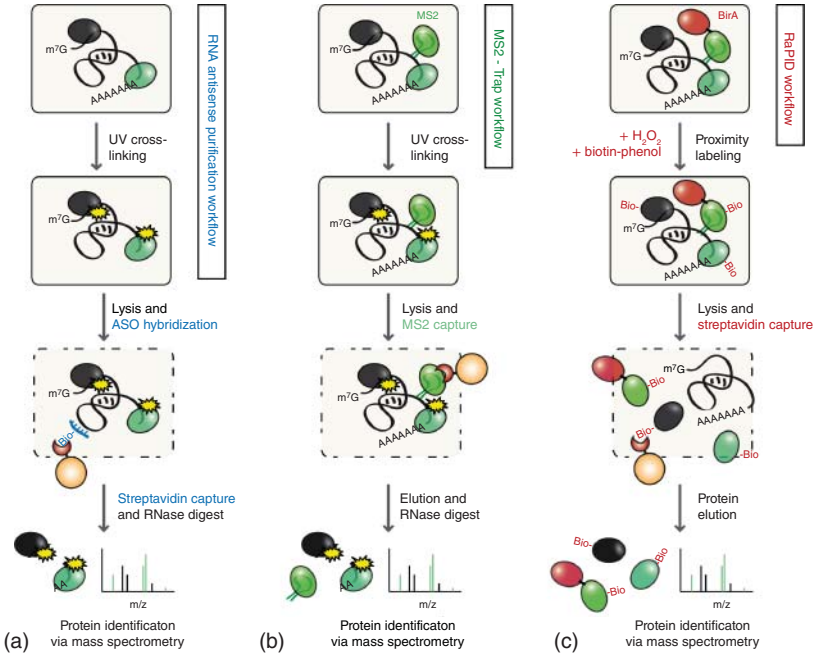


Figure 12.2 Simplified workflows of different RNA-specific RNA interactome capture methods. (a) RNA antisense purification protocols use biotinylated antisense oligos (ASO) to selectively capture an RNA of interest. RNA–protein interactions are stabilized by UV or formaldehyde crosslinking prior to lysate generation. After hybridization of the ASO, RNA–protein complexes are captured on streptavidin beads. RNA-interacting proteins are eluted by RNase digestion and identified by mass spectrometry. Source: Adapted from [21–23, 108, 109]. (b) MS2-Trap is an RNA-tagging-based approach that can be used to purify an RNA of interest. It requires heterologous expression of the phage-derived MS2 coat protein and the insertion of an aptamer sequence – in this case, MS2 stem-loops – into the RNA that is to be captured. After crosslinking and lysate preparation, tagged RNA–protein complexes can be selected via affinity tag purification. The RNA component of eluted RNPs is removed by RNase digestion and RNA-interacting proteins are identified by mass spectrometry. Source: Adapted from [110]. (c) RaPID combines the MS2-Trap technology with proximity-dependent protein labeling. Here, a fusion of the MS2 coat protein to the biotin ligase BirA is recruited to an RNA that carries an MS2 stem-loop structure. BirA generates activated biotin- 5′-AMP-esters that will biotinylate proteins in the immediate vicinity. After lysate preparation, biotinylated proteins are captured on streptavidin beads and identified by mass spectrometry. Source: Adapted from [111].

As an alternative to RNA antisense purification, various genetic RNA tagging approaches have been developed [116]. These include classical bacteriophage-derived systems that exploit the ability of the MS2 or PP7 phage coat protein to capture the RNA phage genome. The interaction relies on the recognition of a specific RNA hairpin structure present in the phage genome. If this hairpin structure is inserted into an RNA of choice, e.g. in the 3′ untranslated region of an mRNA, the RNA can be efficiently captured by a heterologously expressed, affinity-tagged coat protein [110] (Figure 12.2b). In addition, *in vitro* selection has enabled the creation of a range of RNA aptamers with stable folds that selectively bind proteins

like streptavidin and GFP, or small molecules like tobramycin or thiazole orange [117–119]. After RNA pulldown, specific protein interactors can be identified by MS. RNA tagging has also been combined with proximity-dependent labeling. For this, different stem loop-binding proteins are fused to a biotin ligase to biotinylate proteins in the immediate vicinity of the tagged RNA; biotinylated proteins are then enriched on streptavidin beads and identified by MS ("RNA–protein interaction detection", RaPID) [111, 120] (Figure 12.2c). For example, this strategy has been used to identify proteins associated with ß-actin mRNA, which localizes to protrusions in migrating fibroblasts [120]. Because RNA tagging can alter the function of RNA molecules, any strains generated for RNA tag-dependent purification should be assayed for functionality *in vivo* prior to the experiment.

12.4.2 Which RNA Species Are Bound by my RBP?

12.4.2.1 Copurification Methods: CLIP and Derivatives

To map the RNA-binding sites of a specific RBP across the transcriptome, derivatives of the crosslinking and immunoprecipitation (CLIP) protocol are the method of choice [84, 121]. As in RIC, RNA–protein complexes are first crosslinked *in vivo* using conventional UV or PAR-crosslinking [46, 84]. After cell lysis and partial RNA digestion, the RBP is enriched by affinity purification. The crosslinked RNA is then end-labeled and RNA–protein complexes are purified by gel electrophoresis and membrane transfer. After proteinase digestion of the RBP, the recovered RNA fragments are converted into a cDNA library for high-throughput sequencing (Figure 12.3a). Base conversions and frequent stalling of the reverse transcriptase at crosslink sites allow to map RBP binding at individual nucleotide resolution [46, 127]. While the original protocol was not the type of experiment you would want to hand to an inexperienced graduate student, newer versions of the protocol, such as enhanced CLIP (eCLIP) or iCLIP2, are faster and more robust [122, 123]. eCLIP is part of the methods portfolio of the ENCORE project within ENCODE, which aims to identify RNA-binding elements for all human RBPs in K562 and HepG2 cell lines (www.encodeproject.com). Regardless of the protocol used, high-quality CLIP data are dependent on the availability of either a good antibody or a well-behaved RBP fusion to an affinity tag.

12.4.2.2 Proximity-Dependent Labeling Methods

Proximity-Dependent RNA Editing TRIBE ("targets of RBPs identified by editing") was developed as a crosslinking- and immunoprecipitation-independent alternative to CLIP [124, 125]. Here, the RBP of choice is expressed as a protein fusion to the catalytic domain of the RNA-editing enzyme adenosine deaminase (ADAR), which converts adenosine to inosine; this fusion protein will edit transcripts bound by the RBP (Figure 12.3b). After RNA sequencing, editing sites can be identified as A to C conversions and quantified relative to control to identify RNA targets. Use of a hyperactive version of ADAR further improves sensitivity (HyperTRIBE) [125]. For example, HyperTRIBE was used to identify mRNAs bound by MUSAHI-2, an important regulator of human hematopoiesis [128]. TRIBE works in an analogous fashion

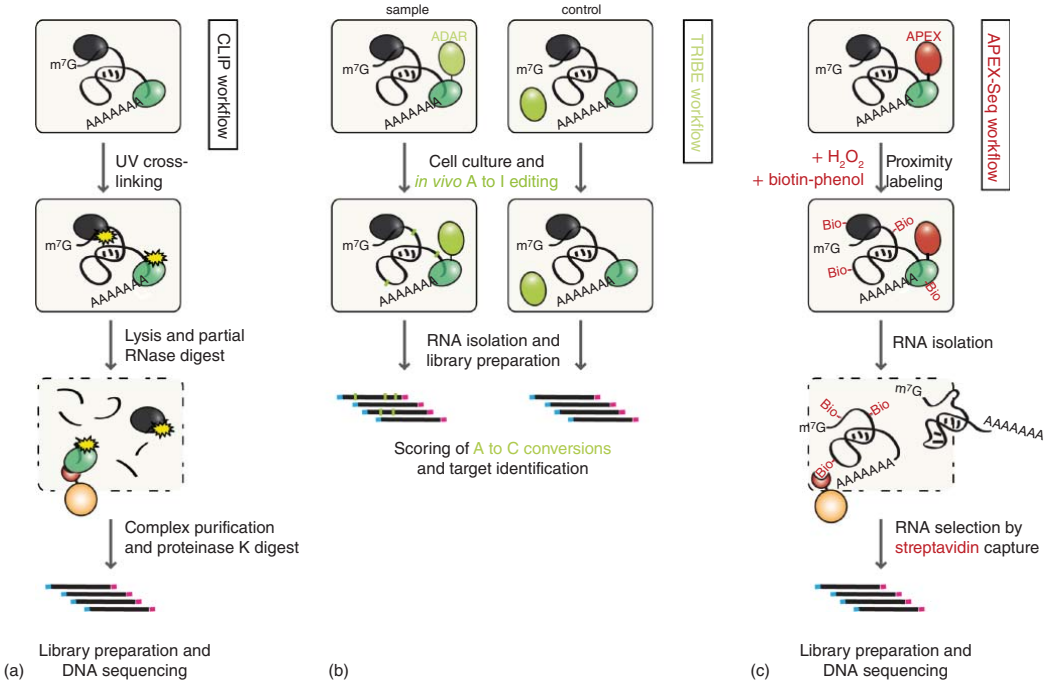


Figure 12.3 Simplified workflows of different protein-centric RNA–protein interaction methods. (a) Crosslinking immunoprecipitation (CLIP) identifies RNAs that interact with a protein of interest. For this, RNA–protein complexes are UV crosslinked *in vivo* and subjected to partial ribonucleolysis after lysate preparation. The RBP of interest is enriched by immunoprecipitation, and RNA–protein complexes are purified further by gel electrophoresis and size selection. Crosslinked RNA fragments are released by proteinase K treatment and converted into a cDNA library for sequencing. Source: Adapted from [122, 123]. (b) TRIBE uses the genetic fusion of the RNA editing enzyme ADAR to an RBP of interest to promote A to I editing of RNA transcripts that are bound by the RBP. After RNA isolation, cDNA library generation and sequencing, editing events can be identified as A to C conversions to identify RBP targets. Source: Adapted from [124, 125]. (c) APEX is an ascorbate peroxidase that has been engineered for optimal proximity-dependent protein labeling. If supplied with H_2O_2, APEX will convert biotin–phenol into a short-lived, highly reactive radical that will biotinylate biomolecules in the immediate vicinity, including RNA. If fused to an RBP of interest, APEX-dependent biotinylation of RNA can be used to identify target RNAs. After RNA isolation, biotinylated transcripts are enriched on streptavidin beads and converted to a cDNA library for sequencing. Source: Adapted from [44, 126].

to DamID, which employs a DNA adenine methyltransferase (Dam) fusion to map DNA-binding sites of a protein across the genome [129], and mirrors its advantages and disadvantages: ADAR fusions can be toxic if they lead to excessive editing that interferes with RNA function. If that is not the case, TRIBE is very sensitive and easy to implement. However, as A to I editing is irreversible, the method is not well suited to observe dynamic changes in RNA–protein interactions.

Proximity-Dependent Biotinylation Proximity labeling approaches have been developed to map transient protein–protein interaction networks [130]. They employ genetic fusions to an engineered enzyme that converts a biotin-containing substrate to a short-lived reactive species with a limited diffusive range that readily attaches to proteins *in vivo*. Two classes of enzymes are most commonly used: (i) Biotin ligases, which convert ATP and biotin to an activated AMP-ester, such as TurboID [131]; and (ii) peroxidases, which generate reactive radicals from exogenously supplied substrates like biotin–phenol in the presence of H_2O_2, such as the engineered ascorbate peroxidase APEX [132]. If APEX is attached to an RBP or a component of RNA granules, APEX-generated radicals will also biotinylate RNAs in a spatially restricted manner. These can be captured by streptavidin pulldown and sequenced, thereby identifying RNAs that are enriched in specific compartments (APEX-Seq) [44, 126] (Figure 12.3c). Alternatively, proximity-dependent biotinylation of proteins can be combined with RNA–protein crosslinking and streptavidin pulldown to isolate RNAs that are enriched at a certain subcellular location (APEX-RIP, proximity CLIP) [133, 134]. Both APEX-Seq and APEX-RIP have been used to map mRNA localization to various intracellular organelles [126, 133]. Because peroxidases like APEX or APEX2 require very short labeling times (<1 minute), proximity labeling approaches can be used to study highly dynamic processes.

12.5 Outlook

By harnessing the increased sensitivity of mass spectrometry instrumentation and the advances in high-throughput sequencing technologies, the field of RNA–protein interactomics has evolved rapidly in the last decade. As a consequence, our view of RNA–protein interactions has diversified: We now realize that (i) depending on the conditions, proteins with primary functions outside of RNA metabolism can significantly contribute to RNA–protein interactions; (ii) RNA can play an important role in scaffolding protein assemblies and coordinating their function. Because we now have a wealth of comparative RNA–protein interactomics data available that were generated with the methods described in this chapter, we know that the RNA–protein interaction landscape is highly dynamic. In the future, we will likely see the development of novel methodologies that help us to monitor the remodeling of RNA–protein interactions on shorter timescales. We also expect to see an increase in approaches that are sensitive to the presence of posttranslational modifications on RBPs and attempt to link these to change in RNA-binding activity on a proteome-wide scale, building on the versatile toolbox that is already available.

Notes

1 Biotinylation of 4sU-labeled RNA with HPDP-biotin or MTS-biotin has been used with great success in the field of transcriptomics [49–51]; however, the prevalence of thiol groups in proteins severely limits the usefulness of thiol-directed biotin conjugation for interactomics studies.
2 Wherever crosslinked RNA is to be sequenced in the course of the experiment, concentrations of irreversible crosslinkers should be kept low and incubation times short, because nucleobase adducts will interfere with reverse transcription.
3 Based on the extensive experimental data now available, machine-learning-based algorithms are able to more faithfully predict RNA binding, for example by scoring the presence of short linear motifs [92].
4 Researchers in the transcriptomics field have an inordinate fondness for method acronyms. Sometimes, different acronyms are used by different labs for what is arguably the same (or a highly related) method. In other cases, a new acronym can be an optimized version of an older protocol. In this chapter, we group methods that are conceptually similar. Readers are referred to the primary literature to identify advantages and disadvantages of the individual protocols.

References

1 Schwanhäusser, B., Busse, D., Li, N. et al. (2011). Global quantification of mammalian gene expression control. *Nature* 473 (7347): 337–342.
2 Bernstein, J.A., Khodursky, A.B., Lin, P.H. et al. (2002). Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.* 99 (15): 9697–9702.
3 Chan, L.Y., Mugler, C.F., Heinrich, S. et al. (2018). Non-invasive measurement of mRNA decay reveals translation initiation as the major determinant of mRNA stability. *eLife* 7: e32536.
4 Eser, P., Wachutka, L., Maier, K.C. et al. (2016). Determinants of RNA metabolism in the *Schizosaccharomyces pombe* genome. *Mol. Syst. Biol.* 12 (2): 857.
5 Clark, M.B., Johnston, R.L., Inostroza-Ponta, M. et al. (2012). Genome-wide analysis of long noncoding RNA stability. *Genome Res.* 22 (5): 885–898.
6 Tani, H., Mizutani, R., Salam, K.A. et al. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* 22 (5): 947–956.
7 Licatalosi, D.D. and Darnell, R.B. (2010). RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* 11 (1): 75–87.
8 Lécuyer, E., Yoshida, H., Parthasarathy, N. et al. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131 (1): 174–187.

9 Wolin, S.L. and Maquat, L.E. (2019). Cellular RNA surveillance in health and disease. *Science* 366 (6467): 822–827.

10 Castello, A., Fischer, B., Eichelbaum, K. et al. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 149 (6): 1393–1406.

11 Baltz, A.G., Munschauer, M., Schwanhäusser, B. et al. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* 46 (5): 674–690.

12 Shchepachev, V., Bresson, S., Spanos, C. et al. (2019). Defining the RNA interactome by total RNA-associated protein purification. *Mol. Syst. Biol.* 15 (4): e8689.

13 Yao, R.W., Wang, Y., and Chen, L.L. (2019). Cellular functions of long noncoding RNAs. *Nat. Cell Biol.* 21 (5): 542–551.

14 Strein, C., Alleaume, A.-M., Rothbauer, U. et al. (2014). A versatile assay for RNA-binding proteins in living cells. *RNA* 20 (5): 721–731.

15 Fillebeen, C., Wilkinson, N., and Pantopoulos, K. (2014). Electrophoretic mobility shift assay (EMSA) for the study of RNA-protein interactions: the IRE/IRP example. *J. Vis. Exp.* 94: 52230.

16 Urdaneta, E.C., Vieira-Vieira, C.H., Hick, T. et al. (2019). Purification of cross-linked RNA–protein complexes by phenol-toluol extraction. *Nat. Commun.* 10 (1): 990.

17 Licatalosi, D.D., Ye, X., and Jankowsky, E. (2019). Approaches for measuring the dynamics of RNA–protein interactions. *Wiley Interdiscip. Rev.: RNA* 11 (1): e1565.

18 Dominguez, C., Schubert, M., Duss, O. et al. (2011). Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* 58 (1–2): 1–61.

19 Fox, A.H., Nakagawa, S., Hirose, T., and Bond, C.S. (2018). Paraspeckles: where long noncoding RNA meets phase separation. *Trends Biochem. Sci* 43 (2): 124–135.

20 Clemson, C.M., Hutchinson, J.N., Sara, S.A. et al. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol. Cell* 33 (6): 717–726.

21 McHugh, C.A., Chen, C.K., Chow, A. et al. (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* 521 (7551): 232–236.

22 Chu, C., Zhang, Q.C., Da Rocha, S.T. et al. (2015). Systematic discovery of Xist RNA binding proteins. *Cell* 161 (2): 404–416.

23 Minajigi, A., Froberg, J.E., Wei, C. et al. (2015). A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* 349 (6245): 1DUIMMY.

24 Brockdorff, N., Bowness, J.S., and Wei, G. (2020). Progress toward understanding chromosome silencing by Xist RNA. *Genes Dev.* 34 (11–12): 733–744.

25 Harigaya, Y., Tanaka, H., Yamanaka, S. et al. (2006). Selective elimination of messenger RNA prevents an incidence of untimely meiosis. *Nature* 442 (7098): 45–50.

26 Chen, J., Wassarman, K.M., Feng, S. et al. (2017). 6S RNA mimics B-form DNA to regulate *Escherichia coli* RNA polymerase. *Mol. Cell* 68 (2): 388–397.e6.

27 Zhang, Q., McKenzie, N.J., Warneford-Thomson, R. et al. (2019). RNA exploits an exposed regulatory site to inhibit the enzymatic activity of PRC2. *Nat. Struct. Mol. Biol.* 26 (3): 237–247.

28 Liao, Y., Castello, A., Fischer, B. et al. (2016). The cardiomyocyte RNA-binding proteome: links to intermediary metabolism and heart disease. *Cell Rep.* 16 (5): 1456–1469.

29 Castello, A., Fischer, B., Frese, C.K. et al. (2016). Comprehensive identification of RNA-binding domains in human cells. *Mol. Cell* 63 (4): 696–710.

30 Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* 8 (6): 479–490.

31 Kilchert, C., Wittmann, S., Passoni, M. et al. (2015). Regulation of mRNA levels by decay-promoting introns that recruit the exosome specificity factor Mmi1. *Cell Rep.* 13 (11): 2504–2515.

32 Yamashita, A., Shichino, Y., Tanaka, H. et al. (2012). Hexanucleotide motifs mediate recruitment of the RNA elimination machinery to silent meiotic genes. *Open Biol.* 2 (3): 120014.

33 Wang, C., Zhu, Y., Bao, H. et al. (2016). A novel RNA-binding mode of the YTH domain reveals the mechanism for recognition of determinant of selective removal by Mmi1. *Nucleic Acids Res.* 44 (2): 969–982.

34 Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* 18 (5): 285–298.

35 Strom, A.R., Emelyanov, A.V., Mir, M. et al. (2017). Phase separation drives heterochromatin domain formation. *Nature* 547 (7662): 241–245.

36 Larson, A.G., Elnatan, D., Keenen, M.M. et al. (2017). Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* 547 (7662): 236–240.

37 Fei, J., Jadaliha, M., Harmon, T.S. et al. (2017). Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *J. Cell Sci.* 130 (24): 4180–4192.

38 Feric, M., Vaidya, N., Harmon, T.S. et al. (2016). Coexisting liquid phases underlie nucleolar subcompartments. *Cell* 165 (7): 1686–1697.

39 Jain, S., Wheeler, J.R., Walters, R.W. et al. (2016). ATPase-modulated stress granules contain a diverse proteome and substructure. *Cell* 164 (3): 487–498.

40 Decker, C.J., Teixeira, D., and Parker, R. (2007). Edc3p and a glutamine/asparagine-rich domain of Lsm4p function in processing body assembly in *Saccharomyces cerevisiae*. *J. Cell Biol.* 179 (3): 437–449.

41 Reijns, M.A.M., Alexander, R.D., Spiller, M.P., and Beggs, J.D. (2008). A role for Q/N-rich aggregation-prone regions in P-body localization. *J. Cell Sci.* 121 (15): 2463–2472.

42 Nott, T.J., Petsalaki, E., Farber, P. et al. (2015). Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* 57 (5): 936–947.

43 Panhale, A., Richter, F.M., Ramírez, F. et al. (2019). CAPRI enables comparison of evolutionarily conserved RNA interacting regions. *Nat. Commun.* 10 (1): 2682.

44 Padrón, A., Iwasaki, S., and Ingolia, N.T. (2019). Proximity RNA labeling by APEX-Seq reveals the organization of translation initiation complexes and repressive RNA granules. *Mol. Cell* 75 (4): 875–887.e5.

45 Favre, A., Moreno, G., Blondel, M.O. et al. (1986). 4-thiouridine photosensitized RNA–protein crosslinking in mammalian cells. *Biochem. Biophys. Res. Commun.* 141 (2): 847–854.

46 Hafner, M., Landthaler, M., Burger, L. et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141 (1): 129–141.

47 Bao, X., Guo, X., Yin, M. et al. (2018). Capturing the interactome of newly transcribed RNA. *Nat. Methods* 15 (3): 213–220.

48 Jao, C.Y. and Salic, A. (2008). Exploring RNA transcription and turnover in vivo by using click chemistry. *Proc. Natl. Acad. Sci. U.S.A.* 105 (41): 15779–15784.

49 Dölken, L., Ruzsics, Z., Rädle, B. et al. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* 14 (9): 1959–1972.

50 Cleary, M.D., Meiering, C.D., Jan, E. et al. (2005). Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat. Biotechnol.* 23 (2): 232–237.

51 Duffy, E.E., Rutenberg-Schoenberg, M., Stark, C.D. et al. (2015). Tracking distinct RNA populations using efficient and reversible covalent chemistry. *Mol. Cell* 59 (5): 858–866.

52 Miller, C., Schwalb, B., Maier, K. et al. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* 7: 458.

53 Melvin, W.T., Milne, H.B., Slater, A.A. et al. (1978). Incorporation of 6-thioguanosine and 4-thiouridine into RNA: application to isolation of newly synthesised RNA by affinity chromatography. *Eur. J. Biochem.* 92 (2): 373–379.

54 Burger, K., Mühl, B., Kellner, M. et al. (2013). 4-Thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biol.* 10 (10): 1623–1630.

55 Li, J., Huang, S., Chen, J. et al. (2007). Identification and characterization of human uracil phosphoribosyltransferase (UPRTase). *J. Hum. Genet.* 52 (5): 415–422.

56 Miller, M.R., Robinson, K.J., Cleary, M.D., and Doe, C.Q. (2009). TU-tagging: cell type-specific RNA isolation from intact complex tissues. *Nat. Methods* 6 (6): 439–441.

57 Matsushima, W., Herzog, V.A., Neumann, T. et al. (2018). SLAM-ITseq: sequencing cell type-specific transcriptomes without cell sorting. *Development* 145 (13), dev164640.

58 Gay, L., Miller, M.R., Ventura, P.B. et al. (2013). Mouse TU tagging: a chemical/genetic intersectional method for purifying cell type-specific nascent RNA. *Genes Dev.* 27 (1): 98–115.

59 Tallafuss, A., Kelly, M., Gay, L. et al. (2015). Transcriptomes of post-mitotic neurons identify the usage of alternative pathways during adult and embryonic neuronal differentiation. *BMC Genomics* 16: 1100.

60 Roche, K.L., Nukui, M., Krishna, B.A. et al. (2018). Selective 4-thiouracil labeling of RNA transcripts within latently infected cells after infection with human cytomegalovirus expressing functional uracil phosphoribosyltransferase. *J. Virol.* 92 (21), e00880-18.

61 Knüppel, R., Kuttenberger, C., and Ferreira-Cerca, S. (2017). Toward time-resolved analysis of RNA metabolism in archaea using 4-thiouracil. *Front. Microbiol.* 8 (FEB): 286.

62 Hasan, A., Cotobal, C., Duncan, C.D.S., and Mata, J. (2014). Systematic analysis of the role of RNA-binding proteins in the regulation of RNA stability. *PLoS Genet.* 10 (11): e1004684.

63 Jungkamp, A.C., Stoeckius, M., Mecenas, D. et al. (2011). In vivo and transcriptome-wide identification of RNA binding protein target sites. *Mol. Cell* 44 (5): 828–840.

64 Szabo, E.X., Reichert, P., Lehniger, M.K. et al. (2020). Metabolic labeling of RNAs uncovers hidden features and dynamics of the arabidopsis transcriptome. *Plant Cell* 32 (4): 871–887.

65 Sidaway-Lee, K., Costa, M.J., Rand, D.A. et al. (2014). Direct measurement of transcription rates reveals multiple mechanisms for configuration of the Arabidopsis ambient temperature response. *Genome Biol.* 15 (3): R45.

66 Hodson, J.A., Bailis, J.M., and Forsburg, S.L. (2003). Efficient labeling of fission yeast *Schizosaccharomyces pombe* with thymidine and BUdR. *Nucleic Acids Res.* 31 (21).

67 Barrass, J.D., Reid, J.E.A., Huang, Y. et al. (2015). Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biol.* 16 (1): 282.

68 Darr, J., Tomar, A., Lassi, M. et al. (2020). iTAG-RNA isolates cell-specific transcriptional responses to environmental stimuli and identifies an RNA-based endocrine axis. *Cell Rep.* 30 (9): 3183–3194.e4.

69 Hida, N., Aboukilila, M.Y., Burow, D.A. et al. (2017). EC-tagging allows cell type-specific RNA analysis. *Nucleic Acids Res.* 45 (15): e138.

70 Meng, L., Guo, Y., Tang, Q. et al. (2020). Metabolic RNA labeling for probing RNA dynamics in bacteria. *Nucleic Acids Res.* 48 (22): 12566–12576.

71 Beckmann, B.M., Horos, R., Fischer, B. et al. (2015). The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat. Commun.* 6: 10127.

72 Creamer, T.J., Darby, M.M., Jamonnak, N. et al. (2011). Transcriptome-wide binding sites for components of the *Saccharomyces cerevisiae* non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genet.* 7 (10): e1002329.

73 Kilchert, C., Hester, S., Castello, A. et al. (2020). Comparative poly(A)+ RNA interactome capture of RNA surveillance mutants. *Methods Mol. Biol.* 2062: 255–276.

74 Parsa, J.Y., Boudoukha, S., Burke, J. et al. (2018). Polymerase pausing induced by sequence-specific RNA-binding protein drives heterochromatin assembly. *Genes Dev.* 32 (13–14): 953–964.

75 Amorim, M.J., Cotobal, C., Duncan, C., and Mata, J. (2010). Global coordination of transcriptional control and mRNA decay during cellular differentiation. *Mol. Syst. Biol.* 6: 380.

76 Zajaczkowski, E.L., Zhao, Q.Y., Zhang, Z.H. et al. (2018). Bioorthogonal metabolic labeling of nascent RNA in neurons improves the sensitivity of transcriptome-wide profiling. *ACS Chem. Neurosci.* 9 (7): 1858–1865.

77 Hansen, H.T., Rasmussen, S.H., Adolph, S.K. et al. (2015). *Drosophila* Imp iCLIP identifies an RNA assemblage coordinating F-actin formation. *Genome Biol.* 16 (1): 123.

78 Kwasnieski, J.C., Orr-Weaver, T.L., and Bartel, D.P. (2019). Early genome activation in *Drosophila* is extensive with an initial tendency for aborted transcripts and retained introns. *Genome Res.* 29 (7): 1188–1197.

79 Greenberg, J.R. (1979). Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Res.* 6 (2): 715–732.

80 Wagenmakers, A.J., Reinders, R.J., and van Venrooij, W.J. (1980). Cross-linking of mRNA to proteins by irradiation of intact cells with ultraviolet light. *Eur. J. Biochem.* 112 (2): 323–330.

81 Shetlar, M.D., Christensen, J., and Hom, K. (1984). Photochemical addition of amino acids and peptides to DNA. *Photochem. Photobiol.* 39 (2): 125–133.

82 Shetlar, M.D., Carbone, J., Steady, E., and Hom, K. (1984). Photochemical addition of amino acids and peptides to polyuridylic acid. *Photochem. Photobiol.* 39 (2): 141–144.

83 Darnell, R.B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip. Rev:. RNA* 1 (2): 266–286.

84 Ule, J., Jensen, K.B., Ruggiu, M. et al. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302 (5648): 1212–1215.

85 Sharma, S., Poetz, F., Bruer, M. et al. (2016). Acetylation-dependent control of global poly(A) RNA degradation by CBP/p300 and HDAC1/2. *Mol. Cell* 63 (6): 927–938.

86 Masuda, N., Ohnishi, T., Kawamoto, S. et al. (1999). Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic Acids Res.* 27 (22): 4436–4443.

87 Sutherland, B.W., Toews, J., and Kast, J. (2008). Utility of formaldehyde cross-linking and mass spectrometry in the study of protein–protein interactions. *J. Mass Spectrom.* 43 (6): 699–715.

88 Kamps, J.J.A.G., Hopkinson, R.J., Schofield, C.J., and Claridge, T.D.W. (2019). How formaldehyde reacts with amino acids. *Commun. Chem.* 2 (1): 1–14.

89 Niranjanakumari, S., Lasda, E., Brazas, R., and Garcia-Blanco, M.A. (2002). Reversible cross-linking combined with immunoprecipitation to study RNA–protein interactions in vivo. *Methods* 26 (2): 182–190.

90 Chu, C., Qu, K., Zhong, F.L. et al. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin Interactions. *Mol. Cell* 44 (4): 667–678.

91 Migneault, I., Dartiguenave, C., Bertrand, M.J., and Waldron, K.C. (2004). Glutaraldehyde: behavior in aqueous solution, reaction with proteins, and application to enzyme crosslinking. *Biotechniques* 37 (5): 790–802.

92 Bressin, A., Schulte-Sasse, R., Figini, D. et al. (2019). TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs. *Nucleic Acids Res.* 47 (9): 4406–4417.

93 Huang, R., Han, M., Meng, L., and Chen, X. (2018). Capture and identification of RNA-binding proteins by using click chemistry-assisted RNA-interactome capture (CARIC) strategy. *J. Vis. Exp.* 140: e58580.

94 Queiroz, R.M.L., Smith, T., Villanueva, E. et al. (2019). Comprehensive identification of RNA–protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat. Biotechnol.* 37 (2): 169–178.

95 Trendel, J., Schwarzl, T., Horos, R. et al. (2019). The human RNA-binding proteome and its dynamics during translational arrest. *Cell* 176 (1–2): 391–403.e19.

96 Urdaneta, E.C. and Beckmann, B.M. (2019). Fast and unbiased purification of RNA–protein complexes after UV cross-linking. *Methods* 178: 72–82.

97 Castello, A., Horos, R., Strein, C. et al. (2013). System-wide identification of RNA-binding proteins by interactome capture. *Nat. Protoc.* 8 (3): 491–500.

98 Beckmann, B.M. (2017). RNA interactome capture in yeast. *Methods* 118–119: 82–92.

99 Köster, T., Reichel, M., and Staiger, D. (2020). CLIP and RNA interactome studies to unravel genome-wide RNA–protein interactions in vivo in *Arabidopsis thaliana*. *Methods* 178: 63–71.

100 Kilchert, C., Sträßer, K., Kunetsky, V., and Änkö, M.L. (2019). From parts lists to functional significance—RNA–protein interactions in gene regulation. *Wiley Interdiscip. Rev.: RNA* e1582.

101 Shchepachev, V., Bresson, S., Spanos, C. et al. (2018). Defining the RNA interactome by total RNA-associated protein purification. *Mol. Syst. Biol.* 15 (4): e8689.

102 Asencio, C., Chatterjee, A., and Hentze, M.W. (2018). Silica-based solid-phase extraction of cross-linked nucleic acid-bound proteins. *Life Sci. Alliance* 1 (3): e201800088.

103 Kilchert, C., Kecman, T., Priest, E. et al. (2020). System-wide analyses of the fission yeast poly(A)+RNA interactome reveal insights into organization and function of RNA–protein complexes. *Genome Res.* 30 (7): 1012–1026.

104 Perez-Perri, J.I., Noerenberg, M., Kamel, W. et al. (2021). Global analysis of RNA-binding protein dynamics by comparative and enhanced RNA interactome capture. *Nat. Protoc.* 16 (1): 27–60.

105 Sysoev, V.O., Fischer, B., Frese, C.K. et al. (2016). Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in Drosophila. *Nat. Commun.* 7 (1): 12128.

106 Despic, V., Dejung, M., Gu, M. et al. (2017). Dynamic RNA–protein interactions underlie the zebrafish maternal-to-zygotic transition. *Genome Res.* 27 (7): 1184–1194.

107 Garcia-Moreno, M., Noerenberg, M., Ni, S. et al. (2019). System-wide profiling of RNA-binding proteins uncovers key regulators of virus infection. *Mol. Cell* 74 (1): 196–211.e11.

108 Schmidt, N., Lareau, C.A., Keshishian, H. et al. (2020). The SARS-CoV-2 RNA–protein interactome in infected human cells. *Nat. Microbiol.* 6 (3): 339–353.

109 West, J.A., Davis, C.P., Sunwoo, H. et al. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell* 55 (5): 791–802.

110 Beach, D.L. and Keene, J.D. (2008). Ribotrap: targeted purification of RNA-specific RNPs from cell lysates through immunoaffinity precipitation to identify regulatory proteins and RNAs. *Methods Mol. Biol.* 419: 69–91.

111 Ramanathan, M., Majzoub, K., Rao, D.S. et al. (2018). RNA–protein interaction detection in living cells. *Nat. Methods* 15 (3): 207–212.

112 Rogell, B., Fischer, B., Rettel, M. et al. (2017). Specific RNP capture with antisense LNA/DNA mixmers. *RNA* 23 (8): 1290–1302.

113 Wahlestedt, C., Salmi, P., Good, L. et al. (2000). Potent and nontoxic antisense oligonudeotides containing locked nucleic acids. *Proc. Natl. Acad. Sci. U.S.A.* 97 (10): 5633–5638.

114 Perez-Perri, J.I., Rogell, B., Schwarzl, T. et al. (2018). Discovery of RNA-binding proteins and characterization of their dynamic responses by enhanced RNA interactome capture. *Nat. Commun.* 9 (1): 4408.

115 Schildkraut, C. and Lifson, S. (1965). Dependence of the melting temperature of DNA on salt concentration. *Biopolymers* 3 (2): 195–208.

116 Gemmill, D., D'souza, S., Meier-Stephenson, V., and Patel, T.R. (2020). Current approaches for RNA-labelling to identify RNA-binding proteins. *Biochem. Cell Biol.* 98 (1): 31–41.

117 Leppek, K. and Stoecklin, G. (2014). An optimized streptavidin-binding RNA aptamer for purification of ribonucleoprotein complexes identifies novel ARE-binding proteins. *Nucleic Acids Res.* 42 (2): e13.

118 Shui, B., Ozer, A., Zipfel, W. et al. (2012). RNA aptamers that functionally interact with green fluorescent protein and its derivatives. *Nucleic Acids Res.* 40 (5): e39.

119 Panchapakesan, S.S.S., Ferguson, M.L., Hayden, E.J. et al. (2017). Ribonucleoprotein purification and characterization using RNA Mango. *RNA* 23 (10): 1592–1599.

120 Mukherjee, J., Hermesh, O., Eliscovich, C. et al. (2019). β-Actin mRNA interactome mapping by proximity biotinylation. *Proc. Natl. Acad. Sci. U.S.A.* 116 (26): 12863–12872.

121 Lee, F.C.Y. and Ule, J. (2018). Advances in CLIP technologies for studies of protein-RNA interactions. *Mol. Cell* 69 (3): 354–369.

122 Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A. et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* 13 (6): 508–514.

123 Buchbender, A., Mutter, H., Sutandy, F.X.R. et al. (2020). Improved library preparation with the new iCLIP2 protocol. *Methods* 178: 33–48.

124 McMahon, A.C., Rahman, R., Jin, H. et al. (2016). TRIBE: hijacking an RNA-editing enzyme to identify cell-specific targets of RNA-binding proteins. *Cell* 165 (3): 742–753.

125 Rahman, R., Xu, W., Jin, H., and Rosbash, M. (2018). Identification of RNA-binding protein targets with HyperTRIBE. *Nat. Protoc.* 13 (8): 1829–1849.

126 Fazal, F.M., Han, S., Parker, K.R. et al. (2019). Atlas of subcellular RNA localization revealed by APEX-Seq. *Cell* 178 (2): 473–490.e26.

127 Huppertz, I., Attig, J., D'Ambrogio, A. et al. (2014). iCLIP: protein-RNA interactions at nucleotide resolution. *Methods* 65 (3): 274–287.

128 Nguyen, D.T.T., Lu, Y., Chu, K.L. et al. (2020). HyperTRIBE uncovers increased MUSASHI-2 RNA binding activity and differential regulation in leukemic stem cells. *Nat. Commun.* 11 (1): 1–12.

129 Van Steensel, B., Delrow, J., and Henikoff, S. (2001). Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.* 27 (3): 304–308.

130 Qin, W., Cho, K.F., Cavanagh, P.E., and Ting, A.Y. (2021). Deciphering molecular interactions by proximity labeling. *Nat. Methods* 18 (2): 133–143.

131 Branon, T.C., Bosch, J.A., Sanchez, A.D. et al. (2018). Efficient proximity labeling in living cells and organisms with TurboID. *Nat. Biotechnol.* 36 (9): 880–898.

132 Rhee, H.W., Zou, P., Udeshi, N.D. et al. (2013). Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* 339 (6125): 1328–1331.

133 Kaewsapsak, P., Shechner, D.M., Mallard, W. et al. (2017). Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *eLife* 6: e29224.

134 Benhalevy, D., Anastasakis, D.G., and Hafner, M. (2018). Proximity-CLIP provides a snapshot of protein-occupied RNA elements in subcellular compartments. *Nat. Methods* 15 (12): 1074–1082.

13

Interaction Between Proteins and Biological Membranes

Lorant Janosi[1] *and Alemayehu A. Gorfe*[2,3]

[1] *National Institute for Research and Development of Isotopic and Molecular Technologies, Department of Molecular and Biomolecular Physics, 67-103 Donat Street, PO 5 Box 700, Cluj-Napoca, CJ 400293, Romania*
[2] *University of Texas Health Science Center, Department of Integrative Biology & Pharmacology, McGovern Medical School, 6431 Fannin St, Houston, TX 77030, USA*
[3] *MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, 6431 Fannin St, Houston, TX 77030, USA*

13.1 Introduction

The survival of every cell is dependent upon the maintenance of membrane homeostasis, an actively controlled process of regulating membrane composition, structure, and dynamics. The cell membrane is the boundary that separates a cell from its surroundings and shields the contents of internal organelles. Specifically, the plasma membrane is a semipermeable protective envelope of cells and plays critical role in the regulated movement of material between the interior of the cell and its surroundings. Similarly, internal or endomembranes compartmentalize intracellular cell components into various organelles, such as the Golgi apparatus, endoplasmic reticulum, and the nucleus. Membranes also serve as vesicles to transport cargo across the cytosol, as well as into and out of the cell, as two-dimensional structural frameworks to host and organize proteins involved in substrate transport, signaling, and other functions. The focus of this chapter is on the fundamental interactions responsible for the targeting and stability of these proteins in membranes.

The topic of the interactions of proteins with biological membranes is too broad to cover in full in a single chapter. Therefore, we will focus on two major classes of proteins, peripheral and integral membrane proteins, and describe in general terms the most common mechanisms of targeting, stabilizing, and reorganizing these proteins inside or on the surface of membranes. Moreover, most of our discussion will focus on the plasma membrane, with only occasional references to endomembranes. This is because, while the plasma membrane and endomembranes differ in lipid composition and protein content, the fundamental interactions relevant to the structural integrity and function of membranes and membrane proteins are conserved.

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

The chapter is organized as follows. First, an overview of the structure, composition, and function of the plasma membrane is provided, followed by a short review of its structural heterogeneity and the key factors responsible for its dynamics. Then, a somewhat detailed description of peripheral membrane proteins is provided, with emphasis on the protein-based and lipid-based targeting motifs that are responsible for their interaction with membrane. Turning to the larger family of transmembrane proteins, some examples are used to discuss the key roles of membrane and membrane lipids for the structural integrity and conformational changes that drive function in both single-pass and multi-pass transmembrane proteins that are most common in the plasma membrane of eukaryotic cells. This is followed by a brief discussion of antimicrobial peptides as an example of proteins that are conditionally peripheral or transmembrane and closes with a summary of the main points discussed in the chapter.

13.2 The Plasma Membrane: Overview of Its Structure, Composition, and Function

The primary constituents of the plasma membrane are lipids. While there are thousands of lipid species in every cell, the most abundant lipids in the plasma membrane are phospholipids. When mixed with water, phospholipids spontaneously assemble into bilayers due to the opposing forces from their "water-loving" (hydrophilic) head group and "water-hating" (hydrophobic) tails. The resulting nonbonded phospholipid–phospholipid interactions in bilayers are weak, making membranes fluid-like (i.e. they have structural integrity while being flexible). Beyond their structural role, phospholipids in the plasma membrane are the "pickets" that form a "fence" around the cell, with the hydrophobic interior (core) of the membrane forming a barrier for water-soluble substances to maintain different mixtures and concentrations of solutes inside and outside of the cell.

Other major constituents of the plasma membrane are cholesterol, carbohydrates, and proteins. Cholesterol tucks in-between phospholipid molecules and contributes to both the fluidity and stability of the plasma membrane, while carbohydrates, which serve as self-identification markers of cells, "sugar coat" the outer surface by attaching to proteins and lipids. The plasma membrane is also rich in proteins. In fact, while lipids are about 50 times more numerous, the more massive proteins account for nearly half of the total mass of the plasma membrane. Cells differ in the number, distribution, and activity of the proteins they possess, and proteins are continuously recycled into and from the plasma in a controlled manner. There are two types of membrane proteins: integral or transmembrane proteins (TMPs) and surface-bound or peripheral membrane proteins (PMPs).

As the name suggests, TMPs span the entire thickness of the host membrane and perform many diverse functions. Some serve as highly selective water-filled passageways or channels that transport specific ions across the plasma membrane (e.g. Na^+ and K^+ channels). TMPs also serve as: (i) carriers to transport specific substances that cannot cross the membrane core on their own (e.g. iodine is transported from

blood into thyroid gland cells by carrier proteins); (ii) receptors to recognize and bind specific molecules in the cell's environment (such as hormones in the blood stream); (iii) cell-adhesion molecules to link the extracellular surroundings of the cell with the intracellular cytoskeleton (e.g. integrins); (iv) some TMPs protrude from the cell surface and form loops or hooks by which cells grip each other (e.g. cadherin zippers holding cells between tissues and organs). Unlike TMPs, peripheral membrane proteins are attached to one side of the plasma membrane. However, just like TMPs, PMPs have various important functions, including roles as enzymes, scaffolding proteins, and regulators of cell signaling events. In fact, many kinases, phosphatases, and GTPases are PMPs. Other PMPs serve as docking markers for secretory vesicles.

13.3 Lipid-Based and Protein-Based Sorting of Plasma Membrane Components

The plasma membrane is highly dynamic, capable of rapidly reorganizing to form local substructures in response to generalized or local changes in the interior content or external environment of the cell. Membrane substructures range from nanometer to micrometer in size and function as transient signaling platforms or more permanent immunological synapses. At the fundamental level, membrane structural heterogeneity arises from variations in lipid composition and content or activity of specific membrane proteins. Therefore, despite the complexity of the plasma membrane, basic properties of its substructures or domains can be studied by using much simpler model membranes made up of, for example, specific phospholipids, sterols, and selected membrane proteins. Based on such studies, two broad mechanisms of domain formation have been described: lipid-based and protein-based sorting. Below, the molecular interactions that underlie these sorting mechanisms, operating either independently or in tandem, are briefly reviewed.

13.3.1 Lipid-Based Sorting and Domain Formation

A fundamental property of biological membranes is phase separation. Phase transition is a process of lipid sorting that occurs at temperatures characteristic of the constituent lipids and allows for the formation of co-existing membrane domains (or substructures) characterized by distinct biophysical properties. The most common lipid phases are gel or solid ordered (So), liquid disordered (Ld), and liquid ordered (Lo). In the So phase, lipids are tightly packed and have reduced lateral mobility, whereas, in the Ld phase, lipids are loosely packed and have faster lateral mobility. The Lo phase shares tight lipid packing with So and high lipid diffusion with Ld. Two or more of these phases can co-exist in bilayers of mixed lipids. For instance, in model membranes composed of sphingomyelins, unsaturated phosphatidylcholine (PC), and 10–30% cholesterol, Lo and Ld phases co-exist with the Lo phase enriched in cholesterol and sphingomyelin, and the Ld phase enriched in the unsaturated PC. The molecular basis of this process can be

studied using a mixture of saturated PC and cholesterol. The total area occupied by this mixture is smaller than would be expected from the sum of the constituents because cholesterol "condenses" the PCs. The condensing effect of cholesterol is related to its spontaneous negative curvature [1] as well as its reversible interactions with PC to form condensed complexes of defined stoichiometry [2]. In such complexes, PCs pack more tightly to shield the hydrophobic cholesterol from the hydrophilic membrane–water interface [3]. Molecular dynamics (MD) simulations of cholesterol/sphingomyelin/phosphatidylcholine mixtures [4–6] have shown that cholesterol localizes at the interface between sphingomyelin-enriched and PC-enriched regions [6], with the saturated sphingomyelin acyl chains packing against the smooth α-face of cholesterol, while the disordered acyl chains of PCs pack more easily against the methylated and hence rougher β-face of the cholesterol.

13.3.2 Protein-Based Sorting and Membrane Curvature

Many plasma membrane proteins partition to nanoscopic lipid domains called lipid rafts, roughly defined as Lo domains surrounded by Ld domains [7–9]. Characterized by unique biophysical and thermodynamic properties, lipid rafts recruit specific lipids and proteins to facilitate cell signaling and other functions [9]. Along with palmitoylation (discussed later), the hydrophobicity and size of transmembrane domains in TMPs are key determinants of raft affinity [10, 11]. Also, many PMPs have an intrinsic preference for specific membrane substructures or cause curvature via one or more of three mechanisms: (i) mechanical bending: caused by a protein or network of proteins with a rigid tertiary structure and a curved surface; (ii) local deformation: caused by proteins that embed amphipathic helices into a monolayer; and (iii) area difference between leaflets: caused by protein domains that penetrate only one leaflet.

An example of mechanical bending is that caused by the Bin-Amphiphysin-Rvs (BAR) domains that arrange on the membrane surface in various ways to generate distinct membrane curvatures. Mechanical bending is often complemented by interactions of basic amino acids with membrane lipids, or by the action of an amphipathic helix causing local deformation of a monolayer (mechanism ii). Amphipathic helices tend to lie on the membrane surface with the hydrophobic face partitioning into the membrane core and the polar face interacting with lipid head groups and solvent, thus causing a local disruption of lipid packing. For example, in proteins containing a particular BAR domain, N-BAR, sensing of membrane curvature involves a dimeric BAR domain plus a disordered N-terminus from each protomer forming an amphipathic helix upon binding to membranes containing negatively charged phospholipids [12]. Finally, membrane curvature and domain formation through inter-leaflet surface area difference can be exemplified by the action of caveolins, a family of proteins that insert helical hairpins into the inner leaflet and generate nanometer-sized plasma membrane pits [13].

13.3.3 Proteolipid Sorting and Membrane Domain Stabilization

Domain formation in biological membranes, including the plasma membrane, typically involves a combination of both lipid-based and protein-based sorting processes. An illustrative example of such proteolipid sorting is provided by Ras proteins, a group of post-translationally lipid-modified signaling proteins that form nanometer-sized protein–lipid clusters on the plasma membrane ([14] and references therein). MD simulations of Ras tethered to various lipid bilayer models provided insights into the physical basis for the clustering and nonoverlapping distribution of Ras isoforms in membrane domains [15–17]; it was found that the nature of the lipid modification dictates lateral organization of Ras proteins. Specifically, simulations of the dually palmitoylated and farnesylated lipid anchor of H-Ras in a bilayer made up of dipalmitoylphosphatidylcholine (DPPC), dilinoleoylphosphatidylcholine (DLiPC), and cholesterol [15] showed that the peptides spontaneously assemble into clusters of 4–10 molecules, and clustering leads to segregation of the peptides to the boundary between the Lo and Ld domains. Removal of the palmitoyl resulted in accumulation at the Ld phase, while removal of the farnesyl group segregated the peptides to the Lo phase. Thus, lateral segregation is primarily driven by the differential affinity of the saturated palmitoyl chain for the saturated DPPC lipids and of the polyunsaturated farnesyl for the unsaturated DLiPC lipids, while cholesterol modulates lipid domain stability and thereby Ras nanocluster stability. Importantly, the asymmetric incorporation and aggregation of Ras curves multi-domain membranes by expanding the surface area of the host monolayer [18–21], with the Ras lipid anchors acting as linactants to reduce the line tension at the domain boundary [21].

While these studies explained how lipid modification modulates domain preference, other studies showed that protein sequence and structure also contribute to the process. As an example, using a combination of experiments and simulations on another Ras isoform, K-Ras, it was found that the identity of the side chains and the structure of the intrinsically flexible polybasic domain, together, dictate preferences for specific membrane lipids, clustering, and lateral dynamics of K-Ras [22]. As discussed later, other lipid-modified proteins also utilize a combination of a lipophilic motif and an amphipathic helix to bind membranes, which, as in Ras, can cause curvature via both lipid-based and protein-based sorting mechanisms.

13.4 Interaction of Peripheral Membrane Proteins with Membrane Lipids

There are two major classes of signals that target PMPs to the surface of membranes: protein-based modular domains (Figure 13.1) and lipid-based motifs (Figure 13.2). High-affinity membrane binding of PMPs often requires the use of more than one of these motifs or complementation by a polybasic domain or an amphipathic helix. The most important atomic interactions between each class of these motifs and membrane lipids are described below using specific examples

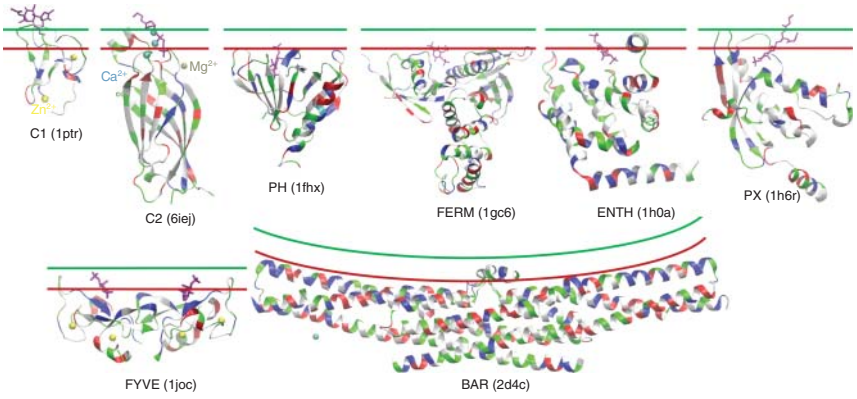


Figure 13.1 Modular protein-based membrane-targeting motifs found in many peripheral membrane proteins. Left to right from top, C1 (conserved homology-1), C2, PH (pleckstrin homology), FERM (4.1/Ezrin/Radixin/Moesin), ENTH (Epsin N-terminal homology), PX (Phox homology), FYVE (Fab1/YOTB/Vac1/EEA1), and BAR (Bin/Amphiphysin/Rvs) domain structures from the protein data bank (PDB; ID in bracket). The secondary structure is shown in ribbon with basic residues in blue, acidic in red, polar in green and hydrophobic residues in white. For each domain except BAR, a bound lipidic ligand is shown in purple stick representation: phorbol acetate in C1, phosphatidylcholine in C2, inositol-1,3,4,5-tetrakisphosphate in PH, inositol 1,4,5-trisphosphate or IP3 in FERM and ENTH, phosphatidylinositol 3-phosphate or PI3P in PX, and inositol-1,3-phosphate in FYVE (one on each monomer with the dimerization helices omitted). The approximate location of each domain inserted into one leaflet of an idealized membrane is shown schematically using red line (representing the approximate position of lipid phosphate groups) and green line (the topmost location of the hydrocarbon core just beneath the carbonyl oxygens). Metal ions are shown as balls with zinc in yellow, calcium in light blue, and magnesium in dark green.

for illustration. Where relevant, the additional signals used to increase affinity are also noted.

13.4.1 Protein-Based Membrane-Targeting Motifs

Many PMPs contain one or more C1, C2, PH, FERM, ENTH, PX, FYVE, or BAR domains to dock onto the surface of the plasma membrane, typically at the cytosolic side. Studies using various techniques, such as X-ray crystallography and solution or solid-state NMR, typically focusing on the smallest autonomous domain embedded in a model membrane or micelle, have provided useful insights into how PMPs are targeted to membranes [24–27]. These insights include measures of the membrane insertion depth (how deep a protein penetrates the hydrocarbon core of a bilayer), angle of insertion (the angle between the principal axis of a protein and the bilayer normal), and, in favorable cases, the atomic interactions with lipids. These data, coupled with modeling and molecular simulations [27, 28], showed that membrane targeting by modular protein domains involves various combinations of few structural and sequence features (e.g. [29]): (i) a polar pocket or groove to specifically target a ligand; (ii) a cluster (or clusters) of arginine and lysine residues to interact with phospholipid head groups; and (iii) a cluster or clusters of hydrophobic surface

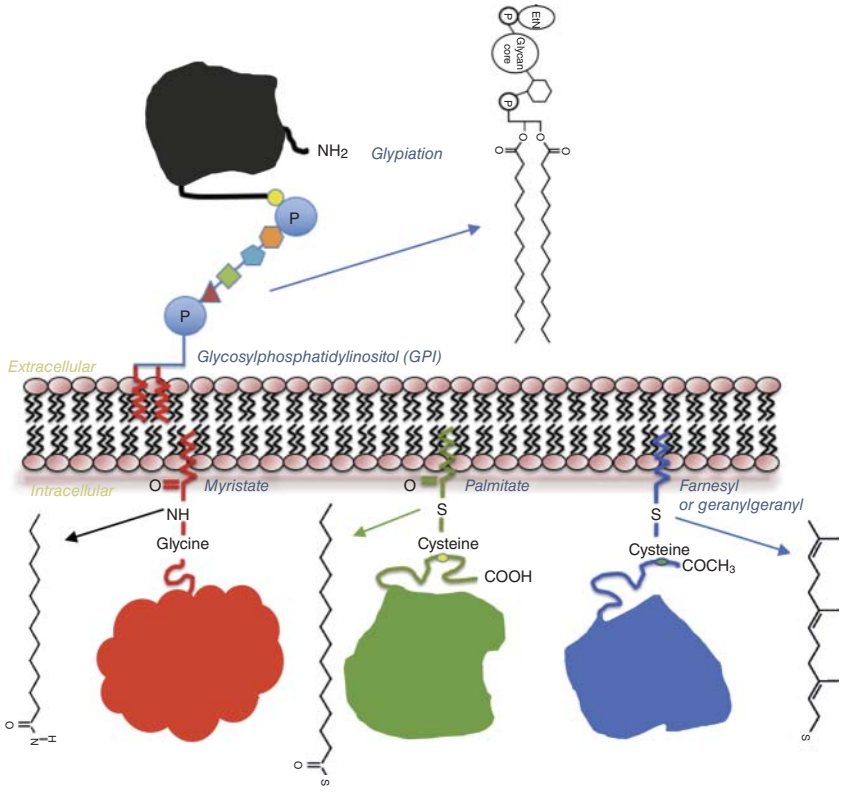


Figure 13.2 Lipid-based membrane-targeting motifs. Schematics illustrate the targeting of proteins to the cytosolic and extracellular leaflets of the plasma membrane by the four most common lipidation motifs in humans. Also shown are chemical structures of the fatty acids that modify a C-terminal cysteine (S-prenylation), an N-terminal glycine (N-myristoylation), and a nonterminal cysteine (S-palmitoylation) residue on proteins at the inner surface of the plasma membrane, as well as a glycophosphoinositol (GPI) modification of proteins at the extracellular side**.** Source: Figure adapted from Lobo [23].

residues to penetrate the hydrocarbon core of the membrane to provide an anchor for the protein.

An example of a domain with all three of these features is C1. The conserved homology-1 (C1) domain recognizes a phorbol ester or a diacylglycerol ligand via a zinc finger possessing a narrow polar groove, interacts with negatively charged lipid head groups via a belt of basic residues, and two hydrophobic loops penetrate into the hydrocarbon core of the membrane (Figure 13.1). Similarly, a combination of electrostatic and shape complementarities is critical for lipid recognition and high-affinity membrane binding by proteins containing C2 domains. Typical C2 domains are characterized by an eight-stranded beta-sandwich structure (Figure 13.1) and are frequently found in proteins with enzymatic functions, including phospholipases and phosphatases. Most, but not all, C2 domain families have two or more calcium-binding sites near the ligand recognition site (see Figure 13.1).

While most other protein-based targeting motifs are selective for closely related ligands, C2 domains are promiscuous and bind to most of the major membrane components, including phosphatidylserine (PS) and phosphatidylcholine (PC). As for C1, proteins containing C2 domains can target acidic phospholipids in membranes. This may be, at least in part, due to the positive electrostatic surface generated by the calcium ions. Therefore, proteins containing C2 domains can be targeted to both the PS-enriched inner surface of the plasma membrane and a variety of endomembranes involved in cellular traffic.

Pleckstrin homology (PH) domains are used for membrane binding by a diverse group of proteins involved in lipid signaling. As shown in Figure 13.1, typical PH domains lack the large hydrophobic protrusions found in C1 domains, but they have a deep pocket that binds phosphoinositides (PIs) with variable affinity and specificity. Since electrostatic interactions are the key determinants of ligand recognition, the preference and strength of PH domains for ligand binding depends on the distribution of basic residues that populate loop regions surrounding the binding pocket.

Similar to PH domains, FERM (from 4.1 protein, ezrin, radixin, and moesin) domains lack a hydrophobic protrusion (Figure 13.1). They are commonly found in proteins that link the cytoskeleton with the plasma membrane by binding to the cytoplasmic regions of transmembrane proteins. FERM consists of three compact modules: A, B, and C. It has been shown that FERM domains bind phosphatidylinositol 4,5-bisphosphate (PIP2) and inositol-1,4,5-trisphosphate (IP3) at the C module and between the A and C modules, respectively. Both of these modules contain basic surface residues that interact with the phosphates of the PIs. PI binding causes conformational changes that allow binding of FERM-containing proteins to the cytosolic regions of integral membrane proteins.

The endomembrane-localized ENTH, PX, and FYVE domains utilize the same general principles described above to recognize various endosomal lipids, particularly phosphatidylinositol-3-phosphates (PI3Ps). Epsin N-terminal homology (ENTH) domains are commonly found at the N-terminus of endocytic proteins. The ENTH domain has nine alpha-helices, with three stacked helical hairpins and a flexible N-terminal helix that together form a groove for binding to PI3P ligands (Figure 13.1). As in N-BAR, the intrinsically disordered N-terminus of ENTH forms a helix upon membrane binding, which is achieved by inserting the hydrophobic face of the amphipathic helix into the membrane (Figure 13.1). This interaction anchors the protein and may promote positive membrane curvature through monolayer area expansion.

PX domains bind PIs in a deep pocket, interact with negatively charged lipid head groups via a basic surface surrounding the pocket, and insert into the membrane with proximal hydrophobic protrusions (Figure 13.1). The FYVE domain has a shallow basic pocket that specifically recognizes PI3P, plus additional basic residues for nonspecific interaction with acidic phospholipid head groups, as well as a hydrophobic protrusion that penetrates the membrane. As shown in Figure 13.1, FYVE can form a homodimer with each monomer binding one ligand.

BAR domains do not have a pocket for binding to specific membrane lipids, and therefore, many BAR domain-containing proteins contain additional motifs for target specificity. BAR domains form a banana-shaped helical dimer to sense and bind to curved membranes (Figure 13.1). The concave face of the dimer is enriched with Lys and Arg residues whose nonspecific electrostatic interactions with negatively charged phospholipids enable BAR domains to target membranes enriched with PS. As noted in the previous section, induction of membrane curvature by BAR domains requires an N-terminal disordered region forming an amphipathic helix that inserts into the membrane core.

13.4.2 Lipid-Based Membrane-Targeting Motifs

A large number of PMPs (and some TMPs) with diverse structures and functions undergo covalent modification by one or more lipid-based motifs for regulated binding to the plasma membrane [30]. The most important of these lipid-based targeting motifs are S-prenylation (addition of a farnesyl (15-carbon) or geranylgeranyl (20-carbon) unsaturated branched fatty acid to one or two C-terminal Cys residues), S-palmitoylation (addition of a palmitoyl (16-carbon) saturated fatty acid to a Cys residue), N-myristoylation (addition of a 14-carbon saturated fatty acid (myristate) at an N-terminal Gly), and glypiation (addition of a glycosylphosphatidylinositol (GPI) group with various number and saturation level of lipid tails) (Figure 13.2). Most lipidated PMPs are involved in a variety of signal transduction pathways and include kinases (e.g. Src family and AKAPs), G-proteins (e.g. Ras family and Gα subunits), and cell surface receptors (such as G-protein-coupled receptors (GPCRs) and the transferrin receptor). Key to the activity of these proteins is proper localization to specific regions of the plasma membrane or endomembranes. This is achieved by the interaction of each lipid-based targeting motif with distinct membrane lipids. Perhaps the best-studied example in this context is the Ras family of small GTPases.

Ras proteins function as molecular switches controlling a wide variety of signal transduction pathways, including the mitogen-activated protein kinase (MAPK) and the PI3K/AKT/mTOR (phosphatidyl-inositol 3-kinase/protein kinase B/mammalian target of rapamycin) pathways. Malfunction of Ras due to mutation was documented for ∼20% of all human cancers [31]. Ras proteins are tethered to the inner leaflet of the plasma membrane by a farnesylated and carboxy-methylated C-terminus plus a proximal palmitoyl (N-Ras and H-Ras) or polybasic domain (K-Ras). These prenyl/palmitoyl or prenyl/polybasic domain combinations, termed lipid anchors, are the fundamental determinants of Ras-membrane interactions. The first molecular model showing the organization of a Ras lipid anchor on a membrane emerged from a combined FTIR, ssNMR, and neutron diffraction study on a palmitoylated and hexadecylated heptapeptide representing the C-terminus of N-Ras [32]. Although it lacked atomic resolution, the study provided insights into how deep the lipidated residues penetrate the model DMPC (dimyristoylphosphatidylcholine) bilayer used in the study, and how the backbone localizes at the bilayer-water interface. A subsequent MD simulation study of the same system provided the missing details [33]. Insertion of about five terminal carbon atoms of

the palmitoyl moiety at Cys181 or the farnesyl at Cys186 into the core of the DMPC bilayer was found to be sufficient to cross the barrier at the bilayer-water interface, allowing for the rest of the anchor to spontaneously partition into the bilayer. The resulting vdW interactions between the hydrophobic lipid tails of the peptide and the membrane lipids are the major driving force for the binding. Hydrogen bonds between the peptide backbone and lipid head groups provide additional stabilization [33]. Similarly, in H-Ras, the farnesylated (Cys186) and palmitoylated (Cys181 and Cys184) cysteines plus Met182 are engaged in vdW interactions with lipid acyl chains, while the peptide backbone and polar side chains form hydrogen bonds with lipid head groups [34]. Additional MD simulations [35] and potential of mean force calculations showed that the two palmitoyl modifications of H-Ras do not contribute equally to the free energy of membrane insertion [36], and that insertion is dominated by an enthalpy-driven hydrophobic effect [37]. In the case of the K-Ras lipid anchor, the insertion of the farnesyl tail and electrostatic interactions of six proximal lysine residues with anionic lipids stably tether it to the membrane [22, 38, 39].

Essentially the same interplay between vdW and electrostatic/polar interactions underlies membrane targeting of other lipid-modified proteins. This includes, with few exceptions, the Ras, Rho, and Rab family of proteins in the Ras superfamily. All of these proteins are C-terminally prenylated (i.e. undergo one or more farnesyl or geranylgeranyl modifications) and harbor a polybasic domain or a palmitoylatable Cys near the site of prenylation. For example, the Rho family protein RhoA is geranylgeranylated at Cys190 and contains a proximal polybasic domain, whereas the Rab family protein Rab11A is dually geranylgeranylated at Cys212 and Cys213. As a result, RhoA interacts with membranes (mostly the plasma membrane but also endomembranes) through a combination of vdW and electrostatic interactions similar to K-Ras. Membrane interaction of Rab11A (typically in recycling endosomes), on the other hand, is dominated by hydrophobic interactions, as in H-Ras.

Other examples of regulatable membrane binding by lipid modification include the myristoylated alanine-rich C-kinase substrate (MARCKS) [40, 41] and the proto-oncogene tyrosine-protein kinase Src (c-Src) [42, 43]. In both of these proteins, a glycine-myristoyl moiety at the N-terminus, complemented by a polybasic region, enables targeting and tight binding to the plasma membrane. As in K-Ras, the driving force for the binding involves a combination of hydrophobic interactions between the myristoyl chain and lipid acyl chains and electrostatic interactions between the polybasic domain and anionic lipid head groups. K-Ras and MARCKS can get phosphorylated at residues close to the polybasic domain, with the negative charge on the phosphate reducing the strength of interaction with membranes. In some cases, membrane binding via lipid modification is accompanied by the insertion of an amphipathic helix into the core of the membrane (as discussed above, an amphipathic helix is a common membrane-targeting motif in peripheral proteins). An example is the Arf family of proteins, which interact with membranes via a myristoylated N-terminus and a proximal amphipathic helix [44–46].

While prenylation, myristoylation, and palmitoylation are used to target proteins to the cytosolic side, the GPI anchor localizes glycoproteins to the extracellular side of the plasma membrane [30]. Modification of a protein with a GPI that harbors two or three long fatty acyl chains enables tight binding of proteins to lipid bilayers due to the extensive vdW interactions between the acyl chains of the GPI and membrane lipids. The fatty acyl chains in most mammalian GPI anchors are saturated, which, driven primarily by lipid sorting, segregate GPI-anchored proteins to ordered membrane domains (or lipid rafts) that are enriched with saturated lipids. In short, the membrane affinity and lateral organization of GPI-anchored proteins are largely determined by the number, length, and saturation level of the hydrocarbon tails of the GPI anchor.

It is clear from the foregoing discussion that membrane association of lipid-modified proteins is primarily determined by the lipid anchor. However, recent reports suggest that the soluble domains of lipidated proteins also engage membranes, although those interactions are typically short-lived [39, 47, 48].

13.5 Interactions and Conformations of Transmembrane Proteins in Lipid Membranes

As already introduced, TMPs are integral membrane proteins that sit across the height of membranes. The structure of the transmembrane (TM) region is typically composed of single or multiple TM helices or a bundle of beta-sheets. The TM surface that is in contact with the core of the membrane is largely hydrophobic (nonpolar), whereas the TM residues facing away from lipids and into the protein core can be neutral polar or charged. The intra- and extra-cellular regions of TMPs are mainly hydrophilic. Usually, the hydrophobic environment of the biological membrane restrains the tertiary structure of TMPs to a relatively small number of stable conformations. These conformations can be altered by mutation, changes in the membrane environment, or interaction with other proteins or ligands. In fact, many TMPs undergo conformational changes or association reactions within the lipid matrix to drive various biochemical processes, including signal transduction [49, 50]. The following subsections discuss the membrane interactions and attendant conformational changes of two categories of helical TMPs that are most common in the plasma membrane of eukaryotic cells: single-pass and multi-pass TMPs.

13.5.1 Glycophorin A and EGFR as Examples of Single-Pass Transmembrane Proteins

Glycophorin A (GpA) is a widely studied seven amino-acid pattern (LIxxGVxxGVxxT) single TM domain (see Figure 13.3) that is extensively used as a model for experimental, theoretical, and simulation studies of the association of TM proteins within membranes. Because it folds into a single-pass TM helix, it is highly hydrophobic, and therefore tilting is the main mechanism to compensate for hydrophobic

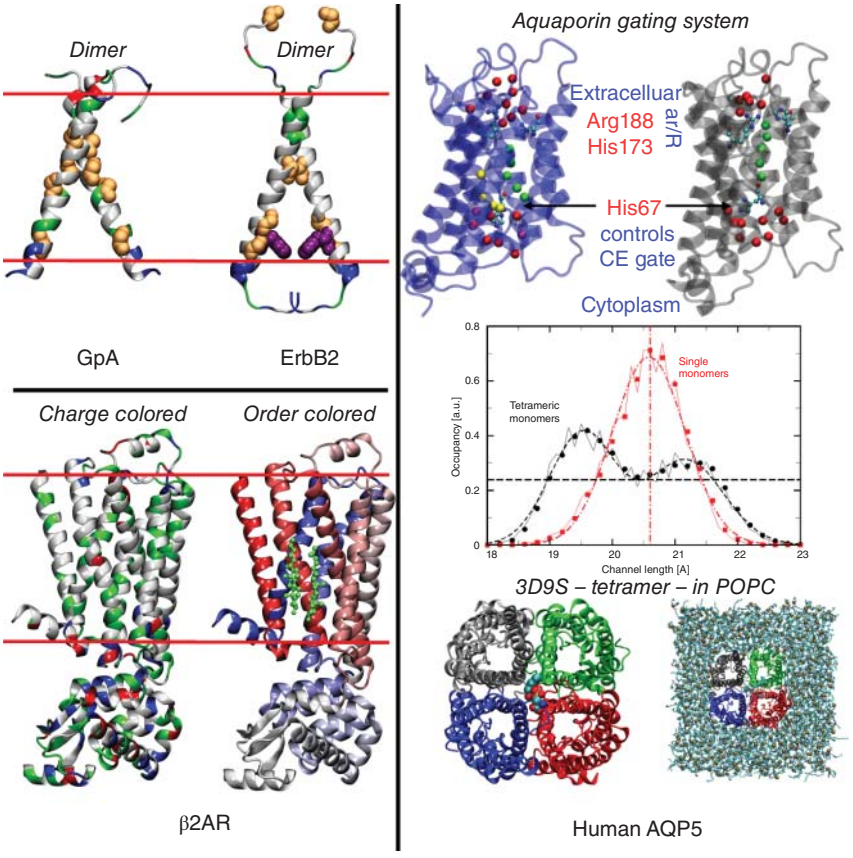


Figure 13.3 Examples of TMPs. (Left Top) Glycophorin A (GpA) and **Epidermal Growth Factor Receptor ErbB2** homodimer crystal structures color-coded by electrostatic charge (red-negative, blue-positive, green-polar, white-hydrophobic) are shown in a cartoon representation. Gly of the GxxxG motifs and Phe residues are shown in ball representation colored in orange and violet. **(Left Bottom) Beta-2 Adrenergic Receptor (β2AR)** in cartoon representations colored by charge (see above), to highlight the polarity of residues, and order (N-terminus in red, C-terminus in blue), to emphasize the monomeric structure of **β**2AR. Two crystallographic binding sites of cholesterol are shown in green ball-and-stick representation. In these figures, the average position of the membrane phosphate group is marked by red lines. **(Right) Human Aquaporin 5** system. (Top): Open (Left) and closed (Right) states are controlled by His67 at the cytoplasmic end (CE); Aromatic ring and Arg (ar/R) motifs control the water-flow rate. (Middle): Channel length distribution for single (red) and tetrameric (black) monomers shows a significant increase in pore flexibility when tetrameric assembly allows for gating mechanisms due to conformational switch (see text). (Bottom Left) Tetrameric crystal structure 3D9S of human AQP5 with cartoon-represented monomers in different colors and POPS central lipid in vdW representation. (Bottom Right) Equilibrated tetramer solvated in POPC bilayer (cyan stick lipids, P atoms in tan). Source: Janosi and Ceccarelli [51]/Public Library of Science/CC BY Attribution.

mismatch. The three Gly residues in the TM region do not have the role of helix breakers, as usually found in hydrophilic environments. Instead, they are part of a crucial sequence-specific (GxxxG) motif that stabilizes the GpA homodimers [52, 53]. While the dimer interface is hydrophobic and dimerization itself is suggested to be driven by van der Waals interactions [54], the membrane environment plays a key role in modulating the GpA TM domains. Specifically, it has been shown that membrane thickness, composition, and lipid acyl chain ordering modulate the range of accessible dimer conformations [55–57]. This dimerization mechanism is applicable to other single-pass TM proteins harboring the GxxxG dimerization motif as well, such as the epidermal growth factor receptors (EGFRs), which are discussed below.

EGFR represents a family of four closely related receptor tyrosine kinases that include the ErbB subfamily: ErbB1 (also called Her1), ErbB2 (Her2/Neu), ErbB3 (Her3), and ErbB4 (Her4) [58]. Overexpression of ErbBs leads to many types of cancers, while insufficient ErbB signaling leads to neurodegenerative diseases, such as multiple sclerosis and Alzheimer's diseases [59, 60]. ErbBs are activated by growth factors that bind to the extracellular region [61], which induces dimerization in the TM region, as shown in Figure 13.3 [62], as well as trans-phosphorylation of tyrosine residues at the intracellular kinase domain [63–66]. This leads to the activation of a wide variety of downstream partners [63] in the MAPK and PI3K/Akt pathways, leading to cell proliferation and division. Deregulation of ErbB and other EGFR signaling due to mutation or overexpression can therefore cause cancer [58]. As GpA, EGFRs are single-pass TM proteins characterized by mostly hydrophobic residues with the exception of a few polar ones. However, the TM helix of EGFR contains two GxxxG motifs that, together, form "pockets" for cholesterol binding. The side-chain entropy of a conserved C-terminal Phe residue and its orientation perpendicular to the membrane normal can lead to cholesterol depletion around the GxxxG motif near the C-terminus [67]. Other important protein–membrane interactions in this class of tyrosine kinases include the interactions of the intracellular domain and the juxtamembrane segment with negatively charged lipids of the plasma membrane [68]. These membrane–protein interactions are built upon key protein features (such as the GxxxG motifs), and are modulated by specific membrane features, such as cholesterol concentration. Due to their role in heterodimerization, these interactions are crucial for EGFR signaling and are addressed below. Therefore, understanding the interactions of the TM helices with each other and with the membrane environment is essential to understanding signal transduction via ErbBs. Heterodimerization of ErbBs is driven by protein–protein interactions that depend on the tilt angle of the helices, which in turn are modulated by enthalpic and entropic factors associated with lipid–protein interactions and lipid dynamics. Specifically, displacement of lipids between the helices and subsequent protein–protein interactions are the driving forces for ErbB TM helix association. Lipids tend to favor parallel inter-helical interactions, and such a face-to-face interaction is favored by small side chains [56, 69–71]. The dimerization pathway is also a function of cholesterol concentration in the membrane. In the absence of cholesterol, dimerization proceeds via the hydrophobic

Phe residue at the C-terminus. In the presence of an intermediate (20%) cholesterol concentration, recognition occurs through contacts at the N-terminus [67]. This example illustrates how protein–lipid interactions modulate functionally important conformational changes in addition to their crucial role in the structural integrity and stability of TMPs.

13.5.2 GPCR as an Example of Multi-Pass TM Helical Proteins

G-protein-coupled receptors (GPCRs) are the largest family of human membrane proteins. The 7-TM helix proteins are also the largest target (∼35%) of drugs on the market [72, 73]. GPCRs are located in phospholipid-rich bilayers and are involved in signal transduction across the plasma membrane by changing their conformation from an inactive (antagonist bound) to an active (agonist and G protein-bound) state. The interactions of GPCRs with various lipidic components of the bilayer play an important role in the stability and dynamics of these conformational states. Moreover, binding of cholesterol can affect oligomerization and helical stability and has an allosteric effect on ligand binding [74, 75]. Multiple cholesterol binding sites have been identified by both crystallography (see Figure 13.3) and long-timescale MD simulations [75]. Phospholipids can modulate the association of GPCRs with G-proteins or their binding to specific ligands. For example, for beta 2 adrenergic receptors (β2AR – illustrated in Figure 13.3), phosphatidylglycerol (PG) lipids favor agonist binding, hence activation, while phosphatidylethanolamine (PE) or even PC favors antagonist binding, hence the inactive state. Specifically, the negatively charged PG lipids interact electrostatically with the positively charged residues at the intracellular loop 3 and the intracellular end of TM6, preventing the formation of the so-called "ionic lock" that stabilizes the inactive state of the receptor and stabilize TM6 in the active state of β2AR [76–78]. Similarly, PIP2 favors the active state of adenosine A_{2A} receptor by stabilizing the outward movement of TM6 by binding between TM6 and TM7 [79]. These main features of activation should, in general, be applicable to the entire class of GPCRs [75].

13.5.3 Aquaporin as an Example of Oligomeric Multi-Pass TM Proteins

Oligomeric TMPs perform various functional roles. These include roles as ion and water channels, mechanosensitive channels, transporters, and light-harvesting rings. These proteins are functional as homo- or hetero-oligomers ranging from tetramers to 9/10-mers or more, depending on their function. For some of these TMPs, it has been shown that specific lipids may play a role in either oligomerization and stability or more importantly, modulating their function. A good example of oligomeric TMPs are aquaporins, a group of transmembrane proteins that passively transport water across the membrane-driven response to osmotic gradient [80]. It was revealed that the regulation of these (usually) tetrameric proteins is likely gated by conformational changes (conditioned by the oligomerization state) of specific amino acids that act as on/off "switches" on one side, and water flow "regulators" on the other side [51] (see Figure 13.3). Nonetheless, little is known about

the interaction of the aquaporin outer surface and the host membrane. Studies showed that lipids in the middle of the oligomers usually stabilize the system while interactions at the surface influence function. Hence, strong hydrophobic interactions with the acyl chains of the annular lipids rather than the weaker hydrogen-bonding interaction with head groups are the driving force for lipid positioning near aquaporin [81]. Simulations suggest that aquaporins have little specificity for annular lipids (i.e. local environment) [82].

13.5.4 Antimicrobial Peptides: Peripheral or Integral?

Natural antimicrobial peptides (AMPs) are 10–50 residue-long peptides that function as a type of host defense mechanism by selectively interacting and destroying bacterial but not mammalian cell membranes. As in most protein-based targeting motifs, AMPs employ two important features of their amino acid composition to selectively target bacterial membranes: a large proportion (30–50%) of hydrophobic amino acids and positively charged amino acids (usually in the range +3e to +7e). As is typically the case, the hydrophobic residues enable high-affinity membrane binding while the basic residues allow for preferential binding to anionic bacterial membranes. The secondary structure of AMPs varies from helical (e.g. magainin), ß-strand (e.g. defensin), mixture of helices and strands (e.g. protegrin-1), to extended (e.g. indolicidin). Most AMPs undergo disorder-to-order conformational transition upon membrane binding, as occurs in the unstructured segments of many modular membrane-targeting motifs such as N-BAR and ENTH domains.

Despite the substantial sequence and structural diversity of AMPs, some general conclusions could be drawn from many years of studies regarding their binding to and effect on bacterial membranes. First, Arg and Lys residues ensure the initial engagement with the bacterial outer membrane. In the next phase, the hydrophobic residues facilitate passage through the membrane. However, this process is complicated by the complex composition of bacterial external membranes, including the lipopolysaccharides enriched with Mg^{2+} and Ca^{2+} ions in the outer membrane of Gram-negative bacteria, and the thick layer of negatively charged peptidoglycans in Gram-positive bacteria [83]. It is not fully clear how, after crossing the outer membrane, the peptides reach the cytoplasmic membrane, but concentration gradient and solvent effects likely explain the entire process. AMPs are distinguished from PMPs and TMPs by the extent of their conformational reorganization after binding to the inner membrane. While conformational adaptation of membrane proteins in response to the altered environment is common and even expected, AMPs undergo enormous conformational changes following membrane binding. Some AMPs convert from surface-binder to transmembrane or translocate through the membrane to interact with internal targets. Others form a variety of exquisite oligomeric structures that either remain on the membrane surface or become transmembrane. Not surprisingly, therefore, the way in which AMPs disrupt bacterial membranes is also highly divergent. The most common models to describe this process include the formation of barrel-stove, toroidal pore, and carpet-like structures, or a combination of these. Some AMPs may act as detergents or electroporation agents [84, 85]. Consequently,

multiple mechanisms have been proposed to explain how AMPs kill bacteria, ranging from the formation of transmembrane pores [86, 87] to impact on membrane structural integrity without pore formation [88–91].

13.6 Summary

The plasma membrane is a complex, semipermeable envelope of cells with an actively controlled composition, dynamics, and function; it is a fluid "mosaic" of proteins floating-like icebergs in a moving "sea" of lipids. The plasma membrane allows for the passive diffusion of uncharged nonpolar molecules (e.g. O_2 and CO_2 and fatty acids), and harbors protein channels and pumps for the transport of small ions (e.g. Na^+ and K^+), carrier proteins for the transport of large polar molecules, such as amino acids and glucose, and receptors for hormones, neurotransmitters, and other signals. In addition, the plasma membrane is home to, and an organizing platform for, many transiently binding signaling proteins, enzymes, and scaffolding proteins. In this chapter, we have discussed how proteins use an exquisite combination of charged, polar, and hydrophobic amino acids to recognize and bind to specific regions of membranes, including the plasma membrane. Global shape, local structure, charge distribution, and lipid modification have been discussed as some of the most important determinants of protein–membrane interactions, lateral dynamics, and sensing or induction of curvature. Specifically, a precise mixture of polar and nonpolar protein–lipid interactions define the structure and function of TMPs, including channels and GPCRs. Similarly, a combination of structure, positively charged and nonpolar amino acids, and modification by lipophilic motifs dictate how PMPs are targeted to membrane surfaces. We have used AMPs as an example of proteins that utilize the entire range of protein–membrane interaction patterns. AMPs achieve enormous conformational adaptations by presenting a precise mix of basic, polar and nonpolar residues to alternately interact with the polar or negatively charged surface and hydrophobic core of membranes. This allows most AMPs to conditionally become peripheral and transmembrane, unlike TMPs that are permanently membrane-associated or PMPs that are transiently membrane-tethered.

Acknowledgment

AAG acknowledges financial support from the National Institutes of Health (grant # R01GM124233) and the Cancer Prevention and Research Institute of Texas (CPRIT grant # RP190366). LJ acknowledges financial support from the University of Houston grant No. GEAR IO96691; Marie Curie Actions – International Incoming Fellowship (IIF), project number 254470; Romanian National Authority for Scientific Research, CNDI-UEFISCDI, project numbers PN-II-RU-TE-2014-4-2418 and PN-III-P1-1.1-TE-2016-0032.

References

1 Zimmerberg, J. and Kozlov, M.M. (2006). How proteins produce cellular membrane curvature. *Nat. Rev. Mol. Cell Biol.* 7 (1): 9–19. https://doi.org/10.1038/nrm1784.

2 McConnell, H.M. and Vrljic, M. (2003). Liquid-liquid immiscibility in membranes. *Annu. Rev. Biophy. Biomol. Struct.* 32: 469–492. https://doi.org/10.1146/annurev.biophys.32.110601.141704.

3 Huang, J. and Feigenson, G.W. (1999). A Microscopic interaction model of maximum solubility of cholesterol in lipid bilayers. *Biophys. J.* 76 (4): 2142–2157. https://doi.org/10.1016/S0006-3495(99)77369-8.

4 Niemelä, P.S. et al. (2007). Assessing the nature of lipid raft membranes. *PLoS Comput. Biol.* 3 (2): e34. https://doi.org/10.1371/journal.pcbi.0030034.

5 Pandit, S.A. et al. (2007). Cholesterol surrogates: a comparison of cholesterol and 16:0 ceramide in POPC bilayers. *Biophys. J.* 92 (3): 920–927. https://doi.org/10.1529/biophysj.106.095034.

6 Pandit, S.A. et al. (2008). Cholesterol packing around lipids with saturated and unsaturated chains: a simulation study. *Langmuir* 24 (13): 6858–6865. https://doi.org/10.1021/la8004135.

7 Brown, D.A. (2006). Lipid rafts, detergent-resistant membranes, and raft targeting signals. *Physiology* 21: 430–439. https://doi.org/10.1152/physiol.00032.2006.

8 Lingwood, D. and Simons, K. (2010). Lipid rafts as a membrane-organizing principle. *Science* 327 (5961): 46–50. https://doi.org/10.1126/science.1174621.

9 Sezgin, E. et al. (2017). The mystery of membrane organization: composition, regulation and physiological relevance of lipid rafts. *Nat. Rev. Mol. Cell Biol.* 18 (6): 361–374. https://doi.org/10.1038/nrm.2017.16.

10 Lin, X. and Gorfe, A.A. (2019). Understanding membrane domain-partitioning thermodynamics of transmembrane domains with potential of mean force calculations. *J. Phys. Chem. B* 123 (5): 1009–1016. https://doi.org/10.1021/acs.jpcb.8b10148.

11 Lorent, J.H. and Levental, I. (2015). Structural determinants of protein partitioning into ordered membrane domains and lipid rafts. *Chem. Phys. Lipids* 192: 23–32. https://doi.org/10.1016/j.chemphyslip.2015.07.022.

12 Cornish, J. et al. (2020). Intrinsically disordered proteins and membranes: a marriage of convenience for cell signalling? *Biochem. Soc. Trans.* 48 (6): 2669–2689. https://doi.org/10.1042/BST20200467.

13 Parton, R.G. (2018). Caveolae: structure, function, and relationship to disease. *Annu. Rev. Cell Dev. Biol.* 34: 111–136. https://doi.org/10.1146/annurev-cellbio-100617-062737.

14 Abankwa, D. and Gorfe, A.A. (2020). Mechanisms of Ras membrane organization and signaling: Ras rocks again. *Biomolecules* 10 (11): 1522. https://doi.org/10.3390/biom10111522.

15 Janosi, L. et al. (2012). Organization, dynamics, and segregation of Ras nanoclusters in membrane domains. *Proc. Natl. Acad. Sci. U.S.A.* 109 (21): 8097–8102. https://doi.org/10.1073/pnas.1200773109.

16 Li, Z. et al. (2012). Formation and domain partitioning of H-Ras peptide nanoclusters: effects of peptide concentration and lipid composition. *J. Am. Chem. Soc.* 134 (41): 17278–17285. https://doi.org/10.1021/ja307716z.

17 Lin, X. et al. (2015). Reversible effects of peptide concentration and lipid composition on H-Ras lipid anchor clustering. *Biophys. J.* 109 (12): 2467–2470. https://doi.org/10.1016/j.bpj.2015.11.009.

18 Li, H. and Gorfe, A.A. (2013). Aggregation of lipid-anchored full-length H-Ras in lipid bilayers: simulations with the MARTINI force field. *PLoS One* 8 (7): e71018. https://doi.org/10.1371/journal.pone.0071018.

19 Li, Z. and Gorfe, A.A. (2013). Deformation of a two-domain lipid bilayer due to asymmetric insertion of lipid-modified Ras peptides. *Soft Matter* 9 (47): 11249–11256. https://doi.org/10.1039/C3SM51388B.

20 Li, H. and Gorfe, A.A. (2014). Membrane remodeling by surface-bound protein aggregates: insights from coarse-grained molecular dynamics simulation. *J. Phys. Chem. Lett.* 5 (8): 1457–1462. https://doi.org/10.1021/jz500451a.

21 Li, Z. and Gorfe, A.A. (2014). Modulation of a small two-domain lipid vesicle by linactants. *J. Phys. Chem. B* 118 (30): 9028–9036. https://doi.org/10.1021/jp5042525.

22 Zhou, Y. et al. (2017). Lipid-sorting specificity encoded in K-Ras membrane anchor regulates signal output. *Cell* 168 (1–2): 239–251.e16. https://doi.org/10.1016/j.cell.2016.11.059.

23 Lobo, S. (2019). Protein palmitoylation in cancer. In: *Unravelling Cancer Signaling Pathways: A Multidisciplinary Approach* (ed. K. Bose and P. Chaudhari). Singapore: Springer https://doi.org/10.1007/978-981-32-9816-3_3.

24 Cho, W. and Stahelin, R.V. (2006). Membrane binding and subcellular targeting of C2 domains. *Biochim. Biophys. Acta, Mol. Cell. Biol. Lipids* 1761 (8): 838–849. https://doi.org/10.1016/j.bbalip.2006.06.014.

25 Huang, H. and Cafiso, D.S. (2008). Conformation and membrane position of the region linking the two C2 domains in synaptotagmin 1 by site-directed spin labeling. *Biochemistry* 47 (47): 12380–12388. https://doi.org/10.1021/bi801470m.

26 Hurley, J.H. (2006). Membrane binding domains. *Biochim. Biophys. Acta, Mol. Cell. Biol. Lipids* 1761 (8): 805–811. https://doi.org/10.1016/j.bbalip.2006.02.020.

27 Monje-Galvan, V. and Klauda, J.B. (2016). Peripheral membrane proteins: tying the knot between experiment and computation. *Biochim. Biophys. Acta, Biomembr.* 1858 (7 Pt B): 1584–1593. https://doi.org/10.1016/j.bbamem.2016.02.018.

28 Jaud, S. et al. (2007). Self-induced docking site of a deeply embedded peripheral membrane protein. *Biophys. J.* 92 (2): 517–524. https://doi.org/10.1529/biophysj.106.090704.

29 Lemmon, M.A. (2008). Membrane recognition by phospholipid-binding domains. *Nat. Rev. Mol. Cell Biol.* 9 (2): 99–111. https://doi.org/10.1038/nrm2328.

30 Resh, M.D. (2013). Covalent lipid modifications of proteins. *Curr. Biol.* 23 (10): R431–R435. https://doi.org/10.1016/j.cub.2013.04.024.

31 Prior, I.A. et al. (2020). The frequency of Ras mutations in cancer. *Cancer Res.* 80 (14): 2969–2974. https://doi.org/10.1158/0008-5472.CAN-19-3682.

32 Huster, D. et al. (2003). Membrane insertion of a lipidated Ras peptide studied by FTIR, solid-state NMR, and neutron diffraction spectroscopy. *J. Am. Chem. Soc.* 125 (14): 4070–4079. https://doi.org/10.1021/ja0289245.

33 Gorfe, A.A. et al. (2004). Membrane localization and flexibility of a lipidated Ras peptide studied by molecular dynamics simulations. *J. Am. Chem. Soc.* 126 (46): 15277–15286. https://doi.org/10.1021/ja046607n.

34 Gorfe, A.A. et al. (2007a). H-Ras protein in a bilayer: interaction and structure perturbation. *J. Am. Chem. Soc.* 129 (40): 12280–12286. https://doi.org/10.1021/ja073949v.

35 Gorfe, A.A. et al. (2007b). Free energy profile of H-Ras membrane anchor upon membrane insertion. *Angew. Chem. Int. Ed.* 46 (43): 8234–8237. https://doi.org/10.1002/anie.200702379.

36 Gorfe, A.A. and McCammon, J.A. (2008). Similar membrane affinity of mono- and Di-S-acylated Ras membrane anchors: a new twist in the role of protein lipidation. *J. Am. Chem. Soc.* 130 (38): 12624–12625. https://doi.org/10.1021/ja805110q.

37 Gorfe, A.A. et al. (2008). Water-membrane partition thermodynamics of an amphiphilic lipopeptide: an enthalpy-driven hydrophobic effect. *Biophys. J.* 95 (7): 3269–3277. https://doi.org/10.1529/biophysj.108.136481.

38 Janosi, L. and Gorfe, A.A. (2010). Segregation of negatively charged phospholipids by the polycationic and farnesylated membrane anchor of Kras. *Biophys. J.* 99 (11): 3666–3674. https://doi.org/10.1016/j.bpj.2010.10.031.

39 Neale, C. and Garcia, A.E. (2020). The plasma membrane as a competitive inhibitor and positive allosteric modulator of KRas4B signaling. *Biophys. J.* 118 (5): 1129–1141. https://doi.org/10.1016/j.bpj.2019.12.039.

40 El Amri, M. et al. (2018). MARCKS and MARCKS-like proteins in development and regeneration. *J. Biomed. Sci.* 25: 43. https://doi.org/10.1186/s12929-018-0445-1.

41 Resh, M.D. (1999). Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins. *Biochim. Biophys. Acta, Mol. Cell. Res.* 1451 (1): 1–16. https://doi.org/10.1016/S0167-4889(99)00075-0.

42 Anguita, E. and Villalobo, A. (2017). Src-family tyrosine kinases and the $Ca2^{+}$ signal. *Biochim. Biophys. Acta, Mol. Cell. Res.* 1864 (6): 915–932. https://doi.org/10.1016/j.bbamcr.2016.10.022.

43 Le Roux, A.-L. et al. (2019). A myristoyl-binding site in the SH3 domain modulates c-Src membrane anchoring. *iScience* 12: 194–203. https://doi.org/10.1016/j.isci.2019.01.010.

44 Donaldson, J.G. and Jackson, C.L. (2011). ARF family G proteins and their regulators: roles in membrane transport, development and disease. *Nat. Rev. Mol. Cell Biol.* 12: 362–375. https://doi.org/10.1038/nrm3117.

45 Goldberg, J. (1998). Structural basis for activation of ARF GTPase: mechanisms of guanine nucleotide exchange and GTP–myristoyl switching. *Cell* 95 (2): 237–248. https://doi.org/10.1016/S0092-8674(00)81754-7.

46 Liu, Y. et al. (2009). Structure and membrane interaction of myristoylated ARF1. *Structure* 17 (1): 79–87. https://doi.org/10.1016/j.str.2008.10.020.

47 Gorfe, A.A. et al. (2007c). Structure and dynamics of the full-length lipid-modified H-Ras protein in a 1,2-dimyristoylglycero-3-phosphocholine bilayer. *J. Med. Chem.* 50 (4): 674–684. https://doi.org/10.1021/jm061053f.

48 Prakash, P. and Gorfe, A.A. (2017). Membrane orientation dynamics of lipid-modified small GTPases. *Small GTPases* 8 (3): 129–138. https://doi.org/10.1080/21541248.2016.1211067.

49 MacKenzie, K.R. and Fleming, K.G. (2008). Association energetics of membrane spanning α-helices. *Curr. Opin. Struct. Biol.* 18 (4): 412–419. https://doi.org/10.1016/j.sbi.2008.04.007.

50 Popot, J.L. and Engelman, D.M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry* 29 (17): 4031–4037. https://doi.org/10.1021/bi00469a001.

51 Janosi, L. and Ceccarelli, M. (2013). The gating mechanism of the human aquaporin 5 revealed by molecular dynamics simulations. *PLoS One* 8 (4): e59897. https://doi.org/10.1371/journal.pone.0059897.

52 Lemmon, M.A. et al. (1994). A dimerization motif for transmembrane α-helices. *Nat. Struct. Mol. Biol.* 1 (3): 157–163. https://doi.org/10.1038/nsb0394-157.

53 Russ, W.P. and Engelman, D.M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. *J. Mol. Biol.* 296 (3): 911–919. https://doi.org/10.1006/jmbi.1999.3489.

54 Fleming, K.G. et al. (1997). The effect of point mutations on the free energy of transmembrane α-helix dimerization. *J. Mol. Biol.* 272 (2): 266–275. https://doi.org/10.1006/jmbi.1997.1236.

55 Anbazhagan, V. and Schneider, D. (2010). The membrane environment modulates self-association of the human GpA TM domain–implications for membrane protein folding and transmembrane signaling. *Biochim. Biophys. Acta, Biomembr.* 1798 (10): 1899–1907. https://doi.org/10.1016/j.bbamem.2010.06.027.

56 Janosi, L. et al. (2010). Lipid-modulated sequence-specific association of glycophorin A in membranes. *Biophys. J.* 99 (1): 284–292. https://doi.org/10.1016/j.bpj.2010.04.005.

57 Petrache, H.I. et al. (2000). Modulation of glycophorin A transmembrane helix interactions by lipid bilayers: molecular dynamics calculations. *J. Mol. Biol.* 302 (3): 727–746. https://doi.org/10.1006/jmbi.2000.4072.

58 Herbst, R.S. (2004). Review of epidermal growth factor receptor biology. *Int. J. Radiat. Oncol. Biol. Phys.* 59 (2 Suppl): 21–26. https://doi.org/10.1016/j.ijrobp.2003.11.041.

59 Bublil, E.M. and Yarden, Y. (2007). The EGF receptor family: spearheading a merger of signaling and therapeutics. *Curr. Opin. Cell Biol.* 19 (2): 124–134. https://doi.org/10.1016/j.ceb.2007.02.008.

60 Schlessinger, J. (2000). Cell signaling by receptor tyrosine kinases. *Cell* 103 (2): 211–225. https://doi.org/10.1016/s0092-8674(00)00114-8.

61 Linggi, B. and Carpenter, G. (2006). ErbB receptors: new insights on mechanisms and biology. *Trends Cell Biol.* 16 (12): 649–656. https://doi.org/10.1016/j.tcb.2006.10.008.

62 Yarden, Y. and Schlessinger, J. (1987). Epidermal growth factor induces rapid, reversible aggregation of the purified epidermal growth factor receptor. *Biochemistry* 26 (5): 1443–1451. https://doi.org/10.1021/bi00379a035.

63 Ceresa, B.P. and Vanlandingham, P.A. (2008). Molecular mechanisms that regulate epidermal growth factor receptor inactivation. *Clin. Med. Oncol.* 2: 47–61. https://doi.org/10.4137/cmo.s498.

64 Henriksen, L. et al. (2013). Internalization mechanisms of the epidermal growth factor receptor after activation with different ligands. *PLoS One* 8 (3): e58148. https://doi.org/10.1371/journal.pone.0058148.

65 Schlessinger, J. (2002). Ligand-induced, receptor-mediated dimerization and activation of EGF receptor. *Cell* 110 (6): 669–672. https://doi.org/10.1016/s0092-8674(02)00966-2.

66 Zhang, X. et al. (2006). An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* 125 (6): 1137–1149. https://doi.org/10.1016/j.cell.2006.05.013.

67 Prakash, A. et al. (2011). GxxxG motifs, phenylalanine, and cholesterol guide the self-association of transmembrane domains of ErbB2 receptors. *Biophys. J.* 101 (8): 1949–1958. https://doi.org/10.1016/j.bpj.2011.09.017.

68 Arkhipov, A. et al. (2013). Architecture and membrane interactions of the EGF receptor. *Cell* 152 (3): 557–569. https://doi.org/10.1016/j.cell.2012.12.030.

69 Doura, A.K. and Fleming, K.G. (2004). Complex interactions at the helix-helix interface stabilize the glycophorin A transmembrane dimer. *J. Mol. Biol.* 343 (5): 1487–1497. https://doi.org/10.1016/j.jmb.2004.09.011.

70 Prakash, A. et al. (2010). Self-association of models of transmembrane domains of ErbB receptors in a lipid bilayer. *Biophys. J.* 99 (11): 3657–3665. https://doi.org/10.1016/j.bpj.2010.10.023.

71 Zhang, J. and Lazaridis, T. (2009). Transmembrane helix association affinity can be modulated by flanking and noninterfacial residues. *Biophys. J.* 96 (11): 4418–4427. https://doi.org/10.1016/j.bpj.2009.03.008.

72 Hauser, A., Attwood, M., Rask-Andersen, M. et al. (2017). Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discovery* 16: 829–842. https://doi.org/10.1038/nrd.2017.178.

73 Sriram, K. and Insel, P.A. (2018). G protein-coupled receptors as targets for approved drugs: how many targets and how many drugs? *Mol. Pharmacol.* 93 (4): 251–258. https://doi.org/10.1124/mol.117.111062.

74 Gimple, G. (2016). Interaction of G protein coupled receptors and cholesterol. *Chem. Phys. Lipids* 199: 61–73. https://doi.org/10.1016/j.chemphyslip.2016.04.006.

75 Hedger, G. and Sansom, M.S.P. (2016). Lipid interaction sites on channels, transporters and receptors: recent insights from molecular dynamics simulations. *Biochim. Biophys. Acta, Biomembr.* 1858 (10): 2390–2400. https://doi.org/10.1016/j.bbamem.2016.02.037.

76 Bruzzese, A. et al. (2018). Structural insights into positive and negative allosteric regulation of a G protein-coupled receptor through protein-lipid interactions. *Sci. Rep.* 8 (1): 4456. https://doi.org/10.1038/s41598-018-22735-6.

77 Dawaliby, R. et al. (2016). Allosteric regulation of G protein-coupled receptor activity by phospholipids. *Nat. Chem. Biol.* 12 (1): 35–39. https://doi.org/10.1038/nchembio.1960.

78 Neale, C. et al. (2015). Can specific protein-lipid interactions stabilize an active state of the beta 2 adrenergic receptor? *Biophys. J.* 109 (8): 1652–1662. https://doi.org/10.1016/j.bpj.2015.08.028.

79 Song, W. et al. (2019). State-dependent lipid interactions with the A2a receptor revealed by MD simulations using in vivo-mimetic membranes. *Structure* 27 (2): 392–403.e3. https://doi.org/10.1016/j.str.2018.10.024.

80 Verkman, A.S. (2013). Aquaporins. *Curr. Biol.* 23 (2): R52–R55. https://doi.org/10.1016/j.cub.2012.11.025.

81 Aponte-Santamaria, C. et al. (2012). Molecular driving forces defining lipid positions around aquaporin-0. *Proc. Natl. Acad. Sci. U.S.A.* 109 (25): 9887–9892. https://doi.org/10.1073/pnas.1121054109.

82 Stansfeld, P.J. et al. (2013). Multiscale simulations reveal conserved patterns of lipid interactions with aquaporins. *Structure* 21 (5): 810–819. https://doi.org/10.1016/j.str.2013.03.005.

83 Haney, E.F. et al. (2019). Reassessing the host defense peptide landscape. *Front. Chem.* 7: 43. https://doi.org/10.3389/fchem.2019.00043.

84 Hale, J.D.F. and Hancock, R.E.W. (2007). Alternative mechanisms of action of cationic antimicrobial peptides on bacteria. *Expert Rev. Anti-Infective Ther.* 5 (6): 951–959. https://doi.org/10.1586/14787210.5.6.951.

85 Kumar, P. et al. (2018). Antimicrobial peptides: diversity, mechanism of action and strategies to improve the activity and biocompatibility in vivo. *Biomolecules* 8 (1): 4. https://doi.org/10.3390/biom8010004.

86 Matsuzaki, K. et al. (1998). Relationship of membrane curvature to the formation of pores by magainin 2. *Biochemistry* 37 (34): 11856–11863. https://doi.org/10.1021/bi980539y.

87 Rapaport, D. and Shai, Y. (1991). Interaction of fluorescently labeled pardaxin and its analogues with lipid bilayers. *J. Biol. Chem.* 266 (35): 23769–23775.

88 Andersson, D.I. et al. (2016). Mechanisms and consequences of bacterial resistance to antimicrobial peptides. *Drug Resist. Updates* 26: 43–57. https://doi.org/10.1016/j.drup.2016.04.002.

89 Epand, R.M. et al. (2016). Molecular mechanisms of membrane targeting antibiotics. *Biochim. Biophys. Acta, Biomembr.* 1858 (5): 980–987. https://doi.org/10.1016/j.bbamem.2015.10.018.

90 Shai, Y. (2002). Mode of action of membrane active antimicrobial peptides. *Biopolymers* 66 (4): 236–248. https://doi.org/10.1002/bip.10260.

91 Yeaman, M.R. and Yount, N.Y. (2003). Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.* 55 (1): 27–55. https://doi.org/10.1124/pr.55.1.2.

14

Interactions of Proteins with Small Molecules, Allosteric Effects

Michael C. Hutter

Saarland University, Center for Bioinformatics, Campus Building E2.1, D-66123 Saarbrücken, Germany

Abbreviations

AMPA	α-Amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ATP	Adenosine triphosphate
GABA	γ-Aminobutyric acid
GPCR	G-protein-coupled receptor
hERG	Human ether-a-gogo-related gene
IR	Infrared
NMR	Nuclear magnetic resonance
SCF-MP2	Self-consistent field-Møller–Plesset perturbation theory second order
VSEPR	Valence shell electron pair repulsion

14.1 Introduction

The available repertoire of pharmaceutical substances comprises more than 50 000 small-molecule drugs that either were in use or all still marketed. Repositories, such as the ZINC database, hold more than 750 million purchasable compounds that are categorized as drug-like or at least lead-like [1]. Together with the naturally occurring substrates of enzymes, these molecules share a common property: They all bind more or less strongly, specifically, and persistently to various proteins. The physicochemical principles that mediate these nonbonding interactions are basically the same as in protein–protein or protein–DNA interactions, only the size of the ligands and their possible binding sites can be different.

The prime experimental method to reveal the contact interface between protein and ligand in atomistic detail is still X-ray crystallography. The key characteristic of more than 50 000 protein–ligand complexes being resolved so far is that the ligand has to be in direct contact with its receptor to exert its function, which is a measurable change, for example, in enzymatic activity. The need for direct binding was

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

formulated by Paul Ehrlich decades before the first crystallographic structure of a protein was available: "Corpora non agunt nisi fixata." Likewise, the macroscopic idea of the lock-and-key principle by Emil Fischer regarding the specificity of ligands dates back to the nineteenth century. Both hypotheses still remain valid today, but we have developed a more detailed view of these modes of interaction.

14.2 Modes of Binding to Proteins

The simplest approach is that neither ligand nor protein change their conformation upon binding, which would allow rigid-body docking because only shape complementarity is required according to the lock-and-key principle. The complex of benzamidine in β-trypsin (PDB entry 3PTB) is one of the few enzyme–ligand systems where this assumption is justified. Typically, the ligand contains multiple rotatable bonds and thus will adopt the rotamer in the binding pocket that is energetically most favorable. Likewise, the amino acid side chains of the receptor will undergo conformational changes during binding. Flexible docking can only account for a limited number of them because it is computationally not feasible to simultaneously consider all rotatable bonds of the protein during ligand binding. Instead, the obtained docking pose can be refined using molecular dynamics simulations. In practice, this is, however, only doable for a strongly limited number of ligands owing to the computational demand as well as the manual effort of setting up simulations and evaluating their results. So far, this induced fit corresponds to the mechanical operation of a patent key when opening the lock.

A general aspect of molecular dynamics simulations is their function to reveal the accessible conformations of polypeptides over the respective time span. Thus, transient binding pockets on the surface become apparent. In particular, small-molecule inhibitors of protein–protein interaction bind to such transient pockets, thus giving rise to the so-called conformational selection [2, 3]. Once the appropriate pocket is available, ligand binding stabilizes the corresponding conformation of the protein.

Whereas the physicochemical interactions are generally the same no matter where and how ligands bind, their exerted function heavily depends on the respective mechanism of action at the particular binding site. Furthermore, proteins, especially those considered as drug targets, can be classified into distinct groups not only according to their biochemical function but, in this context, according to the mechanistic effect that these ligands have on them. Due to their high expression level, enzymes have first been exploited as targets for reversible and irreversible inhibitors. Corresponding ligands must exhibit a substantially higher binding affinity than the natural substrate, ideally, several orders of magnitude; thus allowing low dosage, which is, in turn, advantageous for avoiding side effects as less substance is available that could bind to other receptors. In the case of irreversible inhibitors, the binding is not competitive, but instead, a chemical reaction leads to covalent binding to the enzyme, rendering it unfunctional until degradation. The prototypic example is the acetylation of serine by acetylsalicylic acid inside Cyclooxygenase I, which blocks access to the catalytic center.

Receptors, in particular, G-protein-coupled receptors (GPCRs), transfer signals from one side of a membrane to the other one, which eventually causes the desired effect further downstream in the signal transduction cascade. Thus, they are widely exploited as drug targets (about 30% of all prescription medicines). GPCRs possess several conformational states and binding sites for a variety of ligands, which allow multiple ways of modulating receptor response. Whereas agonists intensify the response, antagonists prevent this action, either by competitive binding or indirectly by allosteric effects. This can be achieved either by causing a change in the three-dimensional structure of the protein and modulating the affinity of the receptor for substances that work as agonists or by preventing the receptor response by other mechanisms. Such allosteric effects will be addressed in detail later in this chapter (see Section 14.6).

Finally, ion channels and transporters are also transmembrane proteins involved in either selectively modulating the passage of (small inorganic) ions or active uptake or efflux of compounds that otherwise do not passively diffuse through membranes, such as strongly polar or charged organic substances. Their functionality can be prevented by blockers or inhibitors, whereas openers keep ion channels in a permanently open state, respectively. Their response can be much faster, for example, in depleting the concentration of neurotransmitters in the synaptic cleft.

14.3 Types of Interaction Between Protein and Ligand

The specific physical characteristics of (known) protein–ligand interactions allow to list them in order of decreasing energetic contribution to the energy of binding. Nevertheless, their importance for selectivity is independent of this ranking and rather governed by biochemical and pharmaceutical considerations.

14.3.1 Salt Bridges

Interactions between charged particles follow Coulomb's law and thus depend on distance and magnitude of charge. Therefore, attraction between two oppositely charged ions is strongest in vacuum and media of low dielectric constant, such as the interior of proteins. Here, these short-distance contacts involving the side chains of aspartate, glutamate, lysine, and arginine contribute to the stability of secondary and tertiary structural elements (see Figure 14.1A(a)). In contrast to the situation in aqueous solution where the prior desolvation of ions is energetically unfavorable, the resulting stabilization that goes along with the transfer into the hydrophobic protein interior makes up for the loss of entropy. Since the energies of salt bridges (>400 kJ/mol) are in the same range as ionic bonds in crystals or are even stronger than covalent single bonds, these are, therefore, the dominant terms regarding ligand binding, if corresponding interactions are present. Multiple-charged amino acid side chains can be found, for example, in the binding pockets of kinases, which accommodate the strongly negatively charged adenosine triphosphate (ATP). Since the present charges of the amino acids have to be compensated by ligands to

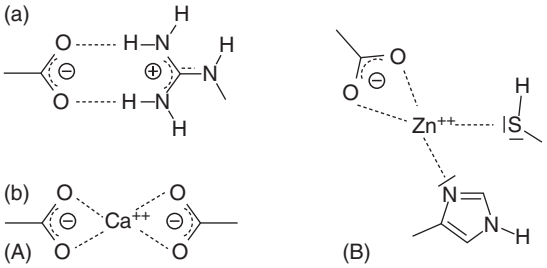


Figure 14.1 Salt bridges between oppositely charged amino acid side chains stabilize secondary and tertiary structural elements in proteins (A), here shown for the carboxylate groups of aspartate or glutamate and arginine (a), or involving an inorganic cation (b). Likewise, metal ions can be complexed by charged groups, or coordinated by lone pairs, typically involving the side chains of cysteine and histidine (B).

yield reasonable binding affinities, this requires likewise inhibitors that contain multiple-charged groups, which hamper passive uptake.

14.3.2 Coordination of Ions via Lone Pairs

Interactions between formally positively charged ions and the lone pair of a ligating atom (N, O, S) are also called "dative bonds" or "coordinate covalent bonds", due to the fact that this is actually a 2-center-2-electron bond and therefore can be attributed as covalent bond. In contrast to usual σ-bonds, the two electrons come from the same atom (provided by one lone pair), which forms the Lewis base, while the metal ion is the Lewis acid. In bulk solution (metal–aquo complexes), as well as in organometallic compounds (including metalloenzymes), this kind of bonding is dominant. Although this ion–dipole interaction is electrostatically less attractive than ion–ion interactions (i.e. salt bridges), it can be as strong as typical covalent bonds. According to molecular orbital theory, *p*-orbitals of the ligand arising from lone pairs overlap with *d*-orbitals of the metal forming π-bonds that correspond to double bonds. Whereas negative charge from the ligand atoms would be transferred onto the more electropositive metal, the latter π-back bonding leads to a depletion of this unfavorable charge density on the metal, while simultaneously the metal–ligand bond is strengthened [4].

In contrast to loosely bound counter ions (e.g. Na^+, K^+, and Ca^{2+}) of charged amino acid side chains at the protein surface, metal ions (especially zinc and iron) are found in the active center of metalloenzymes. According to the valence shell electron pair repulsion (VSEPR) theory, the coordination sphere of metal ions is only partially saturated by interactions with protein atoms, and the remaining places are filled up by water or substrate molecules. Most frequently, such catalytic metal ions are coordinated by the side chains of histidine or cysteine (lone pairs of nitrogen or sulfur, respectively). Furthermore, ionic bidentate coordination by both carboxylate oxygens of aspartate and glutamate is observed (see Figure 14.1B). Likewise, this can be exploited for drug design, where similar functional groups mimic these interactions that otherwise are occupied by the natural substrates (see Figure 14.2). Thus, coordinative binding can provide both affinity and selectivity.

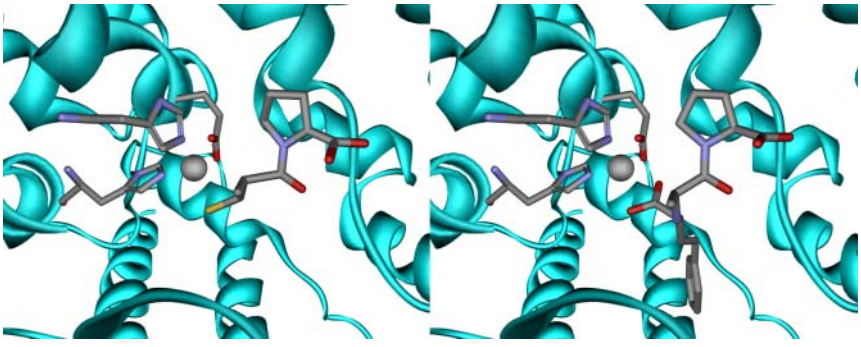


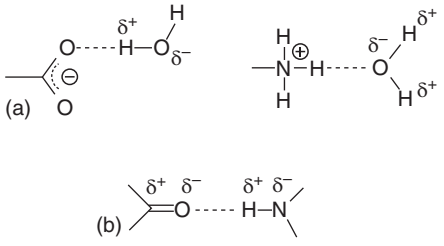
Figure 14.2 The inhibitors captopril (a) vs. enalapril (b) in contact with the zinc ion in the binding pocket of the angiotensin-converting enzyme (PDB entry codes 1UZF and 1UZE, respectively). The coordination of the metal ion is achieved either by the lone pair of sulfur (captopril) or via a carboxylate group (enalapril). Interaction with the zinc ions is likewise found for all later developed ACE inhibitors, whereas their sizes vary largely, aiming to fill the remaining binding pocket.

14.3.3 Hydrogen Bonds

14.3.3.1 Definition

Among noncovalent interactions, hydrogen bonds can be more than 10 times stronger than other van der Waals interactions and therefore have received special attention. Formally, the interaction occurs between the lone (electron) pair of an electron-rich atom (usually nitrogen, oxygen, or sulfur) denoted acceptor (Ac) and the antibonding molecular orbital of the bond between hydrogen and a further electronegative atom, being the donor (Dn) of the hydrogen bond (see Figure 14.3) [5]. Therefore, the interaction involves electrostatic (dipole–dipole or ion–dipole) as well as partially covalent contributions that can lead to resonant structures, particularly if the Dn–H···Ac system at hand is charged and the proton is shared between the two electronegative atoms. In typical protein–ligand interactions, such situations do, however, not arise and the covalent bond between hydrogen and donor can be assigned unambiguously. The character and actual amount of electronic contributions in hydrogen bonds are still heavily debated in the literature [5]. Conversely, the role of hydrogen bonding for selectivity is undisputed.

Figure 14.3 Hydrogen bonds involving charged groups (a) are stronger than those between neutral partners (b) because the former ones involve ion–dipole interactions.



Notable changes upon hydrogen bonding include lengthening of the Dn—H bond by about 0.1 Å, which shifts the bond-stretching frequency toward lower energy, a higher vibrational frequency in the infrared (IR) spectrum (redshift), respectively. The change in electron density can also be detected in the ^{1}H-NMR spectrum, where strong hydrogen bonds cause downfield shifts.

The H· · ·Ac hydrogen bond itself covers a distance range of 1.6–2.0 Å. Lower values lead to repulsion, whereas the attractive interaction vanishes for longer distances. Besides its length, a hydrogen bond is considered to be optimal if the angle between donor, hydrogen, and the respective lone pair of the acceptor is close to 180° (colinear arrangement of Dn–H· · ·Ac). Physical equations that model the shape of hydrogen bond energies make use of trigonometric functions ($\cos^2$ or $\cos^4$) that cause the energy to sharply drop below 170° and to vanish for values ≤90°, also due to steric considerations [6, 7]. Consequently, the colinear hydrogen bonds present in antiparallel β-sheets are energetically stronger than those in parallel β-sheets, which substantially deviate from the collinear Dn–H· · ·Ac orientation.

14.3.3.2 Occurrence and Functionality of Hydrogen Bonds in Biological Systems

Besides intermolecular hydrogen bonds that mediate ligand–receptor interactions, intramolecular hydrogen bonds stabilize secondary structural elements in proteins (i.e. α-helices and β-sheets). Likewise, hydrogen bonding between the nucleic bases is a key element of double-stranded DNA. In contrast to hydrogen bonds in bulk water, these hydrogen bonds are embedded in an otherwise hydrophobic environment of low dielectric constant. For electrostatic reasons, such hydrogen bonds are even stronger than in aqueous solution. Likewise, binding pockets of proteins are mostly hydrophobic, except for those that are located on the surface, where they are exposed to the solvent. Moreover, the lacking option to form alternative hydrogen bonds as there is no surrounding water leads to a loss in entropy compared to the unfolded protein or the unbound ligand, respectively. Regarding ligand binding, in buried pockets, this loss of entropy is at least partially compensated by the expelled water molecules from these cavities. Once in bulk solution, these waters can now form many more alternative hydrogen bonds than inside the binding pocket, which drastically increases their entropy.

It is obvious that ligands must saturate corresponding hydrogen bond interactions in the binding pocket for energetic reasons. Likewise, these hydrogen bonds between receptor and ligand enable selective binding. Depending on the size and number of polar or charged amino acids, the count of these hydrogen bonds can vary drastically. A case where hydrogen bonding accounts for most of the energetic part of the binding affinity is the biotin–streptavidin complex, with an experimentally determined association constant in the order of 10^{14} mol/l, which is unusually high for a noncovalent inhibitor [8]. There, six almost optimally orientated hydrogen bonds are formed in the binding pocket.

Selectivity of inhibitors is, however, typically mediated by only one or two hydrogen bonds between receptor and ligand. Particularly, drugs that work on targets in the central nervous system (CNS) often possess only one hydrogen bond

functionality (either a donor or an acceptor atom), which renders them more lipophilic than other orally administered medications. Moreover, CNS-active drugs are usually chemically less complex than other pharmaceutical agents. Among the guidelines for drug design, Lipinski's rule of five was the first to notice the limiting effect of log *P* (the water–octanol partitioning coefficient as numerical criterion for lipophilicity), count of hydrogen-bonding functionalities, and molecular weight on oral bioavailability [9]. Too many hydrogen bonds render a compound strongly hydrophilic and thus unlikely to diffuse through the lipid bilayers of biological membranes. Therefore, compounds, such as sugars (for example glucose) and oligopeptides, are subject to active uptake by transporters. Drugs that contain corresponding chemical fragments are likewise substrates of these transporters. For example, captopril that contains an L-proline group is recognized by the PEPT1 transporter. Conversely, strongly lipophilic molecules are less water-soluble and tend to remain in hydrophobic environments, such as lipid bilayers. Both extremes have to be avoided, but there are numerous exceptions to this and other suggested rules (about 20% of all marketed drugs).

14.3.3.3 Classification of Hydrogen Bonds

According to Jeffrey, the strength of hydrogen bonds can be classified into the three categories; strong (63–167 kJ/mol involving HF and thus not relevant in biological systems), medium (17–63 kJ/mol that is the typical interaction involving nitrogen or oxygen as acceptor atom and a polar hydrogen atom), and weak (<17 kJ/mol, such as C—H· · ·O hydrogen bonds where the hydrogen atom is nonpolar) [10, 11]. Hydrogen bonds between uncharged partners (e.g. C—O· · ·H—N) are always weaker than those where the acceptor atom is negatively charged (e.g. C(—O)O$^-$· · ·H—O—H), or conversely if the donor atom is positively charged (see Figure 14.3). The reason for this is that the electrostatic ion–dipole interaction is stronger than that between two dipoles (neutral hydrogen bonds).

14.3.3.4 Weak Hydrogen Bonds

Hydrogen bonds involving sulfur are substantially weaker compared to their counterparts with oxygen. First, the distance between donor and acceptor atom is increased as a consequence of the 0.45 Å larger van der Waals radius of sulfur, which weakens the electrostatic attraction according to Coulomb's law, and second, due to the lower electronegativity compared to oxygen. Interestingly, hydrogen bonds involving sulfur as acceptor exhibit more pronounced directionality than corresponding carbonyl oxygen acceptors, as shown by evaluation of X-ray crystallographic structures [11]. According to VSEPR theory, there should be no substantial difference, as both lone pairs are oriented the same way [12]. Moreover, the lone pairs of sulfur are expected to be larger (3*p* compared to 2*p* orbitals) and the angle between them to be slightly widened due to the increased repulsion that goes along with their size. Conversely, by taking the actual electron density around oxygen into account, there is substantial electron density between the lone pairs of oxygen, which are far less evolved than VSEPR theory suggests. Likewise, oxygen is smaller than sulfur, and thus, the proton can come closer to this electron density upon

hydrogen bonding. Consequently, hydrogen bonding in sulfur occurs more via the lone pairs. Moreover, hydrogen bonds in ice are longer and exhibit emphasized directionality compared to liquid water at low temperatures, likewise accounting for the density anomaly of water. Other untypical high physical constants, such as heat capacity, density, dielectric constant, and boiling point of water are also attributed to hydrogen bonding.

Even, weaker hydrogen bonds than those involving sulfur are observed in crystal structures between C–H· · ·O and N–H· · ·C. Their interaction energies are stronger than usual van der Waals contacts, because the interactions involve stronger dipoles, due to the presence of electronegative elements. Nevertheless, these contacts can be important in drug design.

14.3.3.5 Hydrogen Bonds to Fluorine

In principle, fluorine should also be a suitable hydrogen bond acceptor, being more electronegative than nitrogen and oxygen, causing a stronger dipole moment. Corresponding C—F· · ·H—N bonds are, however, rarely seen in X-ray crystallographic structures [13, 14]. Conversely, C–F· · ·H–C van der Waals contacts are observed frequently. Electronegativity, however, increases with the tendency to accept electrons and not protons, which is the key feature of a hydrogen bond acceptor. Covalently bound fluorine is instead a weak Lewis base and an extremely weak proton acceptor. Therefore, corresponding hydrogen bonds to fluorine are weak (6–10 kJ/mol) and exhibit less directionality [13]. Thus, the most common reason for the use of fluorine in medicinal chemistry is to block metabolically labile sites in drugs or to increase the hydrophobicity of the ligand without introducing larger lipophilic groups.

14.3.3.6 Nitrogen vs. Oxygen as Competing Hydrogen Bond Acceptors

In rational ligand design, frequently functional groups have to be exchanged to improve binding affinity. Since heterocycles are widely used building blocks, the question arises if nitrogen or oxygen is the better hydrogen bond acceptor. Böhm et al. analyzed crystal structures of corresponding complexes and also performed *ab initio* self-consistent field Møller–Plesset perturbation theory second-order (SCF-MP2) computations of the interaction energies using triple-zeta plus polarization functions as basis set [15]. Their results clearly show that hydrogen bonds to nitrogen as acceptors lead to much stronger interaction energies, which is reflected by the higher frequency of such contacts in crystal structures. Whereas carbonyl oxygen is a good acceptor, the strength of interaction decreases for sp^3-hybridized oxygen the more aliphatic substituents are attached.

14.3.3.7 Bifurcated Hydrogen Bonds

Since oxygen possesses two lone pairs, it can, therefore, accept two (or even three) hydrogen bonds, making it over-coordinated oxygen. Such arrangements are called bifurcated hydrogen bonds and are found in water clusters (giving rise to a tetrahedral arrangement of four oxygen atoms around a central one), carbonyl oxygens (e.g. in the biotin–streptavidin complex), and carboxylate anions (side chains of aspartate and glutamate). Cases, where single hydrogen participates in more than

one hydrogen bond, seem to be rare, except in bulk water during reorientation of hydrogen bonds.

14.3.4 Halogen Bonds

In halogen bonds (Dn—X· · ·Ac, where X is fluorine, chlorine, bromine, or iodine, while Dn can also be carbon instead of nitrogen), the Lewis acid–base relationship is reversed compared to that in hydrogen bonds with respect to the electron-donor and electron-acceptor atoms, although both hydrogen and halogen gain electron density upon bonding. The resulting interaction energies range from 5 to 180 kJ/mol and thus compete with hydrogen bonds. The X· · ·Ac distance is shorter than the sum of the van der Waals radii of both atoms, which would otherwise cause repulsion. Compared to hydrogen bonds, one observes emphasized directionality. This is due to the anisotropic electron density distribution around the halogen, which gave rise to the definition of the so-called σ-hole [16]. Evaluation of crystallographic structures of proteins and nucleic acids revealed that the relevance of halogen bonds has been overlooked, which is remarkable because halogens are ubiquitously present in drugs (see Figure 14.4) [17]. Beside halogen–oxygen interactions, also contacts with nitrogen and sulfur have been observed, although not as frequently.

Voth et al. showed that halogen bonds to carbonyl oxygens of the protein backbone are most often found in a geometrical arrangement that forms an almost right angle (85°) to the corresponding hydrogen bond to that oxygen atom [18]. They furthermore suggested that such halogen bonds are orthogonal (in a functional sense rather than in Euclidean space) and energetically independent from the classical hydrogen bonds. The local van der Waals surface of both α-helices, as well as β-sheets forms

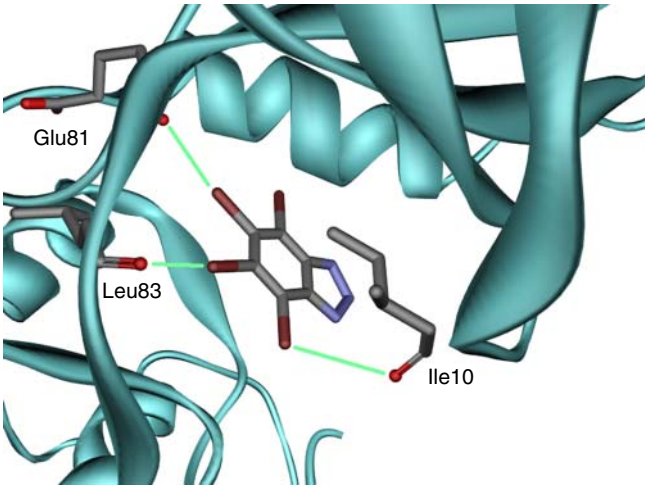


Figure 14.4 Tetrabromobenzotriazole bound to human phospho-CDK2/cyclin (PDB code 1P5E) shows three halogen bonds (light-green lines) to carbonyl oxygens of the protein backbone as prominent interactions. Since this kinase is essential for meiosis, it is a widely exploited target in cancer therapy.

a pocket where the halogen also makes hydrophobic contact with the side chain of the amino acids that contain the hydrogen bond donor.

14.3.5 van der Waals Interactions

The weakest intermolecular interactions are summarized as van der Waals interactions. These range from less than 1 kJ/mol up to values of hydrogen bonds. In contrast to all other ionic, covalent, or dative bonds, these do not arise from a chemical bond-like situation where electrons are shared over two or more atomic centers, and the attraction vanishes rapidly with increasing distance between the considered atoms. Conversely, repulsion occurs at short distances, which increases almost exponentially when the electron density between these (noncovalently connected) atoms is decreased and thus the nuclei repel each other. Further electrostatic forces arise due to interactions between dipole moments that are either permanently present (Keesom interaction), dipoles that are induced by permanent dipoles (Debye force), and finally, the London forces that stem from mutually induced instantaneous dipole moments. Whereas the former dipoles are due to permanently present electric monopoles, these instantaneous dipoles are caused by (random) fluctuations of the charge density within the electron clouds around the atoms, leading to very short-lived multipoles. Therefore, this interaction is also termed London dispersion interaction. The superposition of attraction and repulsion gives rise to the Lennard–Jones potential, where the energetic terms reciprocally depend on powers of the distance r between the considered pairs of atoms. The repulsive part is typically proportional to $1/r^{12}$, although smaller exponents (r^9 or r^{10}) result in a less steep curve. For the attractive part, $1/r^6$ is commonly used.

The strength of interaction increases with the polarizability of the involved distribution of electrons (see Figure 14.5). For example, the interaction between two methyl groups comprises only σ-electrons that are rather confined around the respective single bonds and thus are hardly polarizable. Conversely, electrons of π-systems can be polarized much more easily as the corresponding p-orbitals occupy substantially more space and are furthermore delocalized over multiple atoms. The electronic behavior of aromatic systems closely resembles that of electrons in a coil exhibiting shielding effects due to induced ring currents, which can be detected in nuclear magnetic resonance (NMR) spectra as chemical shifts.

Whereas the attractive part of the van der Waals interactions is rather weak, the repulsive part is the dominating design principle for the overall shape of the ligand. It is easy to understand that a ligand that does not fit into a binding pocket for steric reasons cannot exhibit suitable binding affinity due to steric clashes. Eventually, this is the reason for the lock-and-key principle.

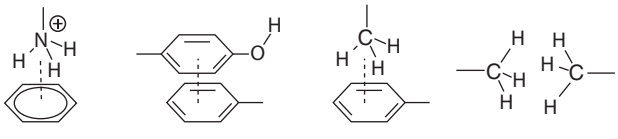


Figure 14.5 Dispersive intermolecular interactions in order of decreasing energy from left to right.

14.3.6 Mutual Interactions of Delocalized π-Electron Systems

At first sight, the interaction between electrons should be always repulsive, according to Coulomb's law. However, this applies only to a static picture of the particles. It has been mentioned above that interactions between instantaneous dipole moments in electrons lead to the attractive London forces. This attraction becomes even larger once π-electrons are involved. The simplest system to elucidate stacking effects of aromatic π-systems is the benzene dimer, which was extensively studied [19]. The appropriate treatment of the electron dispersion requires the use of theoretical methods that account for electron correlation (e.g. SCF-MP2 or coupled-cluster), as well as rather large counterpoise-corrected basis sets to reflect dimerization and diffuse electron distribution. The computed interaction energy is in the range of 8–12 kJ/mol, with the two most stable arrangements being the parallel displaced dimer and a perpendicular T-shaped conformation, whereas the sandwich-like complex that has both ring centers aligned on top of each other is less stable. This can be explained by the competing quadrupole moments that arise from the local dipole moments of the C—H bonds, which destabilize this sandwich configuration. Similarly, X-ray structures of proteins and double-stranded DNA show the displaced dimer being the most often adopted arrangement between corresponding aromatic side chains (i.e. phenylalanine, tyrosine, histidine, and tryptophan) or nucleic bases, respectively.

The reason why particularly heterocyclic aromatic rings are ubiquitously found in drugs is not primarily due to the stronger interaction compared to other nonaromatic rings, but rather for their synthetic accessibility and their preferred metabolic profile compared to benzene rings. Furthermore, heterocycles are less hydrophobic than benzene, because they can form hydrogen bonds, which helps to keep the log *P* low.

14.3.7 Cation–π Interaction

Even, stronger electrostatic attraction is expected for the interaction of a cation (monopole) with the quadrupole moment of delocalized π-systems (see Figure 14.5, leftmost). Corresponding interaction energies are computed to be around 45 kJ/mol in the gas phase [20]. Interesting for ligand design is the interaction of fragments that contain quaternary nitrogen; for example, the neurotransmitter acetylcholine to the nicotinic acetylcholine esterase, where the positively charged $N(Me)_3$ group binds to tryptophan. The side chains of arginine and lysine side can conversely interact with aromatic rings of inhibitors, whereby the more localized charge density in lysine causes the stronger interaction [21].

14.3.8 Anion–π Interaction

To enable substantial attractions of an anion, the charge distribution of the delocalized π-system has to be (in principle) reversed so that the charge density on the sp^2-hybridized carbon atoms is largely depleted. This can be achieved by strongly electron-withdrawing substituents, e.g. fluorine, or the use of π-electron-poor aromatic moieties. The latter approach has been applied to the design of specific sensors

to detect simple inorganic anions, i.e. chlorine [22]. However, no corresponding protein inhibitor that makes use of this kind of interaction is known as of yet.

14.3.9 Unusual Protein–Ligand Contacts

The intermolecular contacts so far have been presented in order of decreasing energetic strength from the point of electrostatics. The specific nature of synthetic ligands, however, leads to a bias of certain interactions that are mainly due to the frequency of preferred chemical fragments in pharmaceutical drugs; for example, sulfonyl groups, or nitrile substituents, which form weak hydrogen bonds [23]. Likewise, halogen bonding to the carbon atom of backbone peptide linkers was observed. The statistical significance of such unusual contacts was investigated in detail by Kuhn et al. [24]. In contrast to the above-mentioned weak hydrogen bonds, the C—F· · ·H—N bond was not found to be relevant.

14.4 Modeling Intermolecular Interactions by Force Fields and Docking Simulations

Whereas covalent bonds are modeled in force fields using more or less elaborated mathematical descriptions of the corresponding potential shapes for effects due to bond stretching, bond bending, and rotation around single bonds, the representation of nonbonding interactions relies on extensive parameterization of rather simple functions. Electrostatic interaction is computed by applying Coulomb's law using point charges on the nuclear centers and sometimes using distance-dependent dielectric constants to account for shielding effects due to water or protein environment. Basically, all methods for generating the required partial atomic charges will result in positive values for the more electropositive atoms and conversely in negative values for the more electronegative elements, to maintain the overall net charge. In turn, negative charge distributions due to *p*-orbitals of π-systems or lone electron pairs are neglected, unless additional dipole representations are used that enable iterative treatment of polarization effects (so-called polarizable force fields) [25]. As a consequence, all interactions involving such distributions of electrons, e.g. cation–π attraction or π–π stacking, cannot be represented properly by using point charges on the nuclei only. The same holds for corresponding scoring functions used in protein–ligand docking. On the other hand, these induced electrostatic interactions are substantially weaker than salt bridges and can be, at least partially, accounted for by suitable parameterization of the van der Waals interactions. Even explicit terms for hydrogen bonding seem to be expendable if corresponding optimized parameters for the respective atom types are used [26]. In fact, tuning and calibration of parameters to the experimental binding affinity have shown to be a successful strategy for interaction terms that cannot be addressed otherwise, such as effects due to solvation or entropy [27, 28].

The strong steric repulsions of overlapping van der Waals radii as given by the Lennard–Jones potential will cause even slightly wrong ligand conformations to be ranked far inferior than they should be. As a work-around, the repulsive potential can either be cut off at a given energy threshold or by using a less steeply rising term that allows moderate clashes, while preserving the continuity of the potential form.

14.5 Entropic Aspects

Except for hydrogen bonds, so far only purely energetic aspects of protein–ligand interactions have been addressed, which would correspond to the static picture once the ligand is bound to the protein. The actual process of binding goes, however, along with substantial changes in the entropy of both enzyme and ligand, which enter into the free energy of binding [29]. The most obvious change arises from the loss of translational and rotational degrees of freedom of the ligand that is no longer free to move around in solution. For any inhibitor, this term is, however, very similar to that of the natural substrate and thus negligible. The major difference and what can rationally be addressed is the count of rotatable single bonds in the inhibitor, which will also be conformationally restricted. The recommended upper limit is eight rotatable single bonds, not counting those that are members of ring systems because the flexibility of cyclic systems is limited anyhow [30].

Entropic changes upon expelling water molecules from the binding site have been mentioned earlier in the context of hydrogen bonding. This process positively contributes to binding affinity, because these water molecules gain entropy due to increased mobility in solution. The situation for hydrophobic surface areas of the ligand is similar to that inside the binding pocket. Water molecules cannot form hydrogen bonds to the corresponding atoms of the ligand and are thus restricted. Once these hydrophobic surface parts are buried inside the equally hydrophobic binding pocket, these water molecules now can form more dynamic hydrogen bonds to other solvent waters, which increase their entropy. The corresponding desolvation term is proportional to the hydrophobic surface area of the ligand and can be included in scoring functions for docking [27]. While it seems advantageous to maximize the hydrophobic surface of inhibitors from this point of view, the worse solubility in aqueous media is the limiting factor for oral bioavailability [9].

14.6 Allosteric Effects: Conformational Changes Upon Ligand Binding

The rational design of ligands by energetically optimizing the interactions at the binding interface has been addressed over decades. The gained experience and principles can be used in virtual screening, for example, by applying filtering steps according to rules for bioavailability and other guidelines, as well as molecular docking. On the other hand, there are far less rules and approaches to predict conformational changes of proteins upon ligand interaction, in particular,

if these go beyond the change of rotameric states of amino acid side chains, which can be accounted for in flexible docking [31]. A further assumption is that competitive binding to such orthosteric sites does not substantially change the overall conformation of the protein.

Local changes of secondary structural elements have been observed in X-ray structures, for example, upon binding of statins to 3-hydroxy-3-methylglutaryl-CoA reductase [32]. The situation for GPCRs, ion channels, and transporters is even more difficult, because there are fewer crystallographic structures available, on top of the presence of (several) allosteric binding sites. These areas are different from the orthosteric site where the endogenous agonist of the receptor binds, which corresponds to the binding pocket for substrates of enzymes. Conversely, allosteric modulators comprise substances that modulate the receptor's response by binding to other sites than the orthosteric site. In particular, GPCRs offer multiple possibilities for making use of corresponding ligands. This functionality requires a coupling process through space that seems to be conformationally triggered. The resulting effect will be rather different from that of a typical antagonist that blocks access of the agonist to the orthosteric site as it can be either positive, negative, or neutral. A positive allosteric modulator (PAM) can increase both the likelihood of binding the agonist, its ability to activate the receptor or both simultaneously. Neutral allosteric ligands (NAL, also called silent allosteric modulators; SAM) do not alter the activity of the agonist but prevent other effectors from binding to one of the possible allosteric sites, but without exerting a negative effect themselves. Only negative allosteric modulators (NAM) show the same effects as antagonists by decreasing the affinity or efficacy of the agonist.

So far, discovery of most allosteric modulators has been made by chance, although the number of known effectors and binding sites is drastically rising [31]. In general, allosteric modulators cause a change in the conformation of the receptor. For example, they can stabilize one of the conformations that are naturally occurring during the activation and deactivation cycle of GPCRs. Often the shape of the orthosteric site is also affected, which has direct consequences for the binding of the agonist. Such modulation of the binding affinity has been observed for benzodiazepines and barbiturates, which are both PAMs of the $GABA_A$ receptor, thereby enhancing the agonistic effect of the neurotransmitter γ-aminobutyric acid (GABA).

Allosteric modulators are interesting for drug development, because they are typically more specific than orthosteric ligands and thus show less adverse effects due to decreased binding affinity to off-targets. The reason for this is the higher degree of conservation of orthosteric sites, which accommodate endogenous ligands. Mutations in these regions are, therefore, more likely to result in severe functional consequences. Conversely, allosteric sites, which are less important for the function of the receptor, allow larger variability of the amino acid sequence during evolutionary processes.

An example how allosteric modulators can overcome a crucial drug-resistant mutation of the orthosteric site in the BCR–ABL1 fusion product is the allosteric inhibitor asciminib. The so-called "gatekeeper mutation" of threonine in position 315 to isoleucine causes a whole series of orthosteric inhibitors, including imatinib,

bosutinib, nilotinib, and dasatinib, to become non-functional. Asciminib exerts its inhibitory activity by binding to the C-lobe of the kinase that is otherwise the myristoyl-binding pocket required for auto-inhibition, and thus locks BCR–ABL1 into an inactive conformation [33].

Besides stabilizing a particular conformation directly, allosteric modulators can also affect the unbinding of the agonist, causing the receptor to remain longer in the active conformation. For example, piracetam is a PAM of the α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor, which is activated endogenously by glutamate, thereby opening the ion-channel mediating fast synaptic transmission in the CNS. Moreover, cyclothiazide also works as PAM but prevents desensitization of the receptor. This phenomenon is observed during repeated or intense exposure to agonists. In the AMPA receptor, the ligand-binding domain is a dimer, which can be disrupted, but is stabilized by cyclothiazide. The therapeutic spectrum of AMPA receptor ligands furthermore comprises antagonists and NAMs.

14.7 Aspects of Ligand Design Beyond Protein–Ligand Interactions

From the thermodynamic point of view, an effective inhibitor must show substantially larger binding affinity than the present natural ligand. Competing against an endogenous substance that is available in the cell in high concentration, such as pyruvate or ATP, therefore requires the binding constant of the inhibitor to be higher by several orders of magnitude. Moreover, such ubiquitously used catabolites are substrates of numerous enzymes, which are often vital for homeostasis. A typical example is kinases. Nevertheless, successful inhibitors for various kinases in cancer treatment have been designed that bind to subpockets that are not occupied by ATP itself. Typically, one strives for selective inhibitors to avoid obvious side effects that arise from binding to other enzymes. In particular notorious and seemingly unspecific antitargets, such as the human ether-a-gogo-related gene (hERG) channel have to be avoided. On the other hand, too much selectivity for the target at hand can be disadvantageous when mutations in the binding pocket can easily cause resistance, which was observed for numerous antibiotics and anti-HIV drugs. Conversely, a certain degree of promiscuity may lead to synergistic effects, if the affected enzymes are therapeutic targets belonging to the same category. For example, the vasopeptidase inhibitor omapatrilat binds to the angiotensin-converting enzyme and neutral endopeptidase, which, both, are part of the regulatory system that lowers blood pressure.

The straightforward application of protein–ligand interactions for the design of ligands with maximum potency will furthermore result in rather large and lipophilic inhibitors that fill up the binding pocket as much as possible. Again, this is unfavorable because solubility and bioavailability will be decreased and likewise undesired metabolization will be increased [34]. Finally, limitations due to synthetic accessibility have to be kept in mind.

14.8 Conclusions

Based on the physicochemical aspects that promote binding of ligands to proteins shown here, it becomes apparent that one has to make use of all possibilities of forming attractive interactions to maximize binding affinity. Most important are ionic interaction and polar interactions, such as salt bridges and contacts to metal ions, because these contribute most to the binding energy. Additionally, hydrogen bonds mediate selectivity. So far, the importance of halogen bonds seems to be underestimated, although their energetic contribution is similar to that of hydrogen bonds. Finally, repulsive van der Waals interactions due to steric clashes in the binding pocket are a limiting factor in ligand design, which can be, however, addressed by docking simulations. Conversely, conformational changes of the receptor upon ligand binding, both orthosteric and allosteric sites, are difficult to predict and account for.

References

1 Irwin, J.J. and Stoichet, B.K. (2005). ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45 (1): 177–182. https://doi.org/10.1021/ci049714.

2 Tsai, C. J., Kumar, S., Ma, B. et al. (1999) Folding funnels, binding funnels, and protein function, *Protein Sci.* 8 (6): 1181−1190. doi:https://doi.org/10.1110/ps.8.6.1181

3 Motlagh, H. N., Wrabl, J. O., Li, J. et al. (2014) The ensemble nature of allostery. *Nature* 508: 331−339. doi:https://doi.org/10.1038/nature13001

4 Huheey, J.E., Keiter, E.A., and Keiter, R.L. (1993). *Inorganic Chemistry: Principles of Structures and Reactivity*, 4th ed. Addison Wesley Pub. Co. Inc. ISBN: 978-0060429959.

5 Weinhold, F., Klein, R. A. (2014) What is a hydrogen bond? Resonance covalency in the supramolecular domain. *Chem. Educ. Res. Pract.* 15: 276–285. doi:https://doi.org/10.1039/c4rp00030g

6 P.J.A. Goodford (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* 28 (8): 849–857. doi:https://doi.org/10.1021/jm00145a002

7 Vedani, A. (1988) Yeti: an interactive molecular mechanics program for small-molecule protein complexes. *J. Comput. Chem.* 9 (3): 269–280. doi:https://doi.org/10.1002/jcc.540090310

8 Chaiet, L., Wolf, F.J. (1964) The properties of streptavidin, a biotin-binding protein produced by *Streptomycetes*. *Arch. Biochem. Biophys.* 106: 1–5. doi:https://doi.org/10.1016/0003-9861(64)90150-X

9 Lipinski, C.A., Lombardo, F., Dominy, B.W. et al. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* 23 (1–3): 3–25. doi:https://doi.org/10.1016/S0169-409X(96)00423-1

10 Jeffrey, G.A. (1997). *An Introduction to Hydrogen Bonding*. Oxford University Press. ISBN: 978-0-19-509549-4.

11 Steiner, T. (2002) The hydrogen bond in the solid state. *Angew. Chem. Int. Ed.* 41: 48–76. doi:https://doi.org/10.1002/1521-3773(20020104)41:1

12 Gillespie, R. J. (2004) Teaching molecular geometry with the VSEPR model. *J. Chem. Educ.* 81 (3): 298–304. doi:https://doi.org/10.1021/ed081p298

13 Howard, J. A. K., Hoy, V. J., O'Hagan, D. et al. (1996) How good is fluorine as a hydrogen-bond acceptor? *Tetrahedron* 52 (38): 12613–12622. doi:https://doi.org/10.1016/0040-4020(96)00749-1

14 Dunitz, J. D., Taylor, R. (1997) Organic fluorine hardly ever accepts hydrogen bonds. *Chem. Eur. J.* 3 (1): 89–98. doi:https://doi.org/10.1002/chem.19970030115

15 Böhm, H.-J., Brode, S., Hesse, U. et al. (1996) Oxygen and nitrogen in competitive situations: Which is the hydrogen-bond acceptor? *Chem. Eur. J.*, 2 (12): 1509–1513. doi:https://doi.org/10.1002/chem.19960021206

16 Clark, T., Hennemann, M., Murray, J. S. et al. (2007) Halogen bonding: the σ-hole. *J. Mol. Model.* 13 (2): 291–296. doi:https://doi.org/10.1007/s00894-006-0130-2

17 Auffinger, P., Hays, F.A., Westhof, E. et al. (2004). Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci. U.S.A.*, 101 (48): 16789–16794. doi:https://doi.org/10.1073/pnas.0407607101

18 Voth, A. R., Oishi, K., Ho, P. S. (2009) Halogen bonds as orthogonal molecular interactions to hydrogen bonds. *Nat. Chem.* 1 (1): 74–79. doi:https://doi.org/10.1038/nchem.112

19 Sinnokrot, M.O., Valeev, E. F., Sherrill, C.D. (2002) Estimates of the ab initio limit for π–π interactions: the benzene dimer. *J. Am. Chem. Soc.* 124 (36): 10887–10893. doi:https://doi.org/10.1021/ja025896h

20 Dougherty, D. A., Ma, J.C. (1997) The cation–π interaction. *Chem. Rev.* 97 (5): 1303–1324. doi:https://doi.org/10.1021/cr9603744

21 Pitt, W. R., Parry, D. M., Perry, B. G. et al. (2009) Heteroaromatic rings of the future. *J. Med. Chem.* 52 (9): 2952–2963. doi:https://doi.org/10.1021/jm801513z

22 de Hoog, P., Gamez, P. Mutikainen, I. et al. (2004) An aromatic anion receptor: anion–π interactions do exist. *Angew. Chem. Int. Ed.* 43 (43): 5815–5817. doi:https://doi.org/10.1002/ange.200460486

23 Bissantz, C., Kuhn, B., Stahl, M. (2010) A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* 53 (14): 5061–5084. doi:https://doi.org/10.1021/jm100112j

24 Kuhn, B., Gilberg, E., Taylor, R., et al. (2019) How significant are unusual protein–ligand interactions? Insights from database mining. *J. Med. Chem.* 62 (22): 10441–10455. doi:https://doi.org/10.1021/acs.jmedchem.9b01545

25 Wang, Z. X., Zhang, W., Wu, C. et al. (2006) Strike a balance: optimization of backbone torsion parameters of amber polarizable force field for simulations of proteins and peptides. *J. Comput. Chem.* 27 (6): 781–790. doi:https://doi.org/10.1002/jcc.20386

26 Wang, J., Wolf, R. M., Caldwell, J. W. et al. (2004) Development and testing of a general amber force field. *J. Comput. Chem.* 25 (9): 1158–1174. doi:https://doi.org/10.1002/jcc.20035

27 Morris, G.M., Goodsell, D.S., Halliday, R.S. et al. (1998). Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. *J. Comput. Chem.* 19: 1639–1662. https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14%3C1639::AID-JCC10%3E3.0.CO;2-B.

28 Böhm, H.-J. (1994). The development of a simple empirical scoring function to estimate the binding-constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* 8: 243–256. https://doi.org/10.1007/BF00126743.

29 Du, X., Li, Y., Xia, Y.-L. et al. (2016) Insights into protein–ligand interactions: mechanism, models, and methods, *Int. J. Mol. Sci.* 17(2): 144–177. doi:https://doi.org/10.3390/ijms17020144

30 Oprea, T. I. (2000) Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* 14 (3): 251–264. doi:https://doi.org/10.1023/a:1008130001697

31 Lu, S., He, X., Ni, D. et al. (2019). Allosteric modulator discovery: from serendipity to structure-based design, *J. Med. Chem.*, 62 (14): 6405–6421. doi:https://doi.org/10.1021/acs.jmedchem.8b01749

32 Istvan, E. S., Deisenhofer, J. (2001) Structural mechanism for statin inhibition of HMG-CoA reductase. *Science* 292 (5519): 1160–1164. doi:https://doi.org/10.1126/science.1059344

33 Hughes, T. P., Mauro, M. J., Cortes, J. E. et al. (2019) Asciminib in chronic myeloid leukemia after ABL kinase inhibitor failure. *N. Engl. J. Med.* 381 (24): 2315–2326. doi:https://doi.org/10.1056/NEJMoa1902328

34 Hann, M. M. (2011) Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* 2: 349–355. doi:https://doi.org/10.1039/C1MD00017A

15

Effects of Mutations in Proteins on Their Interactions

Alexander Gress[1] *and Olga V. Kalinina*[1,2,3]

[1] *Helmholtz Institute for Pharmaceutical Research Saarland (HIPS) / Helmholtz Centre for Infection Research (HZI), Saarbrücken, Germany*
[2] *Saarland University, Medical Faculty, Homburg, Germany*
[3] *Center for Bioinformatics, Saarland Informatics Campus, Saarbrücken, Germany*

15.1 Introduction

One way a mutation can have an effect on a protein interaction is when it changes the amino acid sequence of one of the proteins. Single-nucleotide variants (SNVs) are the most common type of genetic variation [1], and when talking about mutations in this chapter, we always mean genetic variations that lead to the substitution of a single amino acid (*non-synonymous SNVs*).

From a biochemical point of view, a change in an amino acid can affect interaction in many different ways. There can be obvious mechanisms, for example, if the wild-type amino acid formed an important non-covalent bond (e.g. a salt bridge) with a residue on the interface of the interacting protein chain and the mutant amino acid can no longer do that (e.g. it is not charged). This will typically result in lower total interaction strength between the two proteins. A less obvious mechanism would be a mutation that leads to a change in the overall conformation of the protein that, in turn, changes the interaction interface and weakens the total strength of the interaction. On the other hand, mutations can also have a stabilizing effect on protein–protein interactions (PPIs), for example, by introducing a new bond at the interface to the interaction partner or by simply enlarging the interaction interface. However, it is well known that most SNVs are either functionally neutral [2] or weakly deleterious [3], and thus the corresponding protein mutations do not affect PPIs at all or have only a minor negative effect on binding.

The great challenge is to differentiate mutations that weaken, strengthen, or do not affect an interaction without making an expensive biochemical experiment by just employing computational biology methods. Such methods differ, for example, in how much information is needed for prediction. As a starting point, in all cases, the amino acid sequence of the protein has to be known – without it, one would not be able to identify the mutation in the first place. Given the sequence of the protein (and potentially protein interaction partners), one can resort to another very useful

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

source of information: sequences of evolutionarily related proteins or homologs. The benefit of employing this information depends on the quantity and diversity of these sequences, but thanks to ever-growing amounts of sequencing data, the quantity and diversity of homologous sequences are constantly improving. Later in this chapter, we will show how computational methods leverage these data, and why it is important to have a large and diverse set of homologs. The next level of information that can be used by computational tools is experimentally resolved three-dimensional (3D) structures of proteins, which typically renders the tools more accurate than purely sequence-based approaches. Many methods can only be applied to structures that contain both partners of the PPI, which drastically limits the number of cases where those comparatively more precise methods can be applied. If an experimentally resolved 3D structure of a homologous complex is available, sometimes it can be used as a source of structural data as well, and comparing multiple related structures can shed additional light on the consequences of a mutation.

This necessitates development of tools that efficiently search and store data related to the availability of sequences of homologs or structural data for a given mutation. The latter problem is solved by so-called *structural annotation methods* that are described in detail in Section 15.2. Later, in Section 15.3, we move on to an overview of computational methods that perform the actual *mutation effect prediction*.

15.2 Structural Annotation of Mutations in Proteins

Structural annotation can be defined as the task of relating experimentally resolved protein structures to protein sequences. In its simplest form, this means assigning one or more 3D structures to a given amino acid sequence of a protein. In most cases, the protein is not given by its sequence specifically but by an identifier of a database storing the sequences of proteins or the corresponding genes that encode the protein, and hence information about the amino acid range covered by a particular structure is required. These methods, as a rule, rely on protein there-dimensional structures represented by entries in the Protein Data Bank (PDB) [4].

The simple protein-to-structure mapping can be solved by a sequence similarity search, but structural annotation has more facets than this. The more interesting problem setting is the mapping of individual amino acids in a given protein to individual residues in a protein structure. This is a nontrivial task, since certain 3D structures may lack certain parts of sequences due to biological (signal peptide cleavage and posttranslational modifications) or technical reasons. If a residue-level mapping between a protein sequence and a 3D structure is built, this can be automatically used in the most common application scenario: structural annotation of nsSNVs. Hence most structural annotation methods focus on such mutations.

Structural annotation methods differ in many aspects: what type of input they can process; the type and quantity of annotated structures; and the complexity of subsequent structural analysis of the results. While technically structural analysis is not necessarily a part of structural annotation, it is performed more often than not as the default option. The reason for this is that users that are interested in structural

annotation, especially that of mutations, are usually also interested in the details of the spatial environment of the annotated positions. For example, for mutations, a very important type of analysis is describing their potential impact on interaction interfaces.

Generally, there are two types of structural annotation methods: databases and on-demand automated pipelines. In databases, all annotations and analyses are precomputed; with the obvious caveat that custom inputs, especially mutations not stored in the database, cannot be processed. In pipelines, such input can be processed, but since their annotation and analysis are performed in real time, this could result in long waiting times, especially if the performed analysis is more comprehensive.

15.2.1 Databases for Structural Annotation of Mutations

Over the years, several databases were developed for structural annotation of mutations. The general idea behind them was to somehow map a set of clinically relevant mutations onto three-dimensional structures of the same or homologous protein that were resolved experimentally. Such a structural view of clinically described mutations may provide insights into the biochemical mechanisms behind the observed phenotypes. This information was sometimes combined with additional annotations and predictions.

In 2003, MutDB [5] was the first database for structurally annotation mutations to be published, and many more were to follow. MutDB combined all protein sequences from Swiss-Prot [6] and mutation data from dbSNP [7], and employed BLAST [8] to perform a sequence similarity search against the PDB. Then, the results were limited to hits with a sequence identity of 100%, which made position-specific annotation trivial. In 2008, MutDB was updated [9], and in addition to structural annotation also incorporated KEGG [10] pathway annotations, and results of multiple mutation effect prediction methods, including SNPs3D [11], PolyPhen [12, 13], SIFT [14], and others (for a discussion of these methods, see Section 15.3). Another tool, SNPs3D is primarily a method for mutation effect prediction; however, prediction results for all mutations contained in dbSNP and Human Gene Mutation Database (HGMD, [15]) are provided in the form of a database. The structural annotation is only given when an experimental structure of the target protein was available. In those cases, a support vector machine that predicts stability changes introduced by mutations [16] was used. LS-SNP/PDB expands the structural annotation concept to include analysis of 3D structures of homologs by adopting the template search pipeline from MODELLER [17]. LS-SNP/PDB also adds the results of basic structural analysis, including solvent accessibility, secondary structure assignment, and interaction interface assignments. MSV3D included all protein sequences from the OMIM [18] database, in which clinically relevant mutations were annotated using multiple public online databases, including dbSNP [7], SwissVar [19], and several gene-specific databases. For all proteins, if at least one 3D structure was assigned,

MODELLER was used to provide protein structure models, in addition, to experimentally resolved 3D structures. All these databases are not available online anymore. SAAPdb [20, 21] can still be downloaded but is not maintained. Its intention was to provide a more comprehensive structural analysis and visualization than its competitors.

The current landscape of databases (as well as of other tools) for structural annotation of mutations is dynamic and changing, so we focus on a few recent ones (Table 15.1). There are disease-specific and general-purpose databases. As the name suggests, Cancer3D [22, 26] focuses on mutations that are associated with cancer. The basis for the database are mutations coming from cancer cell line encyclopedia (CCLE) [27] and the cancer genome atlas (TCGA) [28]. The protein sequences that contain cancer-associated mutations were compared with all proteins in the PDB [4] using the BLAST [8] sequence similarity search tool. Structures of homologous (nonhuman) proteins were considered, but an imposed threshold of e-value below 10^{-6} strongly limited their amount. The alignments provided by BLAST were used to create position-specific mappings. Later, Cancer3D was extensively updated [22], supplying the community with Cancer3D v2. The update included an expansion on the total amount of cancer-associated mutations using the updates of the underlying databases, as well as an improved structural analysis, including annotation of flexible and disordered regions. An important innovation of Cancer3D was identification of protein–protein, protein–nucleic acid interaction interfaces, and ligand-binding pockets in all annotated structures by finding the residues that lie closer than 5 Å to the corresponding interaction partner (this functionality was also earlier implemented in some dynamic structural annotation methods, see Section 15.2.2). Hence, it became possible to find candidate mutations that could have an impact on these interactions. We discuss dedicated tools in more detail in Section 15.3.

The concept behind mutfunc is similar to Cancer3D but vastly expands on the number of mutations stored in the database by not limiting it to cancer-associated mutations (ExAC [29] and ClinVar [30] were considered) and also by including mutations from two other species, *Escherichia coli* and *Saccharomyces cerevisiae*, extracting them from public genome data and publications. The computational strategy behind the structural annotation in mutfunc is quite different, though. First, the UniProt [31] annotations for each chain in each PDB entry were considered, and structures that maximized the coverage were assigned to each protein. Second, for UniProt entries, to which a structure could not be assigned in this way, homology-based protein structure modeling using the ModPipe [32] software was performed, assigning a structure to a larger portion, but not to all proteins in UniProt. This "one-structure-per-protein" strategy allowed for a more comprehensive structural analysis of the annotated structures due to the limited number of structures. However, this strategy has an obvious disadvantage of losing a lot of information by discarding a lot of experimentally resolved structures. In particular, different interaction partners present in different structures of the same protein are not detected. The variety of analysis methods that are implemented in mutfunc is impressive. First, mutations that participate in interactions are detected by calculating the solvent accessibility area for the isolated protein chains and their

Table 15.1 Currently available databases for structural annotation of mutations.

Database	URL	Year of publication	Species	Homologs	Isoforms	Structural analysis
Cancer3D [22]	http://cancer3d.org/search	2018	Human	Yes	Yes	Interactions
mutfunc [23]	http://www.mutfunc.com/	2018	Human, *E. coli, S. cerevisiae*	Yes	No	Interactions, stability, posttranscriptional modification, transcription factor binding, predicted mutation effect with SIFT
HUMA [24]	https://huma.rubi.ru.ac.za/	2018	Human	No	Yes	None
MISCAST [25]	https://miscast.broadinstitute.org/	2020	Human	No	No	Interactions + secondary structure assignment

complexes, where interface residues are defined as residues for which the area changes. Second, a relatively computationally expensive prediction of the mutation effect on protein structure stability using FoldX [33] is provided. Third, mutfunc includes additional analysis options: evolutionary conservation of the mutation, its potential effect on posttranslational modifications (PTM) predicted with MIMP [34], and predicted phenotypic effects using SIFT [14] are presented. Additionally, for mutations in non-coding regions, their predicted impact on transcription factor binding sites is provided. The mutfunc webserver enables fast queries of one or a few mutations, while the database is also completely downloadable making it useful for large-scale studies.

HUMA [24] is a hybrid method combining structural and sequence annotation by integrating many publicly available resources (dbSNP [7]; UniProt [31]; ClinVar [30]; and OMIM [18]), allowing for quick access and linking of multi-modal data. The structural annotation itself encompasses only structures of the target protein without considering structures of homologs or any modeling attempts. The position-wise mapping is provided, but the information on which structure covers which mutation is missing, so users have to scan manually through the annotated structures. Structural analysis is also not performed; instead, the authors introduce their own tool VAPOR, an integrated consensus method utilizing multiple mutation effect prediction methods: PolyPhen-2 [12], PROVEAN [35], PhD-SNP (single nucleotide polymorphism) [36], FATHMM-XF [37], I-Mutant 2.0 [38], and MUpro [39] (see Section 15.3 for a summary of this class of tools).

A recent addition to the palette of structural annotation databases is MISCAST, which focuses on the generation of features that can be used in machine-learning methods aimed to predict the effect of mutations (further discussed in Section 15.3). The size of the database, compared to the other approaches discussed here, is relatively small due to some incisive filtering. First, similar to HUMA, all proteins without an experimentally resolved structure in the PDB are omitted. Second, all proteins that do not contain any mutation from GnomAD [40], ClinVar [30], or HGMD [15] are also left out. There are few details given on how and if the structure is chosen when there are multiple structures available, or on how the results from the structural analysis are combined. The structural analysis includes identification of secondary structure elements and calculation of solvent accessible area for each residue using DSSP [41], identification of residues participating in interaction interfaces using PDBsum [42], and analysis of nonstructural features, such as physicochemical properties of mutated amino acid and protein function annotations.

Overall, structural annotation databases tend to focus on mutations that can be associated with diseases, which is of great practical importance, but this focus makes them unusable for the structural annotation and analysis of novel mutations. Here structural annotation pipelines take over.

Table 15.2 Pipelines for dynamic structural annotation.

Webserver	URL	Year	Species	Homologs	Isoforms	Structural analysis
MuPIT [43]	http://mupit.icm.jhu.edu/MuPIT_Interactive/	2013	Human	Yes	No	None
dSysMap [44]	https://dsysmap.irbbarcelona.org/	2015	Human	No	No	Protein–protein interactions and solvent accessibility
Mechismo [45]	http://mechismo.russelllab.org/	2015	Human, mouse, yeast, fruit fly, *C. elegans, E. coli, B. subtilis, Mycoplasma pneumoniae*	Yes	No	All interaction partners
PinSnps [46]	https://fraternalilab.kcl.ac.uk/PinSnps/	2016	Human	Yes	No	Protein–protein interactions and solvent accessibility
VarQ [47]	https://varq.qb.fcen.uba.ar/	2018	All	No	Yes	All interaction partners, solvent accessibility, change of stability, aggregability, conservation
VarMap/VarSite [48]	https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/DisaStr/GetPage.pl?varmap=TRUE	2019	Human	Yes	No	All interaction partners
PhyreRisk [49]	http://phyrerisk.bc.ic.ac.uk/	2019	Human	Yes	Yes	None

15.2.2 Dynamic Structural Annotation Pipelines

Similar to structural annotation databases, many pipelines were developed and replaced by successor pipelines over the years (Table 15.2). A number of pipelines are not available anymore: MarkUs [50], snp2structure [51], and G23D [52]. The websites of other pipelines are available but currently not functional: SAAPdap [20], PROSAT+ [53], GenProBis [54], and StructMAn [55].

The web server MuPIT dates back to 2013 but is still functional at the time of this writing. This pipeline is an extension of a now unavailable structural annotation database LS-SNP/PDB [56]. Users can directly enter genomic coordinates regardless of whether the given position belongs to a protein-coding region. The focus of the tool lies in its visualization of the annotated structures, which are obtained by sequence similarity search with BLAST among all proteins in the PDB. Hence, MuPIT can detect structures of homologous proteins, and performs the position-specific mapping by aligning sequences of the obtained proteins with BLAT [57]. Like many pipelines, MuPIT also incorporates a database to store the results of previous queries, thus supplementing the benefits of a pipeline with the benefits of a database, that is allowing for processing of arbitrary new queries and fast access to queries that have once already been processed. Unfortunately, other than visualization, there is no structural analysis performed.

A recent annotation pipeline VarMap [48] has a similar aim as MuPIT of mapping genomic coordinates to the protein sequences and structures. In this case, structural annotations are provided by VarSite [58], which was developed by the same authors and shares the same website. The structural analysis encompasses detecting interactions with all kinds of interaction partners (proteins, nucleic acids, small molecules, and ions), and the results are combined with the results of a comprehensive nonstructural analysis, including evolutionary conservation of each amino acid calculated by ScoreCons [59], functional annotations from CATH [60] and Pfam [61], and disease annotations taken from UniProt and ClinVar. The feature in the structural analysis of VarSite that distinguishes it from other tools is that it combines information obtained from different annotated structures. This results in a view listing all interactions that were detected in any structure in one single plot without requiring users to manually scan through the individual annotated structures. This aggregated results style is also maintained when mutations known to be associated with diseases are supplemented by reporting their participation in interactions. It is also reported in how many structures this interaction could be observed. VarSite cannot annotate alternative protein isoforms.

Extensive structural analysis is the focus of dSysMap [44], since its purpose is to analyze the effects of mutations in PPI interfaces on the scale of PPI networks. dSysMap does not only create a PPI network for given proteins but also places given individual mutations into the network based on their spatial location in the protein complexes. Individual mutations are structurally classified, and information from multiple external databases is gathered: Pfam [61], 3did [62], BIND [63], BioGRID

[64], DIP [65], HPRD [66], InnateDB [67], and IntAct [68]. Only proteins are considered as interaction partners in dSysMap, although alternative protein isoforms are not taken into account, either.

The concept of PinSnps [46] is similar to dSysMap, but it focuses less on the systems biology and more on the structural analysis of individual proteins and their complexes. Hence, the structural annotation pipeline works similarly, but the structural analysis in PinSnps includes additional features: solvent accessible area for every residue is calculated, and all residues are divided into those lying on the surface of the protein and those located in the core of the protein; posttranslationally modified residues and mutations known to be associated with diseases are also annotated. In comparison to dSysMap, PinSnps allows for more comprehensive analysis of individual mutations and proteins, while the systems biology aspect with its nonstructural interaction annotation and its network-style visualization of dSysMap still is superior.

Another structural annotation pipeline is Mechismo [45], whose focus is to predict whether a residue participates in an interaction while considering all kinds of interaction partners. For each protein, the experimentally resolved structures in the PDB are retrieved. To find experimentally resolved structures of proteins evolutionarily related to the given protein, a sequence similarity search is conducted among sequences of proteins in the other UniProt entries. The experimentally resolved structures associated with the entries resulting from this search are also considered for the structural analysis. Interactions are identified by distance: molecules closer than 5 Å are considered to form an interaction. To enhance prediction of interactions, Mechismo also includes the information from databases reporting experimentally identified interactions: BIND [63], BioGRID [64], IntAct [62], and MINT [69]. For the whole input sequence, Pfam [61] domains are retrieved and disordered regions are predicted by IUPred [70]. The predicted interactions are divided by the type of interaction partner into protein–protein, protein–chemical, and protein–DNA/RNA interactions. The protein–chemical interactions are also subdivided into organic, inorganic, and organometallic. The Mechismo webserver has low processing times, due to the fact that its outputs are based on a database of precomputed annotations for all variants in a given protein, even if only a handful of variants are queried by the user. The results of similarity searches are stored, and for all residues in all structures, the corresponding interaction types are precomputed. This limits the scope of possible queries and gives Mechismo a database-like character. But since it also stores intermediate steps and not just the end results, it can be easily expanded by adding precomputed data for other sequences. Currently, about 60 000 sequences of eight organisms are available (March 2021).

VarQ is another recently developed method that is more focused on individual positions [47]. The goal of VarQ is to assess the clinical relevance of a given nsSNV. For that purpose, it combines databases comprising data on clinical effects of mutations (dbSNP [7]; BioMuta [71]; humsavar [72]; and ClinVar [30]) with a custom structural annotation pipeline, for which only experimentally resolved structures of the query protein are considered. The structural analysis is performed on a single representative structure for each protein, but if different low molecular

weight ligands are bound in different structures, all such structures are analyzed. For selection of the representative structure, VarQ takes the structure that covers the largest part of the given protein and uses the resolution of a structure as a tie-breaker, when multiple structures have the same coverage. Structural analysis performed by VarQ is the most comprehensive to date in the field. The participation of the wild-type residue in a protein–protein interaction is checked with 3did [62]. The involvement of the residue in an active site is computed with fpocket [73]. The change in the protein structure stability introduced by the amino acid substitution is calculated with FoldX [33]. The relative solvent accessibility (RSA) of the residue is calculated to determine if the residue lies on the surface of a protein or is buried in the protein core. Tango [74] is used to estimate the tendency of the mutation to cause aggregation. Conservation of the wild-type and mutant amino acids is assessed using the frequency in the alignment of the corresponding Pfam family. The results of all performed analyses are reported individually. Since VarQ uses many computationally expensive methods, it is very slow and can only be used in case studies of preselected mutations. Further, the fact that 3D structures of homologs are not considered means that a lot of potentially useful data is omitted.

PhyreRisk [49] was developed to meet a different need. The burden of dynamic structural analysis was substituted by performing an automated protein structure modeling using the in-house homology-based modeling suite Phyre2 [75]. With their support of alternative protein isoforms, PhyreRisk is to date the only annotation method offering structural annotation of multiple protein isoforms, while combining structures of homologous proteins as well.

From the review above, it is obvious that one can perform structural analysis in many different ways, highlighting different aspects of protein interactions. One need that remains unmet to this day is a method that automatically performs a detailed structural analysis (e.g. like the one performed by VarQ or VarSite), but for a larger number of mutations. This analysis could be further used to generate feature databases for training machine-learning methods that predict the effect of mutations.

15.3 Methods for Predicting Effect of Protein Mutations

Predicting effects of a mutation on protein structure, function, and interactions is a notoriously difficult task. Not only can the scale of such effects vary, but also mutations in a protein tend to trigger long cascades of functional effects in the cell. Since there are so many aspects that have to be considered when predicting effects of mutations, many different approaches have been developed to this end in the last two decades. Due to the endless complexity of the effect cascades, there are no methods that attempt to predict them in every detail, therefore every method either concentrates on functional effects in a single protein or tries to predict the final outcome, for example, pathogenicity, without providing insights into the mechanisms behind it. We call such integral effects *phenotypic* and distinguish them from more easily interpretable mechanistic *functional* effects. In this section,

we focus again on mutations that cause changes in proteins encoded in the genome of the affected organism, and primarily on nsSNVs.

Methods that predict phenotypic effects of mutations are usually based on machine learning and their most useful features are derived thanks to the fact that all species on Earth are evolutionary related. One can see every life form as a successful outcome of its genome, and thus a mutation that changes a protein amino acid to one that can be found or is similar to the one found in another species can be regarded as harmless or neutral. The reason for this is that a variant that leads to a cascade with a fatal outcome cannot proliferate in evolution and thus will not be observed in other species. Thus, a variant that cannot be found in genomes of other species can be suspected to have a negative phenotypic effect (this speculation is, however, limited by the amount of sequencing data available from related species). This concept of evolutionary conservation has proven to have high predictive power for the prediction of effects of mutations.

The other types of methods that predict only one step among all biochemical changes that lead to a particular phenotype can also use machine learning but can include concepts that are less based on statistics, but rather on understanding of the underlying molecular biochemistry and directly estimating the mutation's effect on it. One such step is typically the effect of a mutation on particular interactions of proteins, on which we focus in Section 15.3.3.

15.3.1 Prediction of Phenotypic Effect

The field of tools predicting phenotypic effects of mutations expanded immensely over the last decades, and these days comprise tens, if not hundreds of methods [76, 77]. Due to its large size, we will not give a fully comprehensive analysis of the field but provide an overview focused on the most important and recent developments.

Protein function is exerted via interactions of proteins with their molecular counterparts, such as other proteins, nucleic acids, substrates, effectors, or ligands. As described in Section 15.2, mutations in particular protein regions can destroy some or all of these interactions and thus be detrimental to protein function. In case of proteins with crucial functions, this may lead to impaired phenotypes, for example, diseases. Mutations that happen on interaction interfaces (e.g. PPI interfaces) may selectively destroy interactions, and mutations that severely destabilize protein fold can abrogate all of them [78].

The fundamental insight for the field of predicting phenotypic effects was the discovery that evolutionary conservation of a protein position strongly correlates with its ability to harbor mutations that lead to a significant effect on the phenotype: the more conserved a position is in related proteins, the higher are the chances that mutations in it will be detrimental. At the same time, mutations to amino acids that can be found at the corresponding positions in related species, are more likely to be tolerated. One of the first and perhaps most successful model that was based on this concept was SIFT [14], whose predictive power was already great for the method being so simple. SIFT searches for sequences of homologs with PSI-BLAST [79] and

calculates a multiple sequence alignment (MSA) of them. Based on the amino acids that occupy the same position in the MSA as the given mutation, SIFT computes a position-specific scoring matrix (PSSM) and assesses the probability of the mutant amino acid being tolerated based on it.

Most modern methods for phenotypic effect prediction are supervised machine-learning methods that enhance the prediction by adding more and more features into the mix. PolyPhen-2 [12] uses a combination of evolutionary-based features and simple protein structure-based features to train a naive Bayes classifier. The evolutionary features in PolyPhen-2 are PSIC scores [80], a concept similar to PSSMs where related sequences are weighted by their similarity and take into account occurrences of both wild-type and mutant amino acids. The structural features here include RSA and crystallographic beta-factor, but do not consider potential interactions that can be distorted by the mutation. Other methods introduce prior knowledge from annotated databases into the models, e.g. SNAP [81] combines evolutionary features, simple structural features predicted from sequence (RSA, secondary structure assignment, and flexibility), and features derived from database lookups (Pfam [61] and Swiss-Prot [6] annotations) in a neural net. Other machine-learning techniques were also used: FATHMM-XF [37] uses hidden Markov models, CADD [82] uses logistic regression, DANN [83] uses a deep neural net, and M-CAP [84] uses a random forest.

In all these tools (SIFT, FATHMM-XF, CADD, and DANN) the protein 3D structure is not considered. When it is used, features based on protein structure are comparatively simple and/or are predicted from a sequence. The lack of experimentally resolved structures for many of the human proteins is a major obstacle for the integration of structural features, and thus the inclusion of more complex structural features can lead to the inapplicability of the method for a wide array of input scenarios. However, it has been shown [85] that the more complex structural features can increase the performance of prediction of clinical effects for cases, where they are available. Hence there is a paramount need to improve structural coverage of the human proteome, both with experimentally resolved structures and with high-quality models. It is likely that novel deep learning-based methods, such as AlphaFold [86], will partially satisfy this need, but many important structural features, such as interaction with nonprotein interaction partners, cannot, as of now, be predicted by these tools.

15.3.2 Estimation of Mutation Effects by Modeling Biophysical Properties of Proteins

Experimental methods that determine the stability of a protein structure or protein complex measure the Gibbs free energy ΔG of the folding process. The effect of a mutation on the stability of a structure can be described as the difference in ΔG values when comparing the wild-type and the mutant structures, and hence it is denoted as $\Delta\Delta G$, and is usually given in kcal/mol or kJ/mol. To save time and resources of expensive wet-lab experiments, computational methods for estimating ΔG of biomacromolecules have been developed. One well-known method is FoldX

[33], which calculates a ΔG estimate of a given protein structure. The concept behind FoldX is based on the combination of statistical, physical, and empirical energy functions, which are weighted and added to form a statistical potential to estimate Gibbs free energy. The key terms in this potential represent, for example, Coulomb interaction, steric clashes in the structure, and hydrogen bonds. FoldX is one of the few computational tools for estimating the change of stability that can analyze the change of stability of protein complexes upon mutations.

A more recently developed method attempts to account for protein flexibility in more detail. Flex ddG [87] is based on the Rosetta modeling suite [88] and applies structural optimization with distance constraints to the wild-type complex structure and the corresponding mutant model structure. The key ingredient that distinguishes Flex ddG from other comparable methods is a random sampling of complex structures by probing different backbone and side-chain torsion angles, resulting in fifty different structures of the same two initial complexes. The sampling is followed by energy minimization steps. For the final complex, a ΔG estimation is performed for each ensemble using the Rosetta potential and aggregated to return the $\Delta\Delta G$.

The most ambitious methodology to analyze the effect of a mutation on a PPI is the usage of molecular dynamics (MD) simulations. Since for every MD simulation setup, individual challenges and problems can occur, there is no tool automatizing the whole process. This makes MD simulation the most time-intensive way to study mutations that require extensive expert knowledge. The common approaches here include free-energy calculation protocols (FEP) [89], alchemical calculations [90, 91], and force-field-based MD toolkits [92], but here we will not go into the details of these methods. However, MD simulations are also the most comprehensive way that gives detailed insights into the molecular mechanisms resulting from a mutation, and thus provide invaluable insights into the impact of mutations on proteins and their interactions [93].

15.3.3 Prediction of Mechanistic Effects of Mutations on Interactions of Proteins

In recent years, many machine-learning methods have been developed to predict the change in binding affinity (expressed as $\Delta\Delta G$) of protein interactions, predominantly with other proteins, caused by a mutation (Table 15.3). The different methods differ in the specific machine-learning techniques that they employ, the nature of features that they use based on the given information, the way how they generate these features, and the level of information they expect as input.

SAAMBE-SEQ [99] and MuPIPR [100] only need the amino acid sequence of the wild-type protein and the mutation information, thus they do not use any features based on the 3D structure of the protein. This means that they can be applied in basically every case, but neglecting structural information is likely to decrease their predictive power. SAAMBE-SEQ generates a variety of features based on the sequence, such as the average position-specific scoring matrix (PSSM) value of the interface residues for both interaction partners, and the evolutionary conservation of the mutation, which is directly encoded as a feature vector. Further,

Table 15.3 Methods for predicting change of complex stability upon mutation.

Name	URL	Year	Input	Number of mutations per structure
BindProfX [94]	https://zhanggroup.org/BindProfX/	2016	Structure of the complex	Any number, if all on the interaction interface
iSEE [95]	https://github.com/haddocking/iSee	2019	Structure of the complex	One
mCSM-PPI2 [96]	http://biosig.unimelb.edu.au/mcsm_ppi2/	2019	Structure of the complex	One
MutaBind2 ([97], p. 2)	https://lilab.jysw.suda.edu.cn/research/mutabind2/	2020	Structure of the complex	Up to six
SAAMBE-3D [98]	http://compbio.clemson.edu/saambe_webserver/index3D.php	2020	Structure of the complex	One
SAAMBE-SEQ [99]	http://compbio.clemson.edu/saambe_webserver/indexSEQ.php	2020	Amino acid sequence	Any number
MuPIPR [100]	https://github.com/guangyu-zhou/MuPIPR	2020	Amino acid sequence	One

physicochemical properties of the substituted amino acids are used as features. These features are then used to train a random forest. MuPIPR, in comparison, is more restrained, since it uses only the sequence information as features. On the other hand, a more uniform feature set allows MuPIPR to employ deep neural nets as the machine-learning method.

Other tools require information on the 3D structure of the complex to produce $\Delta\Delta G$ estimations. Among them, BindProfX [94] is a special case, since it does not use a classical machine-learning method to create its statistical prediction model. The idea of this tool is similar to predicting mutation impact using evolutionary conservation, but here structural conservation is used instead. The authors collected a database of experimentally resolved interfaces, which is used to create multiple structural alignments of the interface of the PPI of interest. In this structural alignment, statistical methods similar to calculation of conservation in multiple sequence alignments are used to determine residues that are important for the interface and residues that are exchangeable. Since the structural alignment is done only for the

interface part of the complex, BindProfX can only predict $\Delta\Delta G$ values for mutations that lie directly at the interface.

All other methods require the structure of the PPI complex as input, which is then used to generate a whole new level of features, called structural features. The specific generation of structural features can differ largely among the methods. SAAMBE-3D [98] takes the same physicochemical property features as SAAMBE-SEQ and also encodes the sequence of the protein, but in this method, these features are computed only in a 10 amino acid-long window around the mutation. The structural features are based on the 10 nearest residues lying in a 10 Å sphere around the mutated residue. Further, some quality measures of the corresponding structure resolution experiment are also used as features. For the training of the prediction model, SAAMBE-3D uses gradient boosting as the machine-learning method. Another method that combines the evolutionary features with custom structural features is iSEE [95]. For the computation of the structural features, HADDOCK [101] is used to model the mutant structure, energetically minimize mutant and wild-type structures, and calculate intermolecular energy terms that can then be used as features. MutaBind2 [97] applies a similar strategy. Here, their first step is to model the structure of the mutant complex with FoldX [33] and then apply a number of subsequent steps to both the wild-type and the mutant structures. A short molecular dynamics simulation using the CHARMM36 [102] force field yields energetically minimized structures and then FoldX is applied to produce the energy terms. PROVEAN [35], a phenotypic effect prediction method, was used to generate the evolutionary features, emphasizing the similarity of the two problem settings. A special feature of MutaBind2 is the possibility to predict the effect of multiple mutations in the same protein complex. In MutaBind2 and iSEE, the machine learning part was implemented with a random forest, which allows for the calculation of the relative importance of individual features. This is an interesting way to analyze the model, which is also possible for the gradient-boosting model of SAAMBE-3D. For all three methods, the feature importance factors were calculated and analyzed. They all showed that the evolutionary features are more important for the prediction than the structural features.

The mCSM [103] family of methods addresses many types of prediction problems for which protein 3D structure may provide a useful source of information. They all are based on employing so-called graph-based signatures, structural features that can be extracted from residue interaction networks. Residue interaction networks are defined as graphs that can be calculated from the Cartesian coordinates of a protein structure and contain amino acid residues as nodes, while edges correspond to non-covalent contacts between them. Graph-based signatures are calculated by counting and categorizing all atomic contacts of the wild-type residue of the target mutation in the residue interaction network. In mCSM-PPI2, these signatures are combined with a comprehensive list of other features, including the evolutionary features and physicochemical properties of the substituted amino acids. Further, additional structural features are also computed using other structural analysis methods: FoldX for energy calculations, Bio3D [104] for the estimation of atomic

fluctuations, and Arpeggio [105] for the visual analysis of interacting residues. Interestingly, for none of the analyses, a mutant structure was modeled, all calculations are based exclusively on the wild-type structure.

A recently developed experimental technique called *deep-mutational scanning* (DMS) or *multiplexed arrays of variant effect* (MAVE) enables the training of models that can be thematically placed between the prediction of phenotypic and functional effects [106]. DMS experiments attempt to assess effects of as many mutations as possible on a particular protein function that can be measured in an *in vitro* assay. Cells containing mutated copies of the protein in question are selected based on this function, and afterward, the corresponding fragments of their genomes are sequenced. The proportion of a certain mutation in the sequenced product signifies whether the mutation was beneficial, detrimental, or neutral for the protein. DMS experiments measure the effects of mutations on protein function in a more general sense than just assessing the strength of their interactions, although of course mutations that distort these interactions would appear as deleterious in these screens. Hence, these experiments could be beneficial for training all types of prediction methods in the future. The greatest challenge that has to be overcome along this path is to combine and generalize differences of individual DMS experiments on different proteins. One approach to that aim was presented by Gray et al. in 2017 [107], in a study that created a mathematical framework to normalize the results of different DMS experiments into larger generalized datasets, which can then directly be used to train supervised machine-learning methods. Such a model, Envision, was also later constructed by the same group [108] based on gradient-boosted trees. The dataset that was used in this study contained only a limited number of proteins. For each of them, an experimentally resolved structure was available, which enabled the authors to use evolutionary features in combination with features directly generated from these structures. In general, structural features are more commonly used in models that predict the functional effects, as opposed to phenotypic effects (presented in more detail in Section 15.3.3), which means that the data from DMS experiments has the potential to become indispensable for training methods in that field.

15.4 Conclusion

The computational approaches amenable to assessing the effect of a mutation on protein interactions heavily depend on the available information about the corresponding protein complex. An experimentally resolved structure of the complex gives an irreplaceable advantage; thus the application of a structural annotation method as a first step is inevitable. The range of structural annotation web servers is broad, and to make a correct choice one has to pick a tool that supports the annotation specific for the set of mutations in question (e.g. corresponding to a certain disease). The structural analysis pipeline should detect interaction interfaces to find the structures of potential complexes. Further, one should gauge the number

of mutations that should be annotated and the depth of the provided structural analysis.

While some structural annotation methods provide the results of phenotypic effect prediction models, none of them are coupled with a prediction model that is specialized for predicting the effect on protein interactions. Presumably, the reason for this is that such analysis very much depends on the chosen protein complex structure, and structural annotation methods do not assign them specifically for this purpose. Another potential difficulty is that there are no tools for cases where only the experimentally resolved structures of the interacting proteins are available in different PDB entries in an unbound state. Protein–protein docking would be the required intermediate step, which remains a daunting task (see Chapter 4). A potential approach to skip the necessity for a protein–protein docking method would be training a machine-learning method on isolated structures or combining this with predicted binding interfaces (see Chapter 2).

The quality of models produced by machine learning heavily depends on the quality of the underlying training data. For the prediction of the effect of a mutation on protein interactions, in particular on protein–protein interactions, undoubtedly the most important and widely used dataset is Skempi 2.0 [109]. However, it is limited to only 345 protein complexes with 6187 mutations. Using this restricted dataset is recognized by the authors of the respective methods as a limitation and a source of possible bias in the models.

Acknowledgments

We are grateful to Nadezhda Azbukina and Dr. Vasily Ramensky for their critical reading of the draft and useful discussion. A.G. was funded by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant Sys_CARE [01ZX1908A]) and O.V.K. acknowledges support from the Klaus Faber Foundation.

References

1 The 1000 Genomes Project Consortium (2010). A map of human genome variation from population scale sequencing. *Nature* 467: 1061–1073. https://doi.org/10.1038/nature09534.

2 Ng, P.C. and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* 7: 61–80. https://doi.org/10.1146/annurev.genom.7.080505.115630.

3 Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* 80: 727–739. https://doi.org/10.1086/513473.

4 Berman, H.M., Westbrook, J., Feng, Z. et al. (2000). The protein data bank. *Nucleic Acids Res.* 28: 235–242.

5 Mooney, S.D. and Altman, R.B. (2003). MutDB: annotating human variation with functionally relevant data. *Bioinformatics* 19: 1858–1860. https://doi.org/10.1093/bioinformatics/btg241.

6 Yip, Y.L., Scheib, H., Diemand, A.V. et al. (2004). The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat.* 23: 464–470. https://doi.org/10.1002/humu.20021.

7 Sherry, S.T., Ward, M.-H., Kholodov, M. et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29: 308–311.

8 Altschul, S.F., Gish, W., Miller, W. et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

9 Singh, A., Olowoyeye, A., Baenziger, P.H. et al. (2008). MutDB: update on development of tools for the biochemical analysis of genetic variation. *Nucleic Acids Res.* 36: D815–D819. https://doi.org/10.1093/nar/gkm659.

10 Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28: 27–30.

11 Yue, P., Melamud, E., and Moult, J. (2006). SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinf.* 7: 166. https://doi.org/10.1186/1471-2105-7-166.

12 Adzhubei, I.A., Schmidt, S., Peshkin, L. et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7: 248–249. https://doi.org/10.1038/nmeth0410-248.

13 Ramensky, V., Peer, B.P., and Shamil, S.S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30: 3894–3900. https://doi.org/10.1093/nar/gkf493.

14 Ng, P.C. and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31: 3812–3814.

15 Stenson, P.D., Ball, E.V., Mort, M. et al. (2003). Human gene mutation database (HGMD®): 2003 update. *Hum. Mutat.* 21: 577–581. https://doi.org/10.1002/humu.10212.

16 Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353: 459–473. https://doi.org/10.1016/j.jmb.2005.08.020.

17 Webb, B. and Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* 86: 2.9.1–2.9.37. https://doi.org/10.1002/cpps.20.

18 Amberger, J.S., Bocchini, C.A., Schiettecatte, F. et al. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43: D789–D798. https://doi.org/10.1093/nar/gku1205.

19 Mottaz, A., David, F.P.A., Veuthey, A.-L., and Yip, Y.L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* 26: 851–852. https://doi.org/10.1093/bioinformatics/btq028.

20 Al-Numair, N.S. and Martin, A.C. (2013). The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC Genomics* 14: S4. https://doi.org/10.1186/1471-2164-14-S3-S4.

21 Hurst, J.M., McMillan, L.E.M., Porter, C.T. et al. (2009). The SAAPdb web resource: a large-scale structural analysis of mutant proteins. *Hum. Mutat.* 30: 616–624. https://doi.org/10.1002/humu.20898.

22 Sedova, M., Iyer, M., Li, Z. et al. (2019). Cancer3D 2.0: interactive analysis of 3D patterns of cancer mutations in cancer subsets. *Nucleic Acids Res.* 47: D895–D899. https://doi.org/10.1093/nar/gky1098.

23 Wagih, O., Galardini, M., Busby, B.P. et al. (2018). A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol. Syst. Biol.* 14: e8430. https://doi.org/10.15252/msb.20188430.

24 Brown, D.K. and Bishop, Ö.T. (2018). HUMA: a platform for the analysis of genetic variation in humans. *Hum. Mutat.* 39: 40–51. https://doi.org/10.1002/humu.23334.

25 Iqbal, S., Hoksza, D., Pérez-Palma, E. et al. (2020). MISCAST: MIssense variant to protein StruCture analysis web SuiTe. *Nucleic Acids Res.* 48: W132–W139. https://doi.org/10.1093/nar/gkaa361.

26 Porta-Pardo, E., Hrabe, T., and Godzik, A. (2015). Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res.* 43: D968–D973. https://doi.org/10.1093/nar/gku1140.

27 Barretina, J., Caponigro, G., Stransky, N. et al. (2012). The cancer cell line encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature* 483: 603–607. https://doi.org/10.1038/nature11003.

28 Weinstein, J.N., Collisson, E.A., Mills, G.B. et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45: 1113–1120. https://doi.org/10.1038/ng.2764.

29 Lek, M., Karczewski, K.J., Minikel, E.V. et al. (2016). Analysis of protein-coding genetic variation in 60, 706 humans. *Nature* 536: 285–291. https://doi.org/10.1038/nature19057.

30 Landrum, M.J., Lee, J.M., Benson, M. et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44: D862–D868. https://doi.org/10.1093/nar/gkv1222.

31 The UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43: D204–D212. https://doi.org/10.1093/nar/gku989.

32 Pieper, U., Eswar, N., Webb, B.M. et al. (2009). modbase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* 37: D347–D354. https://doi.org/10.1093/nar/gkn791.

33 Schymkowitz, J., Borg, J., Stricher, F. et al. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.* 33: W382–W388. https://doi.org/10.1093/nar/gki387.

34 Wagih, O., Reimand, J., and Bader, G.D. (2015). MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Methods* 12: 531–533. https://doi.org/10.1038/nmeth.3396.

35 Choi, Y. and Chan, A.P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31: 2745–2747. https://doi.org/10.1093/bioinformatics/btv195.

36 Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22: 2729–2734. https://doi.org/10.1093/bioinformatics/btl423.

37 Rogers, M.F., Shihab, H.A., Mort, M. et al. (2018). FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* 34: 511–513. https://doi.org/10.1093/bioinformatics/btx536.

38 Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33: W306–W310. https://doi.org/10.1093/nar/gki375.

39 Cheng, J., Randall, A., and Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 62: 1125–1132. https://doi.org/10.1002/prot.20810.

40 Karczewski, K.J., Francioli, L.C., Tiao, G. et al. (2020). The mutational constraint spectrum quantified from variation in 141, 456 humans. *Nature* 581: 434–443. https://doi.org/10.1038/s41586-020-2308-7.

41 Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577–2637. https://doi.org/10.1002/bip.360221211.

42 Laskowski, R.A., Jabłońska, J., Pravda, L. et al. (2018). PDBsum: structural summaries of PDB entries. *Protein Sci.* 27: 129–134. https://doi.org/10.1002/pro.3289.

43 Niknafs, N., Kim, D., Kim, R.G. et al. (2013). MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. *Hum. Genet.* 132: 1235–1243. https://doi.org/10.1007/s00439-013-1325-0.

44 Mosca, R., Tenorio-Laranga, J., Olivella, R. et al. (2015). dSysMap: exploring the edgetic role of disease mutations. *Nat. Methods* 12: 167–168. https://doi.org/10.1038/nmeth.3289.

45 Betts, M.J., Lu, Q., Jiang, Y. et al. (2015). Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucleic Acids Res.* 43: e10. https://doi.org/10.1093/nar/gku1094.

46 Lu, H.-C., Herrera Braga, J., and Fraternali, F. (2016). PinSnps: structural and functional analysis of SNPs in the context of protein interaction networks. *Bioinformatics* 32: 2534–2536. https://doi.org/10.1093/bioinformatics/btw153.

47 Radusky, L., Modenutti, C., Delgado, J. et al. (2018). VarQ: a tool for the structural and functional analysis of human protein variants. *Front. Genet.* 9: https://doi.org/10.3389/fgene.2018.00620.

48 Stephenson, J.D., Laskowski, R.A., Nightingale, A. et al. (2019). VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics* 35: 4854–4856. https://doi.org/10.1093/bioinformatics/btz482.

49 Ofoegbu, T.C., David, A., Kelley, L.A. et al. (2019). PhyreRisk: a dynamic web application to bridge genomics, proteomics and 3D structural data to guide interpretation of human genetic variants. *J. Mol. Biol.* 431: 2460–2466. https://doi.org/10.1016/j.jmb.2019.04.043.

50 Fischer, M., Zhang, Q.C., Dey, F. et al. (2011). MarkUs: a server to navigate sequence–structure–function space. *Nucleic Acids Res.* 39: W357–W361. https://doi.org/10.1093/nar/gkr468.

51 Wang, D., Song, L., Singh, V. et al. (2015). SNP2Structure: a public and versatile resource for mapping and three-dimensional modeling of missense SNPs on human protein structures. *Comput. Struct. Biotechnol. J.* 13: 514–519. https://doi.org/10.1016/j.csbj.2015.09.002.

52 Solomon, O., Kunik, V., Simon, A. et al. (2016). G23D: Online tool for mapping and visualization of genomic variants on 3D protein structures. *BMC Genomics* 17: https://doi.org/10.1186/s12864-016-3028-0.

53 Stank, A., Richter, S., and Wade, R.C. (2016). ProSAT+: visualizing sequence annotations on 3D structure. *Protein Eng. Des. Sel.* 29: 281–284. https://doi.org/10.1093/protein/gzw021.

54 Konc, J., Skrlj, B., Erzen, N. et al. (2017). GenProBiS: web server for mapping of sequence variants to protein binding sites. *Nucleic Acids Res.* 45: W253–W259. https://doi.org/10.1093/nar/gkx420.

55 Gress, A., Ramensky, V., Büch, J. et al. (2016). StructMAn: annotation of single-nucleotide polymorphisms in the structural context. *Nucleic Acids Res.* 44: W463–W468. https://doi.org/10.1093/nar/gkw364.

56 Ryan, M., Diekhans, M., Lien, S. et al. (2009). LS-SNP/PDB: annotated non-synonymous SNPs mapped to protein data bank structures. *Bioinformatics* 25: 1431–1432. https://doi.org/10.1093/bioinformatics/btp242.

57 Kent, W.J. (2002). BLAT—The BLAST-like alignment tool. *Genome Res.* 12: 656–664. https://doi.org/10.1101/gr.229202.

58 Laskowski, R.A., Stephenson, J.D., Sillitoe, I. et al. (2020). VarSite: disease variants and protein structure. *Protein Sci.* 29: 111–119. https://doi.org/10.1002/pro.3746.

59 Valdar, W.S.J. (2002). Scoring residue conservation. *Proteins* 48: 227–241. https://doi.org/10.1002/prot.10146.

60 O'Leary, N.A., Wright, M.W., Brister, J.R. et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44: D733–D745. https://doi.org/10.1093/nar/gkv1189.

61 Bateman, A., Coin, L., Durbin, R. et al. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32: D138–D141. https://doi.org/10.1093/nar/gkh121.

62 Mosca, R., Céol, A., Stein, A. et al. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* 42: D374–D379. https://doi.org/10.1093/nar/gkt887.

63 Bader, G.D., Betel, D., and Hogue, C.W.V. (2003). BIND: the biomolecular interaction network database. *Nucleic Acids Res.* 31: 248–250.

64 Oughtred, R., Stark, C., Breitkreutz, B.-J. et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47: D529–D541. https://doi.org/10.1093/nar/gky1079.

65 Xenarios, I., Rice, D.W., Salwinski, L. et al. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res.* 28: 289–291.

66 Peri, S., Navarro, J.D., Kristiansen, T.Z. et al. (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* 32: D497–D501. https://doi.org/10.1093/nar/gkh070.

67 Lynn, D.J., Winsor, G.L., Chan, C. et al. (2008). InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.* 4: 218. https://doi.org/10.1038/msb.2008.55.

68 del-Toro, N., Duesbury, M., Koch, M. et al. (2019). Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set. *Nat. Commun.* 10: 10. https://doi.org/10.1038/s41467-018-07709-6.

69 Chatr-aryamontri, A., Ceol, A., Palazzi, L.M. et al. (2007). MINT: the molecular INTeraction database. *Nucleic Acids Res.* 35: D572–D574. https://doi.org/10.1093/nar/gkl950.

70 Mészáros, B., Erdős, G., and Dosztányi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46: W329–W337. https://doi.org/10.1093/nar/gky384.

71 Dingerdissen, H.M., Torcivia-Rodriguez, J., Hu, Y. et al. (2018). BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery. *Nucleic Acids Res.* 46: D1128–D1136. https://doi.org/10.1093/nar/gkx907.

72 Famiglietti, M.L., Estreicher, A., Gos, A. et al. (2014). Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum. Mutat.* 35: 927–935. https://doi.org/10.1002/humu.22594.

73 Schmidtke, P., Le Guilloux, V., Maupetit, J., and Tufféry, P. (2010). fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* 38: W582–W589. https://doi.org/10.1093/nar/gkq383.

74 Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* 22: 1302–1306. https://doi.org/10.1038/nbt1012.

75 Kelley, L.A., Mezulis, S., Yates, C.M. et al. (2015). The Phyre2 web portal for protein modelling, prediction and analysis. *Nat. Protoc.* 10: 845–858. https://doi.org/10.1038/nprot.2015.053.

76 Cheng, N., Li, M., Zhao, L. et al. (2020). Comparison and integration of computational methods for deleterious synonymous mutation prediction. *Briefings Bioinf.* 21: 970–981. https://doi.org/10.1093/bib/bbz047.

77 Hassan, M.S., Shaalan, A.A., Dessouky, M.I. et al. (2019). A review study: computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. *Gene* 680: 20–33. https://doi.org/10.1016/j.gene.2018.09.028.

78 Sahni, N., Yi, S., Taipale, M. et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161: 647–660. https://doi.org/10.1016/j.cell.2015.04.013.

79 Altschul, S.F., Madden, T.L., Schäffer, A.A. et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

80 Sunyaev, S.R., Eisenhaber, F., Rodchenkov, I.V. et al. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* 12: 387–394. https://doi.org/10.1093/protein/12.5.387.

81 Bromberg, Y. and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 35: 3823–3835. https://doi.org/10.1093/nar/gkm238.

82 Rentzsch, P., Witten, D., Cooper, G.M. et al. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47: D886–D894. https://doi.org/10.1093/nar/gky1016.

83 Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31: 761–763. https://doi.org/10.1093/bioinformatics/btu703.

84 Jagadeesh, K.A., Wenger, A.M., Berger, M.J. et al. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* 48: 1581–1586. https://doi.org/10.1038/ng.3703.

85 Dehiya, V., Thomas, J., and Sael, L. (2018). Impact of structural prior knowledge in SNV prediction: towards causal variant finding in rare disease. *PLoS One* 13: https://doi.org/10.1371/journal.pone.0204101.

86 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873): 583–589.

87 Barlow, K.A., Conchúir, S.O., Thompson, S. et al. (2018). Flex ddG: rosetta ensemble-based estimation of changes in protein-protein binding affinity upon mutation. *J. Phys. Chem. B* 122: 5389–5399. https://doi.org/10.1021/acs.jpcb.7b11367.

88 Ovchinnikov, S., Kinch, L., Park, H. et al. (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* 4: e09248. https://doi.org/10.7554/eLife.09248.

89 Wang, L., Wu, Y., Deng, Y. et al. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* 137: 2695–2703. https://doi.org/10.1021/ja512751q.

90 Crooks, G.E. (1999). Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E* 60: 2721–2726. https://doi.org/10.1103/PhysRevE.60.2721.

91 Jarzynski, C. (1997). A nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* 78: 2690–2693. https://doi.org/10.1103/PhysRevLett.78.2690.

92 Pronk, S., Páll, S., Schulz, R. et al. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29: 845–854. https://doi.org/10.1093/bioinformatics/btt055.

93 Galano-Frutos, J.J., García-Cebollada, H., and Sancho, J. (2021). Molecular dynamics simulations for genetic interpretation in protein coding regions: where we are, where to go and when. *Briefings Bioinf.* 22: 3–19. https://doi.org/10.1093/bib/bbz146.

94 Xiong, P., Zhang, C., Zheng, W., and Zhang, Y. (2017). BindProfX: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *J. Mol. Biol.* 429: 426–434. https://doi.org/10.1016/j.jmb.2016.11.022.

95 Geng, C., Vangone, A., Folkers, G.E. et al. (2019). iSEE: interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins* 87: 110–119. https://doi.org/10.1002/prot.25630.

96 Rodrigues, C.H.M., Myung, Y., Pires, D.E.V., and Ascher, D.B. (2019). mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* 47: W338–W344. https://doi.org/10.1093/nar/gkz383.

97 Zhang, N., Chen, Y., Lu, H. et al. (2020). MutaBind2: predicting the impacts of single and multiple mutations on protein–protein interactions. *iScience* 23: https://doi.org/10.1016/j.isci.2020.100939.

98 Pahari, S., Li, G., Murthy, A.K. et al. (2020). SAAMBE-3D: predicting effect of mutations on protein–protein interactions. *Int. J. Mol. Sci.* 21: 2563. https://doi.org/10.3390/ijms21072563.

99 Li, G., Pahari, S., Murthy, A.K. et al. (2020). SAAMBE-SEQ: a sequence-based method for predicting mutation effect on protein–protein binding affinity. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa761 992–999.

100 Zhou, G., Chen, M., Ju, C.J.T. et al. (2020). Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR Genomics Bioinf.* 2: https://doi.org/10.1093/nargab/lqaa015.

101 van Zundert, G.C.P., Rodrigues, J.P.G.L.M., Trellet, M. et al. (2016). The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol* 428: 720–725. https://doi.org/10.1016/j.jmb.2015.09.014.

102 MacKerell, A.D., Bashford, D., Bellott, M. et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102: 3586–3616. https://doi.org/10.1021/jp973084f.

103 Pires, D.E.V., Ascher, D.B., and Blundell, T.L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30: 335–342. https://doi.org/10.1093/bioinformatics/btt691.

104 Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M. et al. (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 22: 2695–2696. https://doi.org/10.1093/bioinformatics/btl461.

105 Jubb, H.C., Higueruelo, A.P., Ochoa-Montaño, B. et al. (2017). Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.* 429: 365–371. https://doi.org/10.1016/j.jmb.2016.12.004.

106 Fowler, D.M. and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nat. Methods* 11: 801–807. https://doi.org/10.1038/nmeth.3027.

107 Gray, V.E., Hause, R.J., and Fowler, D.M. (2017). Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. *Genetics* 207: 53–61. https://doi.org/10.1534/genetics.117.300064.

108 Gray, V.E., Hause, R.J., Luebeck, J. et al. (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst.* 6: 116–124.e3. https://doi.org/10.1016/j.cels.2017.11.003.

109 Jankauskaitė, J., Jiménez-García, B., Dapkūnas, J. et al. (2019). SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* 35: 462–469. https://doi.org/10.1093/bioinformatics/bty635.

16

Not Quite the Same: How Alternative Splicing Affects Protein Interactions

Zakaria Louadi[1,2], Olga Tsoy[2], Jan Baumbach[2,3], Tim Kacprowski[4,5], and Markus List[1]

[1] Technical University of Munich, TUM School of Life Sciences Weihenstephan, Chair of Experimental Bioinformatics, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany
[2] University of Hamburg, Institute for Computational Systems Biology, Notkestrasse 9, 22607 Hamburg, Germany
[3] Institute of Mathematics and Computer Science, University of Southern Denmark, Department of Mathematics and Computer Science, Campusvej 55, 5000 Odense M, Denmark
[4] Peter L. Reichertz Insitute for Medical Informatics of TU Braunschweig and Hannover Medical School, Division Data Science in Biomedicine, Rebenring 56, 38106 Braunschweig, Germany
[5] TU Braunschweig, Braunschweig Integrated Centre for Systems Biology (BRICS), Rebenring 56, 38106 Braunschweig, Germany

List of Abbreviations

AS	alternative splicing
IDR	intrinsically disordered regions
PPI	protein–protein interactions
SLiM	short linear motif

16.1 Introduction

Eukaryotic genes consist of *exons* and *introns*. During *splicing*, a ribonucleoprotein complex – the *spliceosome* – recognizes splice sites at exon–intron boundaries, joins exons together, and removes introns. *Alternative splicing* (AS) refers to different combinations of exons and introns, where four main types of events are usually considered: an exon could be skipped, an intron could be retained, and alternative 5′ or 3′-splice sites could be chosen. At least 95% of genes with more than one exon undergo AS [1, 2]. For the remainder of this chapter, we will refer to the products of AS as *"transcript variants"* when describing the transcript level; and as "*protein isoforms*" or "*isoforms*" when describing the protein level. As a result of AS, genes can potentially produce thousands of transcript variants and isoforms with

Joint First Authors: Zakaria Louadi and Olga Tsoy
Joint Last Authors: Tim Kacprowski and Markus List

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

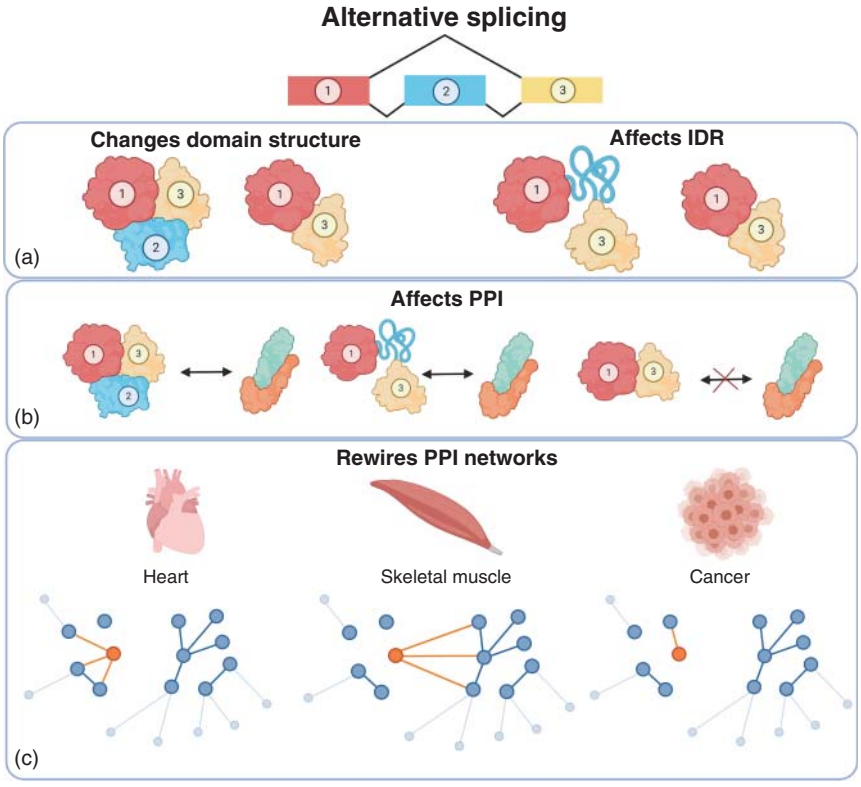


Figure 16.1 The impact of AS on protein structures, PPIs, and PPI networks. (a) The isoforms from the same gene can vary in their domain structure or IDR. (b) Such isoforms could have different PPIs. (c) Since isoforms could be tissue specific, AS could rewire PPI networks. Created with BioRender.com.

widely differing or even opposite functions. AS of caspase-3, for example, generates both pro- and antiapoptotic isoforms [3]. In most cases, only the function of the major isoform is known (if at all), which leads to a considerable knowledge gap concerning the functional repertoire of the transcriptome and proteome. However, the true impact of AS on protein function is heavily debated. In this chapter, we will first explore how AS affects individual proteins and their interactions with other proteins due to changes in binding domains and subsequently consider the systematic perturbations that AS brings to the protein–protein interaction (PPI) network (Figure 16.1).

One theory states that most transcript variants and protein isoforms operate in the organism in a tissue- or condition- or time-specific manner, while another theory is that these variants merely represent stochastic noise of the splicing machinery. The proponents of the stochastic noise theory suggest that although transcript variants are detectable, they do not encode functional proteins, as some evidence exists for the lack of important functional regions, incomplete structures, and structural instability of the resulting isoforms [4–6]. Abascal et al.

analyzed eight large-scale proteomic experiments from 100 different tissues, cell lines, and developmental stages [7]. By integrating eight proteomics datasets and after conservative filtering of peptide quality, they found evidence for 12 716 protein-coding genes but only 236 of them gave rise to more than one isoform. The authors suggested that genes mostly have one main protein isoform, and AS does not contribute much to proteome complexity. However, it is important to note that the limited number of genes with more than one isoform found in the study might be due to the limitations of proteomic technologies. This leads to the main critique of the stochastic theory as detailed below.

The proponents of the AS functionality argue that the classical proteomic approach lacks the necessary level of coverage and sensitivity to detect all isoforms [8]. The emerging proteomic techniques might support the role of AS for proteome complexity: Liu et al. captured the quantitative effect of AS at the proteomic level using a SWATH-MS technology, which combines deep proteome coverage and quantitative accuracy of targeted protein profiling. They compared the transcriptome and proteome before and after mutation of PRPF8, a component of the spliceosome, and identified 3370 transcripts with differential splicing (1284 with differential exon usage, 1449 with intron retention, 637 with both) and 1542 proteins with at least one peptide with differential protein expression. Such transcripts and proteins were enriched in the same functional categories (translation, RNA splicing, mitotic cell cycle, and ubiquitination) suggesting that changes in AS at the transcriptomic level are functionally mirrored at the proteomic level [9]. The other argument is that isoforms are difficult to detect because they could appear only in a particular tissue, condition, or developmental stage [8]. The further development of proteomic technologies and the increase in large-scale data from different conditions will help to answer how AS contributes to the proteome complexity.

Individual experiments demonstrate that AS can change the protein structure and function substantially. For instance, AS modulates the activity of transcription factors, changes the properties of enzymes, changes the subcellular localization of proteins, affects kinetics and sensitivity of ion channels, and disrupts PPIs [10–12]. We review how AS affects protein structure, potentially leading to a loss or gain of interacting partners in Section 16.2. The condition-specific (e.g. tissue- or developmental stage-specific) nature of AS events suggest the rewiring of PPIs [10, 12]. In system biology, PPIs are modeled as networks and used to study biological pathways and mechanisms. Because of difficulties in the current methods for detecting interactions, these networks usually represent only gene-level interactions and ignore isoforms. This simplified approach neglects the complexity of the proteome due to AS and, thus, does not live up to the full potential of systems biology for the study of PPI network rewiring. We will discuss the importance of isoforms for PPI analysis in more detail in Section 16.3. Ideally, all possible interactions between the isoforms of two genes should be tested experimentally. Affinity purification-mass spectrometry is a popular approach for identifying PPI in a high-throughput manner but it generates a high amount of nonspecific interactions [13]. On the other hand, experiments for individual proteins are more accurate but require a prohibitive amount of resources to perform on a large scale. To investigate how AS rewires PPIs in a

particular tissue, condition, or developmental stage, we need a systematic computational approach. We will discuss the benefits and limitations of currently emerging approaches, in Section 16.3.3 and provide our view on further development of this field.

16.2 Effects of Alternative Splicing on Individual Proteins

16.2.1 Alternative Splicing and Protein Structure

AS could potentially disrupt the protein structure, therefore, not all parts of proteins might be equally prone to AS. Several studies have demonstrated that isoforms often only differ in a small number of residues, preserving the domain structure [14, 15]. Wang et al. examined alternatively spliced sequences from a 3D structural point of view and found that alternative splicing mostly affects residues at the protein surface. More specifically, alternatively spliced exons were enriched in coiled regions of the secondary structure [16]. These observations lead to the hypothesis that AS does not dramatically impact the folding and structure of proteins, but rather leads to changes on the surface. Since these changes modulate the affinities to other proteins or ligands, they can, nevertheless, have a functional impact. In some cases, however, AS leads to gain or loss of core parts of a protein. For instance, Sulakhe et al. compared protein features of alternatively spliced isoforms with those of their canonical isoform [17]. They found that 42% of the spliced-out regions in transporters genes are present in a transmembrane domain and 41% of the spliced protein features in the cell cycle proteins correspond to short sequence motifs. Similarly, DNA-binding regions represent 37% of the spliced-out features for DNA-binding proteins. In such cases, the loss or gain introduced by AS can significantly impact the function of their corresponding genes.

16.2.2 Alternative Splicing and Intrinsically Disordered Regions

Intrinsically disordered regions (IDR) are protein segments that lack tertiary structure. Such regions exist in any state from fully ordered to disordered, contributing to protein diversity [18]. They differ in charge, amino acid composition, length, and conformation state. This affects the protein structure, function, and interactions with other proteins, DNA, RNA, and ligands. Therefore, IDRs affect, among other things, signal transduction and molecular recognition [19, 20]. First, auto-inhibition modules (or inhibitory modules) of proteins are often disordered. These modules compete with other biomolecules for interaction with their own protein [21]. Second, IDRs can change their structure and adapt to different interaction partners [22]. Molecular Recognition Features (MoRFs) are a subgroup of IDRs that can become ordered upon binding [22]. Finally, IDRs are enriched with posttranslational modifications [23] and *linear motifs* – short stretches of amino acids (3–10 aa) with a distinct function (e.g. posttranslational modification, binding, localization, and degradation signals, see also Chapter 17) [24]. They tend to locate

within IDRs and accessible protein surfaces [25]. Linear motifs are referred to in the literature as short linear motifs (SLiMs) or eukaryotic linear motifs (ELMs): we will further refer to them as SLiM (not to confuse with the ELM database [26]).

AS of IDRs preserves protein structure while diversifying protein function. For example, proteins of the G protein-coupled receptor (GPCR) family, comprising the extracellular N-terminus, the intracellular C-terminus, seven transmembrane domains, three intracellular loops, and three extracellular loops, have a high proportion of IDRs. The N-, C-termini, and one of the intracellular loops (ICL3) are predicted to be the most disordered and show most AS events and posttranslational modifications. The isoforms resulting from these events have different binding properties. The canonical and the short isoform (with deletion of 29 residues from ICL3) of D2 dopamine receptor 2 (D2R2), for example, bind to different alpha subunits of G protein and activate distinct signaling pathways. Another example is EDNRB which has isoforms without palmitoylation site that cannot activate G-proteins [20].

Colak et al. [27] divided IDRs into "constrained" ones that preserve amino acid sequences in evolution and "flexible" ones that do not. They showed that such a "flexible disorder" allows proteins to have different isoforms without structural disruption; while a "constrained disorder" presents a scaffold for SLiM and posttranslational modifications.

The computational analysis of IDR falls into two groups: prediction of IDR from a protein sequence and analysis of PPIs due to in- or exclusion of IDRs. The first group has been extensively reviewed in several articles [28, 29]; we will focus on the second group of tools and databases.

IDRs affect protein–protein interactions through several mechanisms and each of these motivates dedicated tools and/or databases. Several tools predict MoRFs based on scoring functions (ANCHOR [30]) or different machine learning techniques (e.g. neural networks in en_DCNNMoRF [31]; support vector machines in fMoRF-pred [32]) (Table 16.1). DisoRDPbin [38] predicts residues within IDRs that bind to other proteins, DNA or RNA. The information about experimentally validated and predicted IDRs are stored in dedicated databases (Table 16.2). DIBS [51] represents the largest set of experimentally validated MoRFs. DistProt [52] is a database of manually curated IDRs and their interacting partners. D2P2 [50] collects predictions of disordered proteins. ModiDB [55] collects both experimentally validated and predicted IDRs and related features from various sources. Mutual folding induced by binding (MFIB) [54] database is a repository for protein complexes that are formed exclusively by intrinsically disordered proteins.

Another mechanism for IDRs to affect PPIs is through SLiMs. The in- or exclusion of SLiMs in protein isoforms adds novel functions and introduces new interaction partners including posttranslational modifiers. For example, the proapoptotic member of the Bcl-2 family Bim has several isoforms and one of them, BimS, lacks the dynein-binding motif and differs in its ability to sequester the microtubule dynein complex [57, 58]. The analysis of experimentally validated SLiMs from the ELM [26] database showed that alternatively spliced exons are enriched with particular SLiMs that bind to PDZ, PTB, SH2, and WW domains [59]. Buljan

Table 16.1 The list of tools for investigating the role of IDR and SLiM in PPIs.

Name	Publication	Link	Description
ANCHOR	[30]	http://anchor.enzim.hu	MoRFs prediction based on scoring
en_DCNNMoRF	[31]	http://vivace.bi.a.u-tokyo.ac.jp:8008/fang/home5.html	MoRFs prediction based on convolutional neural networks
fMoRFpred	[32]	http://biomine.cs.vcu.edu/webresults/fMoRFpred/20190527084228/results.html	MoRFs prediction based on support vector machine
IUPred2A	[33]	https://iupred2a.elte.hu/	IDR and MoRFs prediction
MORFchibi	[34]	https://gsponerlab.msl.ubc.ca/software/morf_chibi/	MoRFs prediction based on support vector machine
MoRFPred	[35]	http://biomine.cs.vcu.edu/servers/MoRFpred/	MoRFs prediction based on support vector machine
OPAL+	[36]	https://github.com/roneshsharma/OPAL-plus/wiki/OPAL-plus-Download	MoRFs prediction based on support vector machine
Predict-MoRFs	[37]	https://github.com/roneshsharma/Predict-MoRFs	MoRFs prediction based on support vector machine
DisoRDPbind	[38]	http://biomine.cs.vcu.edu/servers/DisoRDPbind/	Binding residues prediction
DILIMOT	[39]	http://dilimot.russelllab.org/	SLiM search
DSTAR	[40]	https://www.comp.nus.edu.sg/~bioinfo/hugowill/DSTAR.html	Correlated motifs from PPI networks

iELM	[41]	http://i.elm.eu.org/search/	SLiM prediction
iSPOT	[42]	http://cbm.bio.uniroma2.it/ispot/	SLiMs bound to specific domains
MotifCluster	[43]	https://alse.cs.hku.hk/motif_pair/	Motif pairs from PPI networks
MOTIPS	[44]	http://motips.gersteinlab.org/	Predicts protein-domain targets
Scansite	[45]	https://scansite4.mit.edu/4.0/#home	SLiM search
SlimFinder	[46]	http://www.slimsuite.unsw.edu.au/servers/slimfinder.php	SLiM prediction in a group of proteins
SLimPred	[47]	http://bioware.ucd.ie/~compass/biowareweb/Server_pages/slimpred.php	SLiM prediction
SlimProb	Not published	http://www.slimsuite.unsw.edu.au/servers/slimprob.php	SLiM search in protein sequences
SlimSearch	[48]	http://slim.icr.ac.uk/slimsearch/	SLiM search in a proteome

Table 16.2 The list of databases of IDRs and SLiMs.

Name	Publication	Link	Description
ADAN	[49]	http://adan-embl.ibmc.umh.es/default.asp	Database for prediction of PPI of modular domains mediated by SLiMs
D2P2	[50]	http://d2p2.pro/about	Database of integrated IDR predictions
DIBS	[51]	http://dibs.enzim.ttk.mta.hu/help.php	Database of validated MoRFs
DistProt	[52]	https://www.disprot.org/about	Database of manually curated IDR, including interacting partners
ELM	[26]	http://elm.eu.org/	Database of SLiMs
LMPID	[53]	http://bicresources.jcbose.ac.in/ssaha4/lmpid/	Database of SLiMs mediated PPIs
MFIB	[54]	http://mfib.enzim.ttk.mta.hu/	Database of disordered protein complexes
MobiDB	[55]	https://mobidb.bio.unipd.it/about/mobidb	Database of IDR including binding residues prediction
Prosite	[56]	https://prosite.expasy.org/	Database of protein features, including SLiMs

et al. showed that tissue-specific alternatively spliced exons are also enriched in SLiMs [23].

Usually, only 2–5 amino acids are involved in weak transient binding of SLiMs [24]. Two mechanisms can modulate the binding properties: first, post-translational modifications change the binding properties of a SLiM; second, the repetition of SLiMs also modulates specificity and affinity of PPIs [59]. SLiMs could also overlap and form a molecular switch. Posttranslational modifications drive the choice of a SLiM that depends on the biological context: the condition, the oligomeric state, the cell type, or tissue. The CYT-1 isoform of receptor tyrosine-protein kinase ErbB4, for example, has an alternatively spliced exon with a WW domain-binding motif ($PPAY_{1056}$) and an overlapping SH2 domain binding motif ($YTPM_{1059}$). The phosphorylation state of the overlapping tyrosine determines the interacting partner of the protein [58, 60].

Computational analysis of SLiMs is mostly focused on their search or prediction, and most tools for short motif search could be applied here, e.g. MEME [61]. Specialized tools are listed in Table 16.1 and described below. SlimProb searches for known

SLiMs in protein sequences; SlimSearch [48] does the same but proteome-wide. SLiMPred [47], iELM [41], DILIMOT [39], and SlimFinder [46] utilize different approaches to predict potential SLiMs in a protein sequence: SlimFinder and DILIMOT use the overrepresentation of sequence motifs in a group of proteins; iELM uses hidden Markov models trained on manually annotated interactions mediated by SLiMs; SLiMPred uses neural networks. Scansite [45] and iSPOT [42] search SLiMs that are bound to specific domains: Scansite predicts binding to SH2, 14-3-3, and PDZ domains; iSPOT – to SH3, PDZ, and WW domains. MOTIPS [44] predicts binding partners for specific protein domains. MotifCluster [43] and D-STAR [40] suggest algorithms to analyze PPIs to find statistically significant co-occurring pairs of interacting SLiMs, therefore, trying to identify novel SLiMs. SLiM databases are listed in Table 16.2. The largest maintained database in this field – ELM [26] – stores experimentally validated SLiMs independent of their function. The last 2020 release contains 3542 instances of SLiMs from 10 species. LMPID [53] collects only those SLiMs that mediate PPIs. The last release in 2015 included 1762 SLiM instances from 2215 PPIs. Prosite [56] is a general protein motif database but also includes information about SLiMs. ADAN [49] stores information about protein and ligand interactions of binding domains mediated by SLiMs. The last update was in 2009 and contains 3502 entries.

16.3 Effects of Alternative Splicing on Protein–Protein Interaction Networks

16.3.1 Alternative Splicing Rewires Protein–Protein Interactions

Most of the time, proteins do not function individually but rather as a group with interactions between proteins as well as other molecules. The current practice of presenting PPI networks on a one-protein-per-gene level does not capture the complexity of the proteome. As discussed in [62], two main problems result from this simplified approach that often only tests for interactions of major isoforms. First (Figure 16.2b), false-negative PPI: an interaction between two genes could be missed if an interacting isoform was not tested. Second (Figure 16.2a), a false-positive PPI: an experimentally verified interaction between two genes may be specific to a subset of isoforms 16.2. Since AS could be specific to particular tissue, condition, or developmental stage, the consequence of these false-positive and false-negative PPIs can dramatically bias results in network enrichment analysis.

While one might be tempted to dismiss this as a rare issue, recent experimental studies suggest that isoform-specific interactions represent around half of the interactions annotated for genes [11]. Moreover, isoforms tend to behave as functionally distinct proteins rather than minor variants. Colak et al. [27] suggested that tissue-specific alternatively spliced exons encode significantly more often for IDRs compared to alternatively spliced exons found in several tissues. Moreover, Buljan et al. [10] found that tissue-specific protein-coding exons are enriched in IDRs that contain binding motifs. The in- or exclusion of these binding motifs by tissue-specific

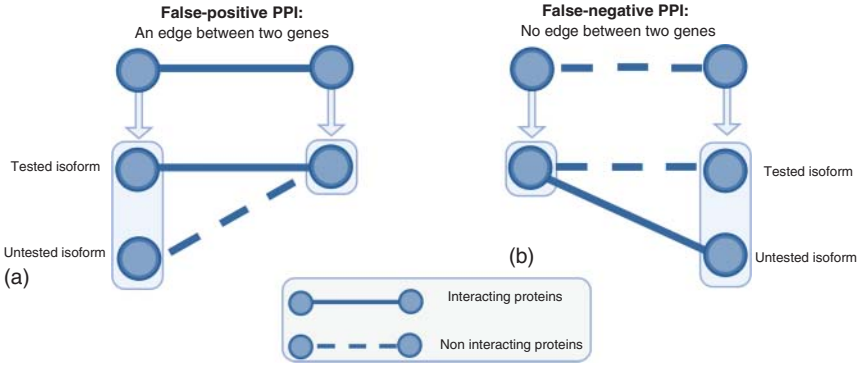


Figure 16.2 Presenting PPI networks on a gene level might lead to (a) a false-positive protein–protein interaction; (b) a false-negative protein–protein interaction. Created with BioRender.com.

splicing can, thus, be expected to rewire interaction networks and modulate signaling pathways. Interestingly, proteins with tissue-specific alternatively spliced IDRs and SLiMs occupy central positions in the PPI network and tend to have more interacting partners than proteins with common alternatively spliced exons found across tissues. Also, their interaction partners differ strongly and rarely overlap between a pair of tissues [23]. In a concrete example, Ellis et al. reported that neural-specific exons are enriched in regions of protein with conserved interaction surfaces, and about a third of them affect PPIs [12]. These findings highlight the importance of considering AS in studying PPIs.

16.3.2 Alternative Splicing in Diseases

Altered isoform–isoform interactions play a role in diseases, for example, cancer. Climente-Gonza et al. found that abnormal splicing in cancer affects domain families that mediate PPIs [63]. These protein domains are also frequently mutated in tumors with a negative correlation between the occurrence of AS and somatic mutations in driver genes. As a consequence, protein domains are often lost with functional consequences that are not apparent at the level of gene expression. Alternative splicing-related changes could hence represent important oncogenic processes that are easily missed in classical gene expression analysis. Kahraman et al. studied isoform switches (the changes of the dominantly expressed transcript variants between cancer and healthy samples) in 27 cancer types. They showed that 20% of isoform switches disrupt PPIs [64]. Their analysis shows that tumor samples with mutations in the spliceosomal complex have a higher number of switches of the dominant transcript, while the expression of only few transcripts was correlated with somatic mutation in *cis*. Kataka et al. analyzed The Cancer Genome Atlas RNA-Seq datasets using the PPIXpress tool [65], generating 642 patient-specific pairs of interactomes, i.e. one for the tumor and one for tumor-adjacent negative controls. In their analysis, they formulate the problem as a differential analysis of the pairs of interactomes to identify patient-specific edgetic perturbation [66].

The perturbed edges refer to the interactions specific to only a set of isoforms from a gene. They found that a set of these perturbations is shared between different patients and highly correlated with survival.

Autism spectrum disorders were also linked to changes in AS [67] based on study of experimentally validated interactions between multiple brain-specific isoforms from nearly 200 genes. In this study Corominas et al. constructed a network, named Autism Spliceform Interaction Network, and revealed new physical associations between genes from pathogenic autism copy number variations, making it more relevant for autism spectrum disorder studies than gene-focused PPI networks. Furthermore, many of the newly discovered interactions were part of related neurodevelopmental disorders pathways. Such a network is a successful example of an isoform-centered view of the interactome serving as a rich resource for neurobiology and drug development.

16.3.3 Resources for Studying the Effect of Alternative Splicing on Protein–Protein Interactions

Separate resources exist to study AS (e.g. APPRIS [68]) and to study PPI (e.g. STRING [69], BIOGRID [70]). Only a few tools and databases (Table 16.3) couple this information and address the impact of AS on the PPIs. Even less do so at a network level.

AS can lead to domain exclusion or inclusion; and in case of interacting domains leads to alterations in PPI. Several databases collect domain–domain interactions. DOMMINO [79] uses domains annotated by SCOP or predicted by an HMM-based approach. The last 2012 release of DOMMINO contains around 200 000 domain–domain interactions among other molecular interactions. 3did contains 14 278 domain–domain interactions based on PFAM domains for which high-resolution 3D structures are available. KBDOCK operates with PFAM domains and 3D structures but specializes in the spatial structure of domain-domain interactions. DIMA integrated experimental and computational resources to form a comprehensive domain–domain interaction database [77].

Other resources provide direct annotations of individual interacting domains or exonic regions involved in AS events. Two databases, ExonOnthology and ExonSkipDB [80, 81], store information about the functional role of alternatively spliced exons. ExonSkipDB collected exon skipping events from GTEx and TCGA [86, 87] and predicted the possible subsequent loss of protein function, including the loss of interactions. ExonOnthology exploits several computational resources and provides a functional description for an exon of interest, including the disordered state, its possible effect on PPI, and posttranslational modifications. The AS-ALPS database collected AS regions where the amino acid sequence is changed by AS [72]. It coupled the knowledge about AS of this region with the information about interactions extracted from the Protein Data Bank (PDB [88]). This information allows AS-ALPS to infer whether AS impairs protein structure and interactions. IRview collects "interacting regions" and for some of them reported observed variants [82]. In the last 2011 release, this database contained 3417 experimentally validated

Table 16.3 The list of resources for investigating the effect of AS on PPIs.

Name	Direct impact on PPI	Publication	Link	Description
3did	—	[71]	http://3did.irbbarcelona.org	Database of domain–domain interactions from PFAM, with 3D structure
AS-ALPS	—	[72]	https://as-alps.nagahama-i-bio.ac.jp/	Database of AS regions
CompleXChange	✓	[73]	https://sourceforge.net/projects/complexchange/	An extension of PPIxpress for differential analysis of protein complexes
CORUM	—	[74]	http://mips.helmholtz-muenchen.de/corum/	Database of protein complexes. The last update included complexes with isoforms
DIGGER	✓	[75]	https://exbio.wzw.tum.de/digger/	Database for isoform-specific and exon-specific interactions. The resource also allows the user to construct a subnetwork of the PPI corresponding to a set of isoforms
DIIP	✓	[76]	http://bioinfo.lab.mcgill.ca/resources/diip	Database for isoform interactions identified from the domain-mediated interactions
DIMA	—	[77]	http://webclu.bio.wzw.tum.de/dima2/index.jsp	The integrated database of conserved domain–domain interactions based on experimental and computational resources
DomainGraph	✓	[78]	https://domaingraph.bioinf.mpi-inf.mpg.de/	Method for the analysis and visualization of exon inclusion/exclusion on domain-mediated interactions and miRNA binding sites
DOMMINO	—	[79]	http://korkinlab.org/dommino	Database of domain–domain interactions from SCOP and predicted by HMM
Exon Ontology	—	[80]	http://fasterdb.ens-lyon.fr/ExonOntology/	Database of alternatively spliced exon function

ExonSkipDB	—	[81]	https://ccsm.uth.edu/ExonSkipDB/	Database of alternatively spliced exons from GTEx and TCGA
IRview	—	[82]	http://ir.hgc.jp/	Database of experimentally validated protein interacting regions and their variants
KBDOCK	—	[83]	https://kbdock.loria.fr/	Database of spatial organization of domain–domain interactions
PPICompare	✓	[84]	https://sourceforge.net/projects/ppicompare/	Based on PPIXpress and extended to compare between two different conditions. The tool identifies significant rewired interactions between the grouped samples
PPIXpress	✓	[65]	https://sourceforge.net/projects/ppixpress/	Constructs condition-specific PPIs from transcript expression data by identifying the interaction of the major expressed isoform. PPIexpress filters the rest of the interactions with nodes of low expressed genes
VastDB	—	[85]	http://vastdb.crg.eu/	Database of AS events, including overlap with protein domains and disordered regions

(using *in vitro* virus and yeast two-hybrid system techniques) interacting regions from human and mouse together with their functional characteristics (InterPro domains, non-synonymous single nucleotide polymorphisms, and variant regions). Vast-DB [85] stores the results of the comprehensive AS profiles of 308 RNA-Seq datasets from human, mouse, and chicken and provides general information about identified AS events including overlap with protein domains, IDRs, and mappings to the protein structure.

Only a few databases report isoform–isoform interactions. DIIP [76] combines domain–domain interactions and PPIs and infers the existence or absence of an interaction between isoforms based on the in- or exclusion of interacting domains. The CORUM database collects manually curated protein complexes [74]. The last update included 58 protein complexes with isoforms, such as the CASP-2S-fodrin complex where only one isoform of CASP interacts with fodrin [89]. This subset from the CORUM database mostly includes isoforms that are associated with diseases or alter function of a protein complex. One database – IIIDB – is currently not available but the idea behind it might contribute to the possible future resources for AS analysis. IIIDB used domain–domain interactions to predict PPI and extended this approach by adding the co-expression of interacting isoforms [90].

While the above methods aimed to reduce the number of false positives among the current PPIs by trying to identify isoform-specific interactions and filter non-interacting isoforms from the network, DomainGraph was the first resource to explore the exon rather than the domain contributions to PPIs [78]. This approach offers the advantage that the impact of exon expression can be studied on a systems level [78]. DomainGraph combines domain–domain interactions with protein–protein interactions to first identify the domains mediating the interaction between two proteins. Subsequently, these domains are mapped to exons. In this way, differential exon usage yields insights about the edgetic changes in the PPI network. This idea was later extended to identify missing interactions in annotated isoforms and known PPIs [75, 76]. PPIXpress [65] constructs a condition-specific PPI on a whole-proteome scale based on transcript expression. The intuition is that interaction should not exist if one of the interacting isoforms in the edge is missing or downregulated even if the two genes are expressed [65, 75]. Combined with RNA-Seq profiles, such methods could extend our understanding of tissue-specific regulation and disease mechanisms related to AS. PPICompare was later built on the output of PPIXpress to extend its functionality for differential analysis of PPIs [84]. The idea is to construct one network for each condition and to compare them with respect to rewiring resulting from isoform switch events. The authors applied PPICompare to different blood development stages. Later on, similar algorithms were extended by the same authors to include differential analysis of protein complexes in their tool CompleXChange [73].

Recently, we developed DIGGER, a database and a web tool for fully exploring the impact of AS at different levels [75]. DIGGER offers the features provided by existing tools such as PPIXpress as a user-friendly database. In addition to PPIs and domain–domain interactions, DIGGER also considers residue-specific interactions inferred from co-resolved protein structures. Users can switch from an

isoform-centric view of the interactome to an exon-centric one and interactively study the impact of AS on both levels.

16.4 Conclusion and Future Work

AS affects protein sequences, hence interfering with protein function and interaction potential. We described mechanisms of how AS affects PPIs via in- or exclusion of interacting domains or binding motifs. The latter are often found within IDRs and do not affect the protein structure, emphasizing the importance of these regions for modulating interactions. As isoforms are often tissue-specific, AS appears to play an important role in rewiring PPIs depending on the biological context. Importantly, PPI rewiring (or in such case, disruption) might also drive pathological conditions, such as cancer or neurological disorders.

Despite the current efforts to understand the functional impact of AS on PPIs and pathways, a few issues remain unaddressed. Isoforms remain difficult to detect in proteome analysis and consequently, functional studies are limited to considering the transcriptome. Here, the majority of RNA-Seq analyses are performed on gene or transcript level using short-read sequencing technologies, where often only gene-level results are considered, ignoring the consequences of AS altogether. Furthermore, the latest benchmarks show that transcript-level RNA-Seq analysis cannot accurately quantify AS events [91]. Importantly, such methods can only detect known transcripts and, as a result, they underestimate the impact and the variation of *de novo* or rare AS events. Methods such as MAJIQ or leafcutter [92, 93], which are tailored toward detecting such *de novo* events are not yet part of most transcriptome analysis pipelines and deserve more attention. Emerging new technologies, such as long-read sequencing and single-cell technologies, lead to more noisy data but offer unique opportunities to reconstruct full-length transcripts and to study cell-type specific effects. Available resources for studying the consequences of AS on PPI networks suffer from the relatively low structural coverage of the interactome. In the best-case scenario, only 20% of the PPIs can be mapped to domain–domain interactions [65, 75] and even with the inclusion of residue-specific information from the PDB, this percentage does not increase significantly [75]. To tackle this issue in the absence of experimental data, few efforts aimed to approach the problem as a supervised machine learning problem [94] with mixed success even though similar ideas were previously applied for other resources such as GO term databases [95]. A logical next step for the field is the systematic integration of AS analysis and the analysis of tissue-specific disordered regions as it will yield tissue-specific PPI networks that are better tailored for knowledge discovery and downstream applications such as network enrichment [96] and drug repurposing [97].

In summary, it is important to switch to an isoform-centered view of the interactome to capture the central role of AS in tissue regulation and to capture the dynamic changes of the proteome in health and disease.

References

1 Pan, Q., Shai, O., Lee, L.J. et al. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40 (12): 1413–1415.

2 Wang, E.T., Sandberg, R., Luo, S. et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456 (7221): 470–476.

3 Vegran, F., Boidot, R., Oudin, C. et al. (2005). Implication of alternative splice transcripts of caspase-3 and survivin in chemoresistance. *Bull. Cancer* 92 (3): 219–226.

4 Tress, M.L., Martelli, P.L., Frankish, A. et al. (2007). The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl. Acad. Sci. U.S.A.* 104 (13): 5495–5500.

5 Melamud, E. and Moult, J. (2009). Structural implication of splicing stochastics. *Nucleic Acids Res.* 37 (14): 4862–4872.

6 Tress, M.L., Abascal, F., and Valencia, A. (2017). Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci* 42 (2): 98–110.

7 Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J. et al. (2015). Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level. *PLoS Comput. Biol.* 11 (6): e1004325.

8 Blencowe, B.J. (2017). The relationship between alternative splicing and proteomic complexity. *Trends Biochem. Sci* 42 (6): 407–408.

9 Liu, Y., Gonzà Lez-Porta, M., Santos, S. et al. (2017). Impact of alternative splicing on the human proteome in brief resource impact of alternative splicing on the human proteome. *Cell Rep.* 20: 1229–1241.

10 Buljan, M., Chalancon, G., Eustermann, S. et al. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol. Cell* 46 (6): 871–883.

11 Yang, X., Coulombe-Huntington, J., Kang, S. et al. (2016). Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164 (4): 805–817.

12 Ellis, J.D., Barrios-Rodiles, M., Çolak, R. et al. (2012). Tissue-specific alternative splicing remodels protein–protein interaction networks. *Mol. Cell* 46 (6): 884–892.

13 Rao, V.S., Srinivasa Rao, V., Srinivas, K. et al. (2014). Protein–protein interaction detection: methods and analysis. *Int. J. Proteomics* 2014: 1–12.

14 Romero, P.R., Zaidi, S., Fang, Y.Y. et al. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl. Acad. Sci. U.S.A.* 103 (22): 8390–8395.

15 Ezkurdia, I., Rodriguez, J.M., Carrillo-De Santa Pau, E. et al. (2015). Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.* 14 (4): 1880–1887.

16 Wang, P., Yan, B., Guo, J.-T. et al. (2005). Structural genomics analysis of alternative splicing and application to isoform structure modeling. *Proc. Natl. Acad. Sci. U.S.A.* 102 (52): 18920–18925.

17 Sulakhe, D., D'Souza, M., Wang, S. et al. (2019). Exploring the functional impact of alternative splicing on human protein isoforms using available annotation sources. *Briefings Bioinf.* 20 (5): 1754–1768.

18 Dunker, A.K., Lawson, J.D., Brown, C.J. et al. (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* 19 (1): 26–59.

19 Wright, P.E. and Dyson, H.J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16 (1): 18–29.

20 Zhou, J., Zhao, S., and Dunker, A.K. (2018). Intrinsically disordered proteins link alternative splicing and post-translational modifications to complex cell signaling and regulation. *J. Mol. Biol.* 430 (16): 2342–2359.

21 Trudeau, T., Nassar, R., Cumberworth, A. et al. (2013). Structure and intrinsic disorder in protein autoinhibition. *Structure* 21 (3): 332–341.

22 Hsu, W.-L., Oldfield, C.J., Xue, B. et al. (2012). Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Science* 22 (3): 258–273.

23 Buljan, M., Chalancon, G., Dunker, A.K. et al. (2013). Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Curr. Opin. Struct. Biol.* 23 (3): 443–450.

24 Davey, N.E., Van Roey, K., Weatheritt, R.J. et al. (2012). Attributes of short linear motifs. *Mol. Biosyst.* 8 (1): 268–281.

25 Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23 (8): 950–956.

26 Kumar, M., Gouw, M., Michael, S. et al. (2020). ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.* 48 (D1): D296–D306.

27 Colak, R., Kim, T.H., Michaut, M. et al. (2013). Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS Comput. Biol.* 9 (4): e1003030.

28 Liu, Y., Wang, X., and Liu, B. (2017). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings Bioinf.* 20 (1): 330–346.

29 Katuwawala, A. and Kurgan, L. (2020). Comparative assessment of intrinsic disorder predictions with a focus on protein and nucleic acid-binding proteins. *Biomolecules* 10 (12): 1636.

30 Dosztányi, Z., Mészáros, B., and Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* 25 (20): 2745–2746.

31 Fang, C., Moriwaki, Y., Tian, A. et al. (2019). Identifying short disorder-to-order binding regions in disordered proteins with a deep convolutional neural network method. *J. Bioinform. Comput. Biol.* 17 (1): 1950004.

32 Yan, J., Dunker, A.K., Uversky, V.N., and Kurgan, L. (2016). Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.* 12 (3): 697–710.

33 Mészáros, B., Erdos, G., and Dosztányi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46 (W1): W329–W337.

34 Malhis, N., Wong, E.T.C., Nassar, R., and Gsponer, J. (2015). Computational identification of MoRFs in protein sequences using hierarchical application of bayes rule. *PLoS One* 10 (10): e0141603.

35 Disfani, F.M., Hsu, W.-L., Mizianty, M.J. et al. (2012). MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* 28 (12): i75–i83.

36 Sharma, R., Sharma, A., Raicar, G. et al. (2019). OPAL+: length-specific MoRF prediction in intrinsically disordered protein sequences. *Proteomics* 19 (6): e1800058.

37 Sharma, R., Kumar, S., Tsunoda, T. et al. (2016). Predicting MoRFs in protein sequences using HMM profiles. *BMC Bioinf.* 17 (Suppl 19): 504.

38 Peng, Z. and Kurgan, L. (2015). High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.* 43 (18): e121.

39 Neduva, V., Linding, R., Su-Angrand, I. et al. (2005). Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.* 3 (12): e405.

40 Tan, S.-H., Hugo, W., Sung, W.-K., and Ng, S.-K. (2006). A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinf.* 7: 502.

41 Weatheritt, R.J., Jehl, P., Dinkel, H., and Gibson, T.J. (2012). iELM–a web server to explore short linear motif-mediated interactions. *Nucleic Acids Res.* 40 (Web Server issue): W364–W369.

42 Brannetti, B. and Helmer-Citterich, M. (2003). iSPOT: A web tool to infer the interaction specificity of families of protein modules. *Nucleic Acids Res.* 31 (13): 3709–3711.

43 Leung, H.C.-M., Siu, M.-H., Yiu, S.-M. et al. (2009). Clustering-based approach for predicting motif pairs from protein interaction data. *J. Bioinform. Comput. Biol.* 7 (4): 701–716.

44 Lam, H.Y.K., Kim, P.M., Mok, J. et al. (2010). MOTIPS: automated motif analysis for predicting targets of modular protein domains. *BMC Bioinf.* 11: 243.

45 van de Kooij, B., Creixell, P., van Vlimmeren, A. et al. (2019). Comprehensive substrate specificity profiling of the human Nek kinome reveals unexpected signaling outputs. *eLife* 8: e44635.

46 Davey, N.E., Haslam, N.J., Shields, D.C., and Edwards, R.J. (2010). SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. *Nucleic Acids Res.* 38 (Web Server issue): W534–W539.

47 Mooney, C., Pollastri, G., Shields, D.C., and Haslam, N.J. (2012). Prediction of short linear protein binding regions. *J. Mol. Biol.* 415 (1): 193–204.

48 Davey, N.E., Haslam, N.J., Shields, D.C., and Edwards, R.J. (2011). SLiMSearch 2.0: biological context for short linear motifs in proteins. *Nucleic Acids Res.* 39 (Web Server issue): W56–W60.

49 Encinar, J.A., Fernandez-Ballester, G., Sánchez, I.E. et al. (2009). ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics* 25 (18): 2418–2424.

50 Oates, M.E., Romero, P., Ishida, T. et al. (2013). D^2P^2: database of disordered protein predictions. *Nucleic Acids Res.* 41 (Database issue): D508–D516.

51 Schad, E., Fichó, E., Pancsa, R. et al. (2018). DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* 34 (3): 535–537.

52 Hatos, A., Hajdu-Soltész, B., Monzon, A.M. et al. (2020). DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* 48 (D1): D269–D276.

53 Sarkar, D., Jana, T., and Saha, S. (2015). LMPID: a manually curated database of linear motifs mediating protein-protein interactions. *Database* 2015.

54 Fichó, E., Reményi, I., Simon, I., and Mészáros, B. (2017). MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics* 33 (22): 3682–3684.

55 Piovesan, D., Necci, M., Escobedo, N. et al. (2021). MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* 49 (D1): D361–D367.

56 Sigrist, C.J.A., de Castro, E., Cerutti, L. et al. (2013). New and continuing developments at PROSITE. *Nucleic Acids Res.* 41 (Database issue): D344–D347.

57 Puthalakath, H., Huang, D.C., O'Reilly, L.A. et al. (1999). The proapoptotic activity of the Bcl-2 family member Bim is regulated by interaction with the dynein motor complex. *Mol. Cell* 3 (3): 287–296.

58 Weatheritt, R.J. and Gibson, T.J. (2012). Linear motifs: lost in (pre)translation. *Trends Biochem. Sci* 37 (8): 333–341.

59 Weatheritt, R.J., Davey, N.E., and Gibson, T.J. (2012). Linear motifs confer functional diversity onto splice variants. *Nucleic Acids Res.* 40 (15): 7123–7131.

60 Sundvall, M., Veikkolainen, V., Kurppa, K. et al. (2010). Cell death or survival promoted by alternative isoforms of ErbB4. *Mol. Biol. Cell* 21 (23): 4275–4286.

61 Bailey, T.L., Boden, M., Buske, F.A. et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37 (Web Server issue): W202–W208.

62 Talavera, D., Robertson, D.L., and Lovell, S.C. (2013). Alternative splicing and protein interaction data sets. *Nat. Biotechnol.* 31 (4): 292–293.

63 Ctor Climente-Gonzá Lez, H., Porta-Pardo, E., Godzik, A. et al. (2017). The functional impact of alternative splicing in cancer. *Cell Rep.* 20: 2215–2226.

64 Kahraman, A., Karakulak, T., Szklarczyk, D., and von Mering, C. (2020). Pathogenic impact of transcript isoform switching in 1,209 cancer samples covering 27 cancer types using an isoform-specific interaction network. *Sci. Rep.* 10 (1): 14453.

65 Will, T. and Helms, V. (2016). PPIXpress: construction of condition-specific protein interaction networks based on transcript expression. *Bioinformatics* 32 (4): 571–578.

66 Kataka, E., Zaucha, J., Frishman, G. et al. (2020). Edgetic perturbation signatures represent known and novel cancer biomarkers. *Sci. Rep.* 10 (1): 1–16.

67 Corominas, R., Yang, X., Lin, G.N. et al. (2014). Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nat. Commun.* 5: 3650.

68 Rodriguez, J.M., Rodriguez-Rivas, J., Di Domenico, T. et al. (2018). APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.* 46 (D1): D213–D217.

69 Jensen, L.J., Kuhn, M., Stark, M. et al. (2009). STRING 8 – a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* 37 (Database issue): D412–D416.

70 Oughtred, R., Stark, C., Breitkreutz, B.-J. et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47 (D1): D529–D541.

71 Mosca, R., Céol, A., Stein, A. et al. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* 42 (Database issue): D374–D379.

72 Shionyu, M., Yamaguchi, A., Shinoda, K. et al. (2009). AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Res.* 37 (Database issue): D305–D309.

73 Will, T. and Helms, V. (2019). Differential analysis of combinatorial protein complexes with CompleXChange. *BMC Bioinf.* 20 (1): 300.

74 Giurgiu, M., Reinhard, J., Brauner, B. et al. (2019). CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* 47 (D1): D559–D563.

75 Louadi, Z., Yuan, K., Gress, A. et al. (2020). DIGGER: exploring the functional role of alternative splicing in protein interactions. *Nucleic Acids Res.* 49 (D1): D309–D318.

76 Ghadie, M.A., Lambourne, L., Vidal, M., and Xia, Y. (2017). Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing. *PLoS Comput. Biol.* 13 (8): e1005717.

77 Pagel, P., Oesterheld, M., Tovstukhina, O. et al. (2008). DIMA 2.0–predicted and known domain interactions. *Nucleic Acids Res.* 36 (Database issue): D651–D655.

78 Emig, D., Salomonis, N., Baumbach, J. et al. (2010). AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res.* 38 (Web Server issue): W755–W762.

79 Kuang, X., Han, J.G., Zhao, N. et al. (2012). DOMMINO: a database of macromolecular interactions. *Nucleic Acids Res.* 40 (Database issue): D501–D506.

80 Tranchevent, L.-C., Aubé, F., Dulaurier, L. et al. (2017). Identification of protein features encoded by alternative exons using exon ontology. *Genome Res.* 27 (6): 1087–1097.

81 Kim, P., Yang, M., Yiya, K. et al. (2020). ExonSkipDB: functional annotation of exon skipping event in human. *Nucleic Acids Res.* 48 (D1): D896–D907.

82 Fujimori, S., Hirai, N., Masuoka, K. et al. (2012). Irview: a database and viewer for protein interacting regions. *Bioinformatics* 28 (14): 1949–1950.

83 Ghoorah, A.W., Devignes, M.-D., Smaïl-Tabbone, M., and Ritchie, D.W. (2014). KBDOCK 2013: a spatial classification of 3D protein domain family interactions. *Nucleic Acids Res.* 42 (Database issue): D389–D395.

84 Will, T. and Helms, V. (2017). Rewiring of the inferred protein interactome during blood development studied with the tool PPICompare. *BMC Syst. Biol.* 11 (1): 44.

85 Tapial, J., Ha, K.C.H., Sterne-Weiler, T. et al. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res.* 27 (10): 1759–1768.

86 Lonsdale, J., Thomas, J., Salvatore, M. et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45 (6): 580–585.

87 Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19 (1A): A68–A77.

88 Berman, H.M., Westbrook, J., Feng, Z. et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1): 235–242.

89 Han, C., Zhao, R., Kroger, J. et al. (2013). Caspase-2 short isoform interacts with membrane-associated cytoskeleton proteins to inhibit apoptosis. *PLoS One* 8 (7): e67033.

90 Tseng, Y.T., Li, W., Chen, C.H. et al. (2015). IIIDB: a database for isoform-isoform interactions and isoform network modules. *BMC Genomics* 16 (Suppl 2): 1–7.

91 Merino, G.A., Conesa, A., and Fernández, E.A. (2019). A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. *Briefings Bioinf.* 20 (2): 471–481.

92 Li, Y.I., Knowles, D.A., Humphrey, J. et al. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50 (1): 151–158.

93 Vaquero-Garcia, J., Barrera, A., Gazzara, M.R. et al. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* 5: e11752.

94 Narykov, O., Johnson, N., and Korkin, D. (2018). Determining rewiring effects of alternatively spliced isoforms on protein-protein interactions using a computational approach. *bioRxiv* 256834.

95 Chen, H., Shaw, D., Zeng, J. et al. (2019). DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning. *Bioinformatics* 35 (14): i284–i294.

96 Lazareva, O., Lautizi, M., Fenn, A. et al. (2020). Multi-omics analysis in a network context. In: *Systems Medicine*, 224–233. Academic Press.

97 Sadegh, S., Matschinske, J., Blumenthal, D.B. et al. (2020). Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nature Communications* 11 (1): 1–9.

17

Phosphorylation-Based Molecular Switches

Attila Reményi

Institute of Organic Chemistry, Biomolecular Interactions Research Group, Research Center for Natural Sciences, H-1117 Budapest, Hungary

17.1 Introduction

There are several posttranslational modifications (PTM), such as phosphorylation, ubiquitination, acetylation, methylation, glycosylation, lipidation, and proteolysis. These chemical modifications on proteins have the ability to regulate activity, affect interactions with other molecules, or influence cellular localization and thus local concentration. PTMs happen fast compared to gene expression level-based mechanisms involving transcriptional regulation. They may be reversible or irreversible and could be categorized based on the chemical nature of the group attached to amino acid side chains: peptide adducts (ubiquitination and sumoylation), small chemical groups (acetylation, methylation, phosphorylation), more complex molecules (glycosylation, isoprenylation), or protein cleavage by proteolysis.

Protein phosphorylation is probably the most widespread posttranslational regulatory mechanism. Modification of amino acids (e.g. aspartate, histidine, but most prominently serine/threonine or tyrosine) by kinases plays a key role in signal transduction or in controlling gene expression. Why is phosphorylation so widespread and effective? First, protein phosphorylation requires the transfer of the γ-phosphoryl group from ATP to an amino acid side chain, and since ATP is abundant and is constantly restocked, the source is readily at hand. Second, the phosphate group on proteins is highly loaded: it has two negative charges that could have large effects on the chemical nature of the phosphorylated protein region [1].

Phosphorylation of the alkyl/aryl hydroxyl group of a serine, threonine, or tyrosine residue results in the formation of an ester bond, which is chemically stable under neutral pH. Phosphorylation on six more amino acids is also chemically feasible: on cysteine by forming a phosphorothiolate bond, on aspartate or glutamate forming a mixed anhydride, or on histidine, arginine or lysine forming a phosphoramidate. However, these latter are chemically more labile in aqueous solution under physiological conditions and are not as widely used as alkyl/aryl hydroxy amino acids.

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

Histidine and aspartate phosphorylation in so-called two-component systems elicits a quick response to extracellular signals since the half-life of histidine phosphoramidate is short and the mixed anhydride bond of phospho-aspartate also spontaneously hydrolyzes. This ancient signaling system, comprising a membrane-bound histidine kinase and the cytoplasmic response regulator modified on aspartate by the former, is frequently used in bacteria, but it is less frequent in eukaryotes, albeit it is present in yeast and is quite common in plants [2]. Due to chemical stability, serine/threonine and tyrosine phosphorylation is experimentally more tractable, and the phosphorylation of these amino acids has been far better studied. This chapter henceforth will discuss serine/threonine and tyrosine phosphorylation.

There are more than 500 protein kinases in the human kinome [3]. The ATP-binding crevice is wedged in-between two compact kinase lobes connected by a more flexible hinge region. Phosphorylation of the so-called activation loop controls kinase activity by affecting substrate and ATP binding. The three-dimensional atomic structure of the first kinase was determined three decades ago [4]; since then, we have learned how kinases themselves are regulated by reversible phosphorylation. It turned out that the structure of all protein kinases is very similar and nucleotide binding and phosphotransfer to substrates are mediated by the well-conserved, common features of the kinase domain core (~250 aa). However, the way how the activity of the different protein kinases is controlled, namely the transition from an inactive state into an active state, is different and there have been different additions to the common kinase core during evolution bringing about new activation mechanisms. In brief, the structures of active kinases are all similar but inactive kinase structures greatly differ [5].

Phosphatases counteract the action of kinases because they are capable of removing the phosphoryl group from phosphorylated proteins [6]. This makes protein phosphorylation a reversible process and this PTM is probably the most widely used mechanism for transmitting signals within the cell. It is estimated that one-third of proteins are subject to regulatory phosphorylation. Because phosphate esters of serine (Ser), threonine (Thr), and tyrosine (Tyr) do not hydrolyze spontaneously under physiological conditions, each protein kinase-mediated phosphorylation needs to be reversed by a protein phosphatase. However, the number of known phosphatase genes is far less than the number of known protein kinases in the human genome (~520 kinases vs. ~180 phosphatases).

Although it is clear that the dual action of kinases and phosphatases is key to the success of alkyl/aryl hydroxy amino acid-based protein phosphorylation as a regulatory mechanism, the molecular logic underlying this type of PTM-based regulation is still not well understood. For example, we have only limited knowledge of how specificity of protein phosphorylation/dephosphorylation is governed and how phosphorylation of specific sites affects function. In summary, we have learned a great deal about how protein kinases as conformational switches work, but how they specifically regulate the phosphorylation of hundreds or thousands of proteins remains far less understood.

Large-scale interactomics and phospho-proteomics are generating data on how protein phosphorylation affects biological function on a global scale. It has become apparent that most protein phosphorylation events occur in disordered protein regions [7]. This top–down strategy complements the classical work addressing individual protein kinases, phosphorylation sites, and their effects on protein function. The output emerging from all these could be best integrated through the exploration of phospho-switches: protein regions whose biophysical/biochemical nature changes upon phosphorylation leading, in turn, to changes in protein activity, interaction capacity, protein abundance, or cellular localization.

17.1.1 Structural and Functional Effects of Protein Phosphorylation

Phospho-amino acids generated by protein phosphorylation act as new chemical entities that do not resemble any natural amino acid. A protein-linked phosphate group can form hydrogen bonds or salt bridges either intra- or intermolecularly and makes stronger hydrogen bonds with arginine than either aspartate or glutamate [1].

Phosphate in a protein was first identified in 1906. Enzymatic phosphorylation of a protein was described 20 years later, and it was 50 years ago when ATP-mediated phosphorylation of a specific serine site was reported. This was in glycogen phosphorylase where phosphorylation allosterically activated the enzyme [8]. The first 3D view of phosphorylation-mediated molecular regulation was then obtained on this enzyme in 1989 [9], which was then followed by the crystal structure of the first protein kinase, PKA, two years later [4]. Protein tyrosine phosphorylation was then discovered in 1979 [10]. This was decades later compared to discovery of Ser/Thr phosphorylation; and after showing that tyrosine phosphorylation may cause similar allosteric changes to what formally had been described in the activation of glycogen phosphorylase, phospho-tyrosine binding domains were also discovered (e.g. SH2 domains) [11]. Their fundamental role in phosphorylation-based biological regulation, parallel to the later discovered phospho-serine/threonine binding domains (e.g. 14-3-3 domains), attracted significant attention, and then they were found in hundreds of signaling proteins [12].

Although protein phosphorylation may happen in structured domains and promote classical allostery, kinases mostly target residues located in flexible regions, for example in an exposed loop or more frequently in a disordered protein region [7]. This is because the latter harbors binding sites for phospho-amino acid recognition domains, may be subject to multisite phosphorylation (see later), and is best suited to be the target of different types of PTMs. These PTM associations (e.g. phosphorylation, acetylation, and mono- and polyubiquitination) may then work together to influence protein function in a hierarchical or combinatorial fashion [13].

Phosphorylation of proteins binding to polyanions, such as DNA, could have a direct impact due to electrostatic repulsion, which in turn may modulate DNA binding. In agreement with this, as early as about three decades ago, the DNA-binding activity of the Oct1 transcription factor was found to be inhibited by

phosphorylation at two distinct phosphorylation sites on its DNA-binding domain. Based on the crystal structure of Oct1 bound to a DNA enhancer, the two serine residues (Ser335 or Ser385) are both located next to the negatively charged DNA phosphate backbone (Figure 17.1a). This suggests a direct mechanistic explanation for decreased DNA binding upon phosphorylation by kinases that, in turn, causes the down-regulation of the mitosis-specific Histone 2B gene [15, 16]. However, such electrostatic repulsion-based direct mechanisms do not seem to be widely used neither for gene expression regulation nor for regulation of RNA–protein binding [17]; these interactions are rather regulated by indirect mechanisms. The C2H2 Zn-finger containing transcription factors constitute the largest family of sequence-specific DNA-binding proteins, and they contain a short-conserved linker (TGEKP) connecting the Zn-fingers. DNA binding requires α-helix formation and the threonine in the linker caps the helices and makes their structure less flexible [18]. It is known that entry into mitosis is accompanied by the cessation of transcription, and apart from RNA polymerase or nucleosome remodeling complex phosphorylation, gene-specific transcription factors may also be inactivated: phosphorylation of the threonine in the conserved linker of C2H2 Zn-finger transcription factors inhibits binding of this large group of DNA-binding proteins [19].

Ribonucleoprotein complexes involved in pre-mRNA splicing and mRNA decay are often regulated by phosphorylation of RNA-binding proteins, which usually occur in dynamic or disordered regions [17]. Many RNA-binding proteins have a signal response segment (SRS) that is disordered but becomes ordered upon phosphorylation and serves as a recognition site for another protein in the ribonucleoprotein complex. SR (serine/arginine) proteins are essential components of the spliceosome that regulate splicing and mRNA export and their processive hyper-phosphorylation plays a key role in mRNA splicing [20]. Furthermore, more recently, phosphorylation was recognized as a major PTM affecting ribonucleoprotein (RNP) granule formation. RNP granules form through liquid–liquid phase separation and phosphorylation can directly weaken or enhance the multivalent interactions between phase-separating macromolecules [21].

The cell membrane is composed of anionic phospholipids that bind polybasic peptides, moreover, this interaction is mediated by bulk electrostatics**.** Apart from structured globular domains responsible for membrane recruitment (e.g. PH, C1 or C2 domains), peripheral membrane proteins often have a positively charged region required for membrane association. Phosphorylation of this region at specific sites could serve as an "electrostatic switch": it attenuates membrane binding by dampening the overall electrostatic attraction between the protein and the membrane. For example, phosphorylation of the yeast Ste5 signaling protein at its N-terminal stretch rich in lysines/arginines by a cyclin-dependent kinase (CDK) demonstrates that phosphorylation can produce changes in protein function through bulk electrostatics, without the necessity of intricate conformational changes [22, 23] (Figure 17.1b)**.** However, it is likely that in most cases a functionally relevant protein phosphorylation event is associated with some specific structural change which will be discussed below.

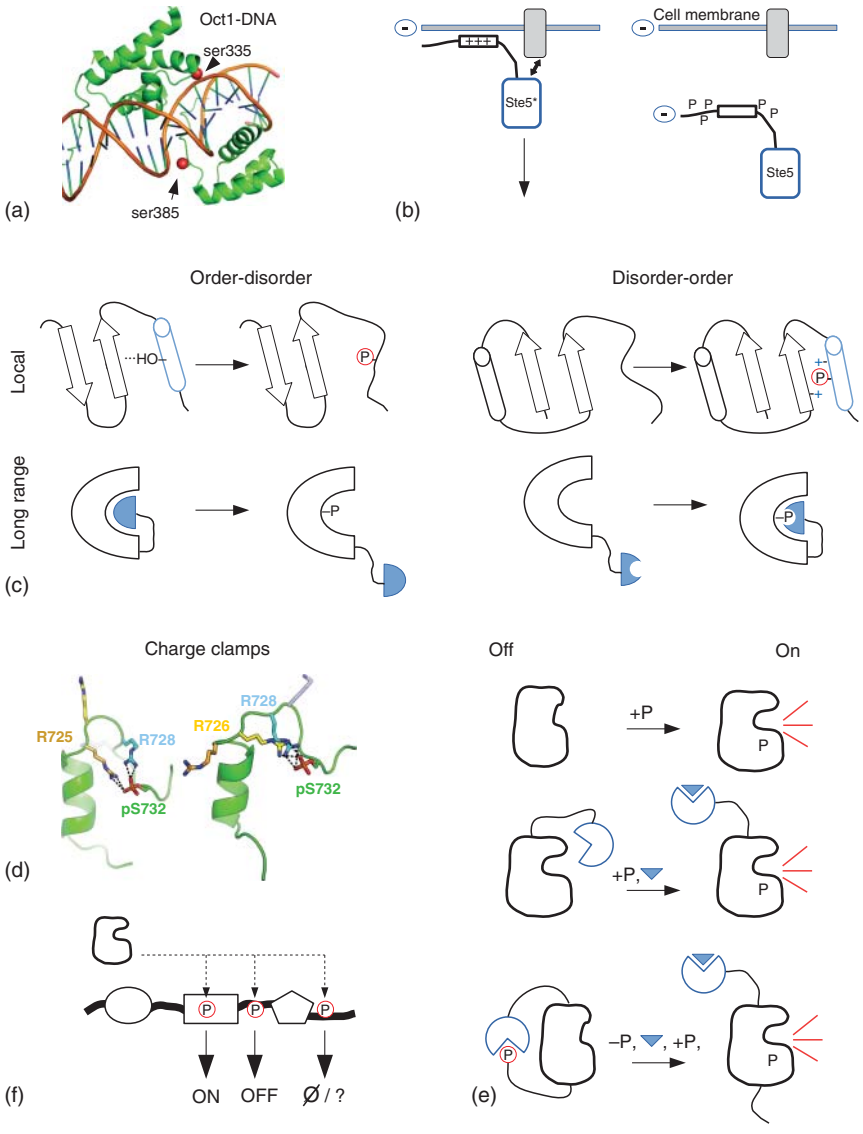


Figure 17.1 Structural effects of phosphate in proteins. (a) Steric clash due to Ser335 or Ser385 phosphorylation in the DNA-binding domain of Oct1 interferes with Oct1-DNA binding in the major or minor groove, respectively (PDB ID: 1GTO), (b) bulk electrostatics (e.g. recruitment of the Ste5 scaffold protein to the cell membrane is blocked when Ste5 is phosphorylated at multiple sites around its membrane binding – originally positively charged – region), (c) local and long-range order-to-disorder or disorder-to-order changes, (d) transient charge clamps (e.g. the arginine/lysine residues in the C-terminal disordered tail of RSK bind to Ser732 when it is phosphorylated; functionally these transient intramolecular charge clamps limit the binding of RSK to one if its partners ERK2). Source: Adapted from Gógl et al. [14], (e) classical vs. modular – conformational or steric – allostery, and (f) no change (e.g. the third phosphorylation event in the hypothetical multi-domain protein may be structurally/functionally neutral).

Phosphorylation may cause local changes in protein structure. In addition to the electrostatic effects, the positively charged amine group in lysine and the guanidium group in arginine side-chains, in particular, can make strong hydrogen bonds. Moreover, the phosphate group can form hydrogen bonds with backbone amides or with different polar amino acid side chains; conversely, it may interfere with hydrophobic interactions. The phosphate group will have a local effect on protein structure by promoting order-to-disorder or disorder-to-order transition in globular proteins and may exert long-range effects through allostery – classical or modular, where the latter involves a phospho-amino acid-binding domain [12] (Figure 17.1c).

Despite that a disordered protein region does not have a structurally well-defined conformation, phosphorylation may affect its conformational dynamics and promote α-helix, turn, or β-sheet formation depending on the neighboring sequence context. For example, lysines or arginines at the right spacing allow the formation of transient charge clamps with the phosphorylated residue, making the region locally less flexible and thus leading to more intra- or intermolecular hydrogen bonds. These latter may promote or limit protein–protein binding; for example, in the disordered C-terminal tail of RSK1, these trigger the disassembly of the ERK2–RSK1 signaling complex [14] (Figure 17.1d). In addition, phosphorylation may play an important role in regulating protein function through modular allostery (Figure 17.1e). The functional outcome of these structural changes naturally will differ from protein to protein; however, it may well be that lots of phosphorylation events remain without any functional consequence: if these do not change the local or global structure, for example, or when there is no system in place that would be able to "read" it out [24] (Figure 17.1f).

17.2 Reversible Protein Phosphorylation in Cellular Signaling: Writers, Readers, and Erasers

Apart from the direct steric or ionic impact of covalently attached phosphate on protein structure, the unique size of the ionic shell and charge properties of this group allows specific and inducible recognition of phosphoproteins by phospho-specific binding domains in other proteins. This promotes inducible protein–protein interaction, therefore phosphorylation may serve as a binary ON/OFF switch in signal transduction networks to transmit signals in response to extracellular stimuli. The appearance of dedicated protein domains capable of binding to phosphorylated amino acids provides a great opportunity for the expansion of classical phosphorylation/dephosphorylation-based mechanisms. According to a simple analogy from an information processing view, protein kinases may be regarded as "writers," phosphatases as "erasers," and modular phospho-amino acid-binding domains as "readers" of phosphorylation-coded intracellular information (Figure 17.2a,b). The latter includes, for example, WW, 14-3-3, and FHA domains [25].

Ser/Thr phosphorylation is more ancient than tyrosine phosphorylation, and the mechanisms of how the writers and readers of these two different systems

work are also somewhat different. For example, tyrosine kinases and phosphatases tend to have more modular domains or specificity elements in addition to their catalytic domains. In contrast to this, Ser/Thr kinases recognize their substrates mostly via their kinase domains, while Ser/Thr phosphatase catalytic domains are promiscuous and their specificity depends on subunits that bind to their specific substrates. Interestingly, an ultradeep human phosphoproteome analysis also revealed the distinct regulatory nature of Tyr and Ser/Thr-based signaling [26].

Tyrosine phosphorylation as a later evolutionary invention became widespread only after the emergence of multicellular organisms [27]. Conversely, tyrosine phosphorylation likely facilitated the evolution of multicellular organisms [28]. The rapid expansion of phospho-tyrosine binding domains – the readers, such as SH2, PTB domains – brought about new cytoplasmic adapters for receptor tyrosine kinases/phosphatases and new non-receptor tyrosine kinases/phosphatases controlled by modular allostery [12] (Figure 17.2c–e).

In summary, a phosphate group in proteins may directly affect protein activity and binding by eliciting a local structural/electrostatic change due to its specific chemical nature – since it is a dianion at neutral pH with good capacity to form salt bridges and hydrogen bonds, but more frequently the functional effect is indirectly "read out" and is mitigated by phospho-amino acid binding domains.

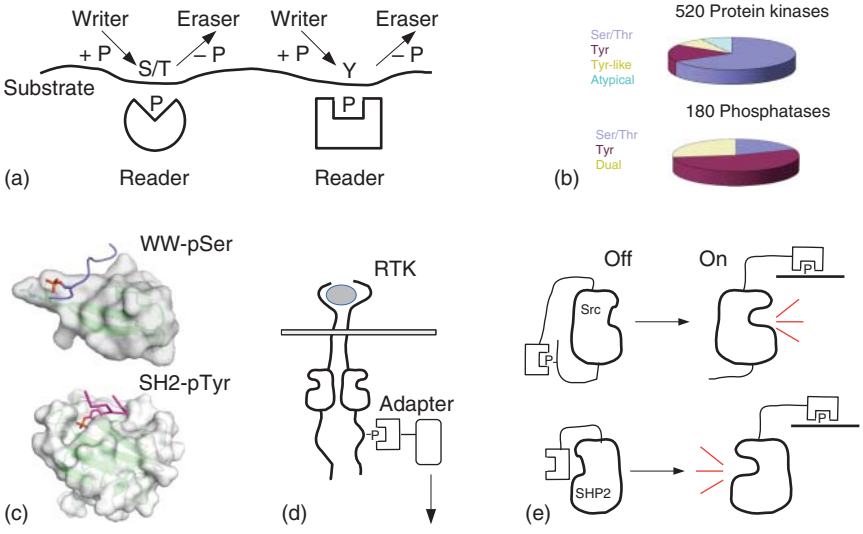


Figure 17.2 Protein kinases, phosphatases, and phospho-amino acid binding domains. (a) Scheme on writers/erasers/readers, (b) Statistics of serine/threonine vs tyrosine kinases and phosphatases from the human proteome, (c) concrete examples for a pSer/Thr or a pTyr binding domain (WW, PDB ID: 2N1O or SH2, PDB ID: 1TZE), (d) SH2 domain-based adapters for receptor tyrosine kinases (e.g. Grb2; RTK: receptor tyrosine kinase, for example, EGF receptor), and (e) examples of SH2 domain-based modular allostery (Src kinase – conformational, SHP2 phosphatase – steric; phosphorylated ligand binding *in trans* relieves the intramolecular autoinhibition; the multi-domain organization of these two enzymes are shown simplified).

17.3 Protein Kinases as Molecular Switches and as Components of Signaling Cascades

Protein kinases could be divided into two primary kinase classes: the eukaryotic protein kinases (ePK) and the atypical protein kinases (aPK). The eukaryotic protein kinases are related via their highly conserved kinase core where twelve of the specific conserved sequence motifs could be classified into twelve subdomains. ePK could be further divided into eight groups: AGC, CAMK, CK1, CMGC, STE, RGC, tyrosine kinase, and tyrosine kinase-like [3]. These groups are distinguished by their specific regulatory mechanisms. Compared to ePKs, atypical protein kinases form a much smaller group and lack the conserved sequence features, albeit they share the same kinase domain fold. This chapter discusses the general features of the ePK group.

There are some protein kinases that are constitutively active, but most protein kinases may be regarded as molecular switches whose activity is universally controlled by phosphorylation at their activation segment [29]. Protein kinases may get phosphorylated at other regions in their catalytic core or in their disordered regions flanking the latter [30]. These auxiliary sites vary among kinases but they have an important but normally indirect role: similarly to other protein phosphorylation sites, they could influence the binding of the kinase to other proteins or the cellular localization, for example. Important structural changes required for turning an inactive kinase into an active one, apart from activation loop phosphorylation, include adjustments between the two kinase lobes to allow nucleotide binding, aligning the ATP phosphates, Mg ions, and the catalytic aspartate to hydrolyze the nucleotide's phospho-anhydride bond, opening of the substrate-binding pocket, and positioning of the phosphorylatable residue for phosphotransfer. There are evolutionary conserved sequence signatures related to any of these processes, and the lack of any of these features in a kinase sequence indicates that the kinase is not active and likely functions as a pseudokinase [30]. Conversely, there are some constitutively active kinases (e.g. PDK1, CK2, and GSK3), where all the catalytic requirements are in place after translation and folding and thus these enzymes are not dependent on activation segment phosphorylation.

Protein kinases may operate in a more complex fashion than simple ON/OFF switches. First, activation segment phosphorylation at one site may not suffice to turn the kinase on, and the phosphorylation of a few more (normally one or two) residues is required in a distributive process (e.g. the two phosphorylation sites in the MAP kinase activation loop). Second, some kinases work as conditional switches and require priming phosphorylation at an allosteric site in addition to activation segment phosphorylation (e.g. AGC kinases with a dual requirement for hydrophobic motif and activation loop phosphorylation). This latter process allows the transmission of the phosphorylation-based signal toward downstream substrates only if two upstream kinases were simultaneously active. From a structural point of view, the activation segment in kinases provides a good example of how phosphates trigger functionally important order-to-disorder or disorder-to-order structural changes: phosphorylation of amino acids in key positions may cause a steric clash with the globular kinase domain core and makes the activation loop

more flexible opening access to the otherwise occluded active site. Conversely, the activation loop may become locked in a more rigid conformation by binding the phosphates to nearby arginines that in turn are required for the formation of the substrate-binding pocket. Activity regulation for some kinases may be very complex: it involves all these mentioned mechanisms in addition to kinase-specific ones; for example, modular allostery due to their extra modular domains (SH2 domain – Src tyrosine kinase, PH domain – PKB) or to subunits binding to different secondary messengers (e.g. cAMP – PKA, DAG – PKC, and Ca/calmodulin – CAMKII) [5]. In some even more complex cases, there could be two kinase domains within one kinase where one activates the other (e.g. RSK) [30] .

Some kinases require that their substrate was pre-phosphorylated by another kinase earlier (e.g. GSK3 binds to and phosphorylates S/T-P-x-x-Phos sites, CK1 binds to and phosphorylates Phos-x-x-S sites, where "Phos" indicates the pre-phosphorylated, "primed," phospho-amino acid and x indicates any amino acid). For example, the JNK mitogen-activated protein (MAP) kinase – as the "master" kinase – responsible for phosphate priming may work together with GSK3 or CK1 to ensure multisite phosphorylation of some of its substrates [31]. Since GSK3 and CK1 – as the "slave" kinase – are both constitutively active, their partnership with the signal-activated JNK MAP kinase may be required for counteracting the action of phosphatases and/or to set up a dynamic threshold for JNK pathway activity-based phosphorylation of specific substrates, probably helping to lower noise. Protein dephosphorylation is simply controlled by recruitment, and one Ser/Thr catalytic domain (e.g. PP1) may involve hundreds of different targeting subunits that place the enzyme next to its substrates [32]. In general, phosphatase catalytic domains are somewhat simpler than most protein kinase domains regarding structural plasticity; however, some tyrosine phosphatase domains are known to be modulated by modular allostery (e.g. SHP2) [33].

Protein kinases often form cascades where they phosphorylate each other. This may be important for signal amplification, to allow protein phosphatases to act at several levels before the phosphorylation of the final effector protein takes place, or to bring about a more complex system that allows the integration of different signaling pathways. In addition, a hierarchically organized kinase cascade provides a simple blueprint for non-linear network behavior, for example for ultrasensitivity, where the protein network shows a non-linear cooperative response, albeit each phosphorylation event between network components is linear [34]. In addition, protein kinases and phosphatases may also establish phosphorylation-based positive and negative feedback loops, where a given enzyme affects the activity of another enzyme found several tiers away from its direct interactors [35]. Because of this complexity, kinase–phosphatase cascades create a challenge for how to handle signaling fidelity even before the phosphorylation of the effector protein happens by the last kinase. To maintain signaling fidelity, namely that an extracellular cue would elicit the right protein phosphorylation pattern, but also to allow some level of required cross-talk for signal integration, kinases, phosphatases, and their substrates form specific interactions that are mostly mediated by so-called linear binding motifs [36]. These interactions depend on protein associations between a

structured globular domain and a short stretch of amino acids located in highly flexible or disordered protein regions. They mediate transient protein–protein binding with only micromolar binding affinity, well suited for fast signal propagation through complex protein kinase/phosphatase networks working with a high turnover rate, and complement those globular domain–domain type interactions that take place when kinases activate each other by phosphorylation or when a phosphatase deactivates a kinase by dephosphorylation [37].

17.4 Mechanisms of Phosphorylation Specificity: the Importance of Short Linear Motifs

A key question in kinase biology is how kinases achieve specificity for their substrates, and conversely, how phosphatases specifically dephosphorylate these. This section will focus on the specificity of protein kinases. An important aspect of substrate recognition is that the phosphorylation site on the substrate falls within a consensus amino acid sequence that is complementary to the active site of the kinase. However, the interaction between the substrate-binding pocket and the phosphorylation target motif normally does not ensure signaling fidelity, because consensus sites are short and occur in hundreds or thousands in the proteome (e.g. an S/TP site for MAP kinases and CDKs or an RxxS/T motif for most AGC kinases) [38, 39].

Kinase specificity can be increased by forming supplementary interactions that recruit the kinase to the substrate. These interactions can be mediated by linear binding motifs – often referred to as docking motifs found in the substrate that binds to the so-called docking groove on the kinase – or by auxiliary modular domains found either on the kinase or on the substrate. The important role of these short linear motif (SLiM)-based protein–peptide type interactions is well-established for many kinases and phosphatases (Figure 17.3). Cyclin-dependent kinases provide an interesting example where the cyclin subunit is not only required for the activity of the kinase but also has substrate-binding docking grooves for different types of

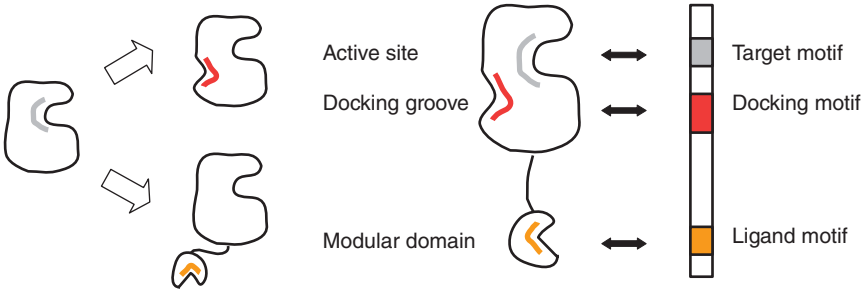


Figure 17.3 Mechanisms of phosphorylation specificity: the importance of protein–peptide type interactions. Phosphorylation specificity is governed by protein–protein interaction specificity between the active site of the kinase and its target motif, kinase-docking grooves and docking motifs, and/or modular binding domains and their ligands.

docking motifs that increase the specificity of S/TP phosphorylation sites [40, 41]. In addition, the phosphorylation of specific substrates may also be increased by anchoring or scaffold proteins [42]. Anchoring proteins are located in specific compartments of the cell and contribute to the spatial pattern of kinase-substrate phosphorylation. The important role of A-kinase anchoring proteins (AKAPs) in directing PKA activity in the cell is well-established [43]. Scaffold proteins bind more than one kinase and facilitate signaling through kinase cascades and/or allow combinatorial regulation between the different tiers of the cascade [12]. Similarly, a protein scaffold that binds a kinase and one of its effector protein substrates at the same time promotes phosphorylation by passive tethering. Therefore, kinase-docking motifs are not only found in direct substrates but also on signaling scaffold proteins where they may indirectly increase the local concentration of the substrate's target motif around the kinase's active site [44].

Mass spectrometry-based phosphoproteomics identified thousands of phosphosites [45]. This huge amount of data is collected and curated in dedicated phosphorylation site resources such as phosphoELM (http://phospho.elm.eu.org) or PhosphoSitePlus (www.phosphosite.org) and in general protein databases such as HPRD (www.hprd.org) and Uniprot (www.uniprot.org) [46]. Kinase–phosphosite predictions are now also possible using sequence-based computational tools. However, our information on this topic is still fragmented, and possibly a large fraction of physiologically relevant phosphorylation sites may not even be known. Fortunately, kinase group-specific or kinome-wide, systems-level experimental studies are adding to our understanding of the molecular basis of protein kinase-mediated phosphorylation specificity [47, 48].

Because most phosphorylation sites are located in disordered protein regions, it is of utmost importance to examine how these regions bind kinases/phosphatases. In contrast to protein interfaces formed between globular domains, where interacting residues come from different parts of the 3D structure, the contact residues of linear motifs are contained within a relatively short (3 to 25 aa) stretch. In principle, this should make the identification of linear motifs binding to the same protein domain straightforward, based on their sequence similarity. Unfortunately, the low information content of these motifs renders them elusive for simple consensus motif-based *in silico* searches [49]. This is due to the fact that only a few fixed positions (as few as 2–3 aa residues, albeit more typically 4–5) define a particular type of linear motifs, called classes. The Eukaryotic Linear Motif (ELM) database currently contains more than 200 motif classes with more than 2000 instances [50].

Linear binding motif discovery will likely facilitate our understanding of phosphorylation specificity because auxiliary protein interactions – beyond substrate target motif binding – play a decisive role. However, natural SLiMs exist as multiple, relaxed versions of a hypothetical optimal consensus. Because of these shortcomings, purely sequence-based *in silico* approaches should be complemented by structure-based scoring schemes whenever possible [51, 52]. Testing three-dimensional complementarity of peptides to the partner domain surface will then help to filter hits before starting their experimental validation.

17.5 Examples of Phospho-Switch-Based Biological Regulation

SLiMs, apart from passively mediating binding between proteins, can play a more active role since they can also be subject to different types of PTMs, which may change their capacity for being able to function as recruitment sites between proteins. The switches ELM resource (http://switches.elm.eu.org) is a useful compendium of short linear motifs that are known to be regulated by phosphorylation [52]. It is important to note that phospho-switches rarely work as classical ON/OFF switches and they rather operate as dimmers where phosphorylation changes their binding affinity to their partners modestly, mostly less than 10-fold [53] (Figure 17.4).

Another important aspect of phospho-switches is that they are often subject to multisite phosphorylation, namely that they are phosphorylated at multiple sites. These more complex phospho-switches may be modified by one kinase at several sites or by different kinases at distinct sites. Thus, their earlier described biophysical/structural sensitivity toward accommodating phospho-amino acids forms the mechanistic basis for higher-level phosphorylation-based biological regulation (e.g. noise suppression, ultrasensitivity, and signal integration). The following sections will describe how CDK activity thresholds may drive cell cycle stage-specific substrate phosphorylation and how human mitogen-activated protein kinases (MAPK) exert their control on the activity of specific transcription

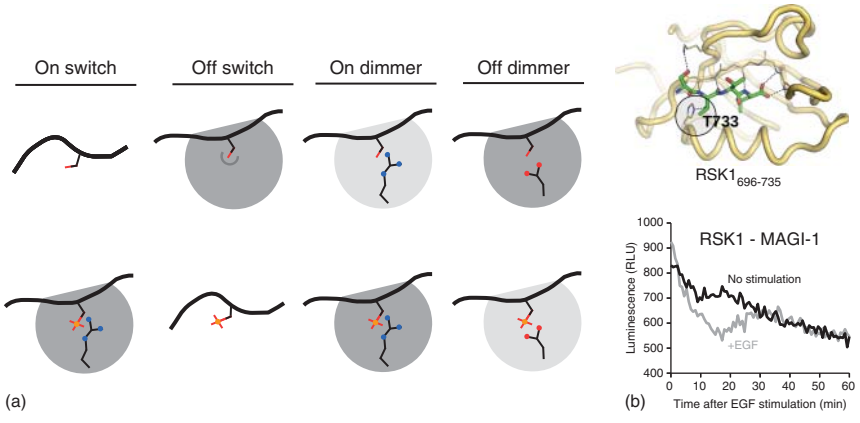


Figure 17.4 Phosphorylation-based ON/OFF switches vs. dimmers. (a) Phosphorylation might have a profound effect on protein–peptide type interactions, either by promoting it (ON switch), disrupting it (OFF switch), or by altering the binding affinity in a more gradual fashion (ON or OFF dimmer). (b) As an example of dynamic regulation, the formation of RSK1–MAGI1 complex is controlled by PDZ domain binding and PDZ ligand phosphorylation (on Thr733). The crystal structure of the RSK1 C-terminus (green) in complex with the MAGI1 PDZ domain (brown) is shown in the upper panel. The lower panel shows the results of a luciferase complementation assay, where the complex formation between RSK1 and MAGI1 was monitored in unstimulated (black) and EGF-stimulated (gray) HEK293 cells. Source: Gógl et al. [14]/John Wiley & Sons.

factors. These examples will demonstrate how the complex regulation of biological activities may come about by combining kinase binding linear motifs and their phosphorylation target motifs in disordered protein regions, and also how the function of the latter could simply change as phosphorylation-based regulatory systems evolve.

Cyclin-dependent kinases (CDKs) drive the major events of the eukaryotic cell division cycle. Substrates of Cdk1 in budding yeast (*Saccharomyces cerevisiae*) were analyzed by phospho-proteomics *in vivo*, which identified 547 phosphorylation sites on 308 proteins. It was found that more than 90% of the identified Cdk1-dependent phosphorylation sites were located in loops and disordered regions [47]. Furthermore, it was also found that Cdk1 targets had multiple phosphorylation sites that tended to cluster, suggesting that multiple phosphorylations modulate the same protein surface. Further analysis of Cdk1 substrate orthologs from 32 different fungal species revealed that many substrates are phosphorylated at rapidly evolving site clusters, which are likely to modify substrate function by simply disrupting or generating protein–protein interactions, possibly allowing diverse cell cycle control mechanisms to adapt rapidly. CDKs are activated by different cyclins at different cell cycle stages, and a quantitative model of CDK function states that cyclins temporally order cell cycle events at different CDK activity levels or thresholds. Three mechanisms determine the phosphorylation rate of a yeast CDK substrate: active site specificity, presence of Cks1 binding sites (phospho-TP), and cyclin docking motifs. CDK1 functions in complex with a specific cyclin (also an allosteric activator of the kinase) and Cks1 (a phospho-threonine binding adapter). The CDK active site recognizes minimal (S/TP) or full consensus motifs (S/TPxK/R), cyclins can bind to specific substrates via linear motifs for substrate targeting and the Cks1 subunit of the CDK complex interacts with phosphorylated TP sites and directs multisite phosphorylation. It was suggested that the cyclin-CDK-Cks1 complex could serve as a scaffold for disordered Cdk1 substrates, mediating an ordered phosphorylation process. According to the multisite phosphorylation code hypothesis, the linear motifs (phosphorylation and docking sites) and their patterns function as a barcode giving a unique identity to each substrate. The cyclin-CDK-Cks1 complex can read the barcode and assigns the execution of any CDK-triggered switch to a specified time point during the cell cycle [54].

We have a substantial understanding of how individual MAPKs bind to transcription factors (TF) since most of the known docking motifs (the so-called D- or F-motifs) were originally found in these proteins decades ago. Physiological responses to extracellular cues depend on a complex activation pattern of different MAPKs (e.g. ERK1/2, JNK, and p38), which is not exclusive but combinatorial, and not binary (i.e. on or off) but quantitative. Immediate-early (IE) gene transcription factors (TF; e.g. Elk1, ATF2) are directly linked to MAPK-signaling cascades, and thus the rate of transcription of a specific gene is coupled to the strength of intracellular-signaling events [55]. The primary response to an extracellular cue is triggered by phosphorylating the TF in its transactivation domain (TAD), which is structurally mostly disordered. The Elk1 TAD binds to one of the subunits of the Mediator complex promoting gene expression, and this binding region is

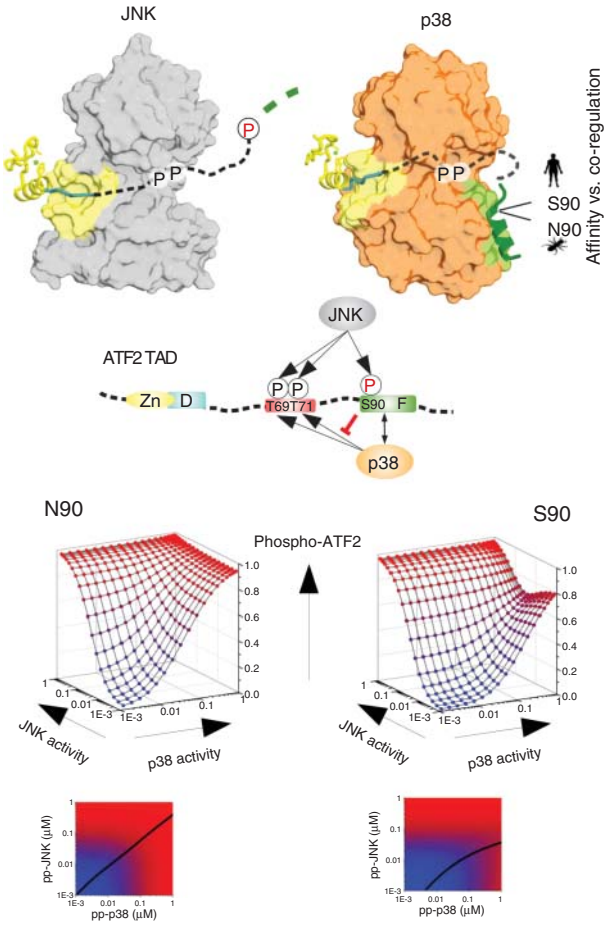


Figure 17.5 An example of a complex, linear motif-based phospho-switch: ATF2 TAD. The ATF2 transcription factor regulates gene expression in response to MAP kinase activation. ATF2 contains a binding site (a Zn finger + D-motif binding to the D-groove, yellow) for JNK and an F-site (green) binding to the F-groove. TAD activity is correlated with the phosphorylation of T69/71 sites. The invertebrate ATF2 TAD contains an asparagine at position 90 which optimally caps the p38 binding helical region: the two kinases activate the TAD in an additive manner. However, all vertebrate ATF2 orthologs have a serine at this position (S90) that is phosphorylated by JNK. Because Ser90 phosphorylation blocks p38 binding, the vertebrate TAD architecture allows a more complex regulation of how p38 and JNK co-regulate ATF2 transcriptional output. Upper panels show the structural models of p38- and the JNK-ATF2 TAD complexes and bottom panels show the results of phospho-ATF2 levels calculated with an in silico model. Source: Kirsch et al. [57]/Springer Nature/CC BY 4.0.

surrounded by ERK phosphorylation target sites and also contains an ERK-binding D-motif as well as an F-motif. The Elk1 TAD is regulated by ERK phosphorylation: the phosphorylation rate of eight distinct sites is determined by their position relative to ERK2 docking motifs, and residues with fast/intermediate or slow phosphorylation rates promote or attenuate co-activator binding, respectively. This

allows the TAD to promote transcription under low/intermediate ERK-signaling pathway flux and to down-regulate it under high ERK flux. In brief, the Elk-1 TAD harbors a MAPK-controlled complex phospho-switch subject to multi-site phosphorylation and demonstrates how progressive Elk-1 TAD phosphorylation causes a self-limiting response to ERK activation [56].

Another study found that the ATF2 TAD architecture is key in determining how the transcription factor will respond to JNK, p38, or the concomitant activation of these two. The ATF2 TAD has two MAPK phosphorylation sites (T69/71), and p38 and/or JNK-mediated phosphorylation promotes the transcription of specific genes. The JNK-binding D-motif and a p38-specific F-motif were mapped N-terminal or C-terminal from the critical T69/71 sites. The topology of the MAPK binding sites relative to the transcription-controlling phosphorylation sites is evolutionarily conserved but in vertebrates the TAD also contains a unique JNK phosphorylation site located in the p38-binding F-site, and JNK phosphorylation at this unique site (Ser90) attenuates p38 binding. Because of this, the ATF2 TAD in vertebrates does not simply integrate the signal coming from the two MAPKs, as this seems to be the case in invertebrate orthologs, but due to its unique architecture, it responds to the relative strength of these two MAPK pathways. This latter capacity of the ATF2 TAD goes beyond additive phosphorylation of the same phospho-switch. The basic integration capacity, emerging from an ancient pre-vertebrate ATF2 architecture comprised of JNK and p38 binding sites, became more complex due to an amino acid change (Asn to Ser at 90) in the critical p38 binding region in vertebrates: one of the MAPKs (JNK) acquired the capacity to have a direct influence on how the other (p38) effects ATF2 mediated transcription [57] (Figure 17.5).

Changing the response of transcription factors to signaling pathways is an important mechanism in the evolution of gene regulation. It was posited that amino-acid changes in the CCAAT/enhancer-binding protein-β (CEBPB) changed the way this TF responded to cAMP/protein kinase A/GSK3β signaling in placental mammals, which in turn changed expression of prolactin hormone [58]. The novel function came about because of amino acid substitutions reorganizing the location of key phosphorylation sites. These simple changes in the architecture of the transcription controlling regulatory domain of CEBPB were sufficient to cause a fundamental change, from repression to activation, in how the TF responded to phosphorylation.

17.6 Conclusion

Phospho-switch-based regulation has been mostly addressed in individual studies where the goal was to reveal the mechanistic basis underlying the function of a known phosphorylation event in a given protein. These studies provided dozens of well-characterized examples, and as the number of biochemically and structurally explored cases increased, some common themes have emerged: (i) most protein phosphorylation occurs in a disordered protein region (although the globular "writers" are also regulated by phosphorylation), (ii) phosphorylation sites with

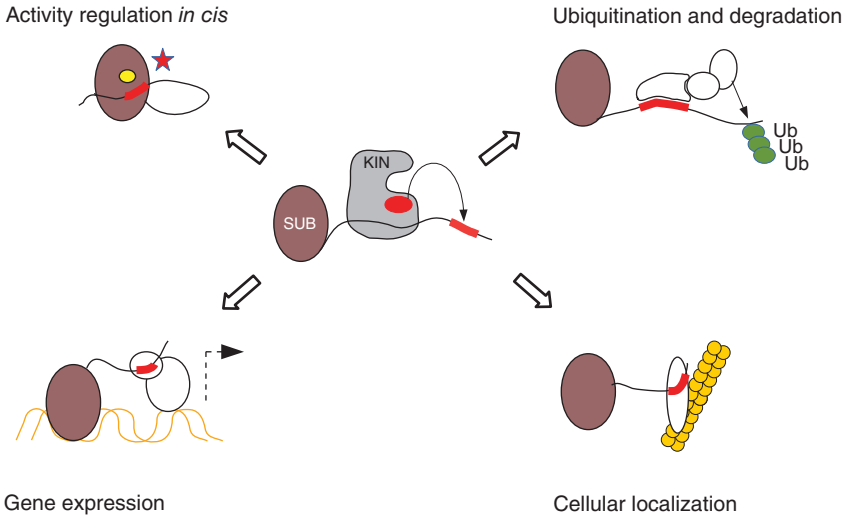


Figure 17.6 Functional modalities regulated by phospho-switches (KIN: kinase; SUB: substrate). Phosphorylation may affect enzymatic activity through specialized phospho-switches that work *in cis*. The mechanistic basis of this type of regulation may be highly diverse and depend on the specific topology of the phospho-regulated protein. However, phospho-switches that work *in trans* may operate similarly: they are phosphorylation sensitive interaction hot spots for effector protein binding involved in protein level control (e.g. ubiquitin ligases or RNA polymerase subunits) or cellular localization (e.g. anchor proteins binding to the cytoskeleton).

functional relevance are often located in known linear binding motifs (for example in the ligands of the "readers", in the SLiMs that bind to phospho-tyrosine or phospho-serine/threonine binding domains), and (iii) phospho-switches are often associated with known linear binding motifs for kinases or phosphatases enhancing the rate of catalytic activity and specificity among the "writers/erasers". Because of these, prediction of phospho-switches in proteins using bioinformatics is inherently linked with linear binding motif and kinase phosphorylation site discovery in general. Fortunately, based on our current structural and biochemical knowledge, individual kinase phosphorylation sites in a protein may be predicted. Moreover, large-scale phospho-proteomics studies provide lots of experimental data on protein phosphorylation globally and under relevant cellular settings. Concurrently, large-scale experimental interactomics, particularly those focusing on the disordered part of the proteome [59], provide more and more data on linear motif mediated protein–protein associations. Moreover, the three-dimensional structure of the writers/erasers and reader domains may be used for the structural modeling of protein–peptide type interactions, and more recently deep-learning computational approaches are also employed to predict these types of interactions [60]. The real challenge in exploring phospho-switches on the large scale is their experimental analysis/validation. There are several protein–protein interaction tools used in protein interactomics: peptide–protein arrays, biological surface display

techniques (e.g. phage display), affinity purification mass spectrometry (AP–MS), protein fragment complementation assays (PCA), two-hybrid screening (Y2H or M2H), signaling pathway reconstruction (e.g. MAPPIT), proximity-based protein labeling, or *in vivo* protein crosslinking that may be used to explore phospho-switch regions as baits; however, the modification of these regions by a kinase in these artificial systems may not be straightforward, moreover, the preys are also mostly unknown [61]. Fortunately, earlier studies on phosphorylation-based biological regulation established that this PTM frequently affects protein abundance (for example through phospho-degrons that bind to ubiquitin ligases), gene expression (for example through binding to phosphorylation controlled TADs), or cellular localization (for example through binding to phospho-amino acid-binding domains from anchoring proteins) and the study of these functional modalities may be addressed with dedicated cell-based assays (Figure 17.6). We foresee that the discovery of phospho-switches will facilitate the understanding of protein phosphorylation not only in individual proteins but also at a higher level, in protein networks and ultimately in intact cellular systems, too.

Acknowledgments

This work was supported by grants from the National Research, Development and Innovation Office, Hungary (KKP_17 126963).

References

1 Hunter, T. (2012). Why nature chose phosphate to modify proteins. *Philos. Trans. R. Soc. London, Ser. B* 367 (1602): 2513–2516.
2 Zschiedrich, C.P., Keidel, V., and Szurmant, H. (2016). Molecular mechanisms of two-component signal transduction. *J. Mol. Biol.* 428 (19): 3752–3775.
3 Manning, G., Whyte, D.B., Martinez, R. et al. (2002). The protein kinase complement of the human genome. *Science* 298 (5600): 1912–1934.
4 Knighton, D.R., Zheng, J., Ten Eyck, L.F. et al. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* 253 (5018): 407–414.
5 Huse, M. and Kuriyan, J. (2002). The conformational plasticity of protein kinases. *Cell.* 109 (3): 275–282.
6 Damle, N.P. and Köhn, M. (2019). The human DEPhOsphorylation database DEPOD: 2019 update. *Database* 2019.
7 Iakoucheva, L.M., Radivojac, P., Brown, C.J. et al. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32 (3): 1037–1049.
8 Pawson, T. and Scott, J.D. (2005 Jun). Protein phosphorylation in signaling – 50 years and counting. *Trends Biochem. Sci* 30 (6): 286–290.

9 Barford, D. and Johnson, L.N. (1989). The allosteric transition of glycogen phosphorylase. *Nature* 340 (6235): 609–616.

10 Eckhart, W., Hutchinson, M.A., and Hunter, T. (1979). An activity phosphorylating tyrosine in polyoma T antigen immunoprecipitates. *Cell* 18 (4): 925–933.

11 Pawson, T. (2004). Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell* 116 (2): 191–203.

12 Bhattacharyya, R.P., Reményi, A., Yeh, B.J., and Lim, W.A. (2006). Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu. Rev. Biochem.* 75: 655–680.

13 Minguez, P., Letunic, I., Parca, L. et al. (2015). PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic Acids Res.* 43 (Database issue): D494–D502.

14 Gógl, G., Biri-Kovács, B., Póti, A.L. et al. (2017). Dynamic control of RSK complexes by phosphoswitch-based regulation. *FEBS J.* 46–71.

15 Segil, N., Roberts, S.B., and Heintz, N. (1991). Mitotic phosphorylation of the Oct-1 homeodomain and regulation of Oct-1 DNA binding activity. *Science* 254 (5039): 1814–1816.

16 Kang, J., Goodman, B., Zheng, Y., and Tantin, D. (2011). Dynamic regulation of Oct 1 during mitosis by phosphorylation and ubiquitination. *PLoS One* 6 (8).

17 Thapar, R. (2015). Structural basis for regulation of RNA-binding proteins by phosphorylation. *ACS Chem. Biol.* 10 (3): 652–666.

18 Laity, J.H., Dyson, H.J., and Wright, P.E. (2000). DNA-induced alpha-helix capping in conserved linker sequences is a determinant of binding affinity in Cys(2)-His(2) zinc fingers. *J. Mol. Biol.* 295 (4): 719–727.

19 Dovat, S., Ronni, T., Russell, D. et al. (2002). A common mechanism for mitotic inactivation of C2H2 zinc finger DNA-binding domains. *Genes Dev.* 16 (23): 2985–2990.

20 Ngo, J.C.K., Giang, K., Chakrabarti, S. et al. (2008). A sliding docking interaction is essential for sequential and processive phosphorylation of an SR protein by SRPK1. *Mol. Cell.* 29 (5): 563–576.

21 Hofweber, M. and Dormann, D. (2019). Friend or foe-Post-translational modifications as regulators of phase separation and RNP granule dynamics. *J. Biol. Chem.* 294 (18): 7137–7150.

22 Serber, Z. and Ferrell, J.E. (2007). Tuning bulk electrostatics to regulate protein function. *Cell.* 128 (3): 441–444.

23 Strickfaden, S.C., Winters, M.J., Ben-Ari, G. et al. (2007). A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. *Cell* 128 (3): 519–531.

24 Beltrao, P., Albanèse, V., Kenner, L.R. et al. (2012). Systematic functional prioritization of protein posttranslational modifications. *Cell* 150 (2): 413–425.

25 Yaffe, M.B. and Elia, A.E.H. (2001). Phosphoserine/threonine-binding domains. *Curr. Opin. Cell Biol.* 13 (2): 131–138.

26 Sharma, K., D'Souza, R.C.J., Tyanova, S. et al. (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* 8 (5): 1583–1594.

27 Miller, W.T. (2012). Tyrosine kinase signaling and the emergence of multicellularity. *Biochim. Biophys. Acta* 1823 (6): 1053–1057.

28 Lim, W.A. and Pawson, T. (2010). Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* 142 (5): 661–667.

29 Nolen, B., Taylor, S., and Ghosh, G. (2004). Regulation of protein kinases; controlling activity through activation segment conformation. *Mol. Cell.* 15 (5): 661–675.

30 Gógl, G., Kornev, A.P., Reményi, A., and Taylor, S.S. (2019). Disordered protein kinase regions in regulation of kinase domain cores. *Trends Biochem. Sci.* 44 (4): 300–311.

31 Beurel, E., Grieco, S.F., and Jope, R.S. (2015). Glycogen synthase kinase-3 (GSK3): regulation, actions, and diseases. *Pharmacol. Ther.* 148: 114–131.

32 Bollen, M., Peti, W., Ragusa, M.J., and Beullens, M. (2010). The extended PP1 toolkit: designed to create specificity. *Trends Biochem. Sci.* 35 (8): 450–458.

33 Lim, W.A. (2002). The modular logic of signaling proteins: building allosteric switches from simple binding domains. *Curr. Opin. Struct. Biol.* 12 (1): 61–68.

34 Huang, C.Y.F. and Ferrell, J.E. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc. Natl. Acad. Sci. U.S.A.* 93 (19): 10078–10083.

35 Kholodenko, B.N. (2006). Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell Biol.* 7 (3): 165–176.

36 Reményi, A., Good, M.C., and Lim, W.A. (2006). Docking interactions in protein kinase and phosphatase networks. *Curr. Opin. Struct. Biol.* 16 (6): 676–685.

37 Alexa, A., Gógl, G., Glatz, G. et al. (2015). Structural assembly of the signaling competent ERK2-RSK1 heterodimeric protein kinase complex. *Proc. Natl. Acad. Sci. U.S.A.* 112 (9): 2711–2716.

38 Ubersax, J.A. and Ferrell, J.E. (2007). Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* 8 (7): 530–541.

39 Miller, C.J. and Turk, B.E. (2018). Homing in: mechanisms of substrate targeting by protein kinases. *Trends Biochem. Sci.* 43 (5): 380–394.

40 Schulman, B.A., Lindstrom, D.L., and Harlow, E. (1998). Substrate recruitment to cyclin-dependent kinase 2 by a multipurpose docking site on cyclin A. *Proc. Natl. Acad. Sci. U.S.A.* 95 (18): 10453–10458.

41 Loog, M. and Morgan, D.O. (2005). Cyclin specificity in the phosphorylation of cyclin-dependent kinase substrates. *Nature* 434 (7029): 104–108.

42 Zeke, A., Lukács, M., Lim, W.A., and Reményi, A. (2009). Scaffolds: interaction platforms for cellular signalling circuits. *Trends Cell Biol.* 19 (8): 364–374.

43 Wong, W. and Scott, J.D. (2004). AKAP signalling complexes: focal points in space and time. *Nat. Rev. Mol. Cell Biol.* 5 (12): 959–970.

44 Zeke, A., Misheva, M., Reményi, A., and Bogoyevitch, M.A. (2016). JNK signaling: regulation and functions based on complex protein-protein partnerships. *Microbiol. Mol. Biol. Rev.* 80 (3): 793–835.

45 Mann, M., Ong, S.E., Grønborg, M. et al. (2002). Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol.* 20 (6): 261–268.

46 Savage, S.R. and Zhang, B. (2020). Using phosphoproteomics data to understand cellular signaling: a comprehensive guide to bioinformatics resources. *Clin. Proteomics* 17 (1): 1–18.

47 Holt, L.J., Tuch, B.B., Villen, J. et al. (2009). Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 325 (5948): 1682–1686.

48 Ptacek, J., Devgan, G., Michaud, G. et al. (2005). Global analysis of protein phosphorylation in yeast. *Nature* 438 (7068): 679–684.

49 Krystkowiak, I. and Davey, N.E. (2017). SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. *Nucleic Acids Res.* 45 (W1): W464–W469.

50 Gouw, M., Michael, S., Sámano-Sánchez, H. et al. (2018). The eukaryotic linear motif resource −2018 update. *Nucleic Acids Res.* 46 (D1): 428–434.

51 Sánchez, I.E., Beltrao, P., Stricher, F. et al. (2008). Genome-wide prediction of SH2 domain targets using structural information and the Fold X algorithm. *PLoS Comput. Biol.* 4 (4): 1–10.

52 Zeke, A., Bastys, T., Alexa, A. et al. (2015). Systematic discovery of linear binding motifs targeting an ancient protein interaction surface on MAP kinases. *Mol. Syst. Biol.* 11 (11): 837.

53 Gógl, G., Biri-Kovács, B., Durbesson, F. et al. (2019). Rewiring of RSK–PDZ interactome by linear motif phosphorylation. *J. Mol. Biol.* 431 (6): 1234–1249.

54 Örd, M., Möll, K., Agerova, A. et al. (2019). Multisite phosphorylation code of CDK. *Nat. Struct. Mol. Biol.* 26 (7): 649–658.

55 Hazzalin, C.A. and Mahadevan, L.C. (2002). MAPK-regulated transcription: a continuously variable gene switch? *Nat. Rev. Mol. Cell Biol.* 3: 30–40.

56 Mylona, A., Theillet, F.X., Foster, C. et al. (2016). Opposing effects of Elk-1 multisite phosphorylation shape its response to ERK activation. *Science* 354 (6309): 233–237.

57 Kirsch, K., Zeke, A., Tőke, O. et al. (2020). Co-regulation of the transcription controlling ATF2 phosphoswitch by JNK and p 38. *Nat. Commun.* 1: 11(1).

58 Lynch, V.J., May, G., and Wagner, G.P. (2011). Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB. *Nature* 480 (7377): 383–386.

59 Davey, N.E., Seo, M.H., Yadav, V.K. et al. (2017). Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. *FEBS J.* 284 (3): 485–498.

60 Lei, Y., Li, S., Liu, Z. et al. (2021). A deep-learning framework for multi-level peptide-protein interaction prediction. *Nat. Commun.* 12 (1): 1–10.

61 Gibson, T.J., Dinkel, H., Van Roey, K., and Diella, F. (2015). Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Commun. Signal.* 13 (1).

18

Summary and Outlook

Volkhard Helms[1] *and Olga V. Kalinina*[1,2,3]

[1] *Saarland University, Center for Bioinformatics, Saarland Informatics Campus, Postfach 15 11 50, 66041 Saarbrücken, Germany*
[2] *Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), 66123 Saarbrücken, Germany*
[3] *Saarland University, Drug Bioinformatics, Medical Faculty, 66421 Homburg, Germany*

18.1 Technical State of the Art

For sure, in all areas (structural biology, interaction assays, spectroscopy, and data sets) we can expect to see continuous improvements of existing methods in terms of spatial and time resolution, sensitivity, and a continuous shift toward *in vivo* techniques and to condition-specific techniques. Also, scientists will continue to introduce novel techniques that enable us to probe interactions on a genome-scale rather than one-by-one.

Some types of protein interactions can already be efficiently approached by experiments at a genome-wide scale, e.g. those of proteins interacting with DNA using ChIP-Seq [1], or with RNA using eCLIP [2]. For decades, protein–protein interactions have been studied by yeast-two-hybrid (Y2H) and tandem affinity purification coupled with mass spectrometry (TAP/MS) techniques. Yet, these methods have their limitations, in that they often lack in sensitivity and cannot differentiate between different isoforms and posttranslational modifications. A promising novel opportunity has arisen in the form of direct protein sequencing using nanopore technologies, Edman degradation, and mass spectrometry that provide resolution down to the single-cell level [3, 4], but these methods are still far from maturity. In this light, computational tools that enable integration of sparse and multimodal experimental data are gaining importance, and machine learning, in particular deep learning, bears a great promise to fulfill this need.

18.2 Role of Machine Learning

The important contributions of machine learning to diverse areas of interactomics were mentioned in several chapters of this book, e.g. Chapters 2 and 3 introduce

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

machine learning approaches that can distinguish biological contacts from crystal contacts or that predict hotspots at binding interfaces. Chapter 4 (as well as other chapters) mention recent breakthroughs in protein structure prediction made by deep learning approaches [5, 6], and note that similar techniques will likely become helpful for predicting biomolecular contacts as well. Also mentioned in Chapter 4 is the potential of machine learning in deriving scoring functions. A special type of machine learning methods (Boltzmann generators, see [7]) have been developed to generate putative protein conformations along protein folding pathways as alluded to in Chapter 9. Chapter 15 then explores the application of machine learning to predict changes of binding affinity upon mutations as well as for predicting their phenotypic effects. Finally, Chapter 16 mentioned machine learning tools that predict Molecular Recognition Features (MoRFs) as a subgroup of intrinsically disordered regions. One can certainly anticipate that the role of machine learning and the diversity of its applications in the field of interactomics will steadily increase in the coming years.

The key challenge for powerful machine learning techniques, such as deep learning, in the field of interactomics currently lies in the sparsity of experimental data available for training. Additionally, these data often come from different experimental approaches. Luckily, standardized high-throughput experimental tools (as outlined in Section 18.1) should alleviate this problem in the future. The data sparsity issue can be also mitigated by emerging machine learning tools that allow to train models on very sparse and multimodal data. Machine learning should become particularly helpful in integrating data stemming from different sources.

18.3 Challenges

Experimental studies addressing protein interactions do not routinely characterize whether the involved proteins are subject to posttranslational modifications although, in the light of Chapter 17, such details appear to be of critical importance. A particularly noteworthy example is the C-terminal domain of RNA polymerase. It consists of multiple repeats of the peptide sequence YSPTSPS, in which five out of seven residues can be phosphorylated. Peck et al. [8] described how dynamic alterations in the phosphorylation status of these residues enable the formation of specific interactions with different regulatory proteins during the transcription cycle. Further examples mentioned, e.g. in Chapter 10 describe how phosphorylation affects transcription factor–DNA interactions.

More work is also needed to map the conditions under which, e.g. protein–protein interactions are relevant, and when they are not. It has become common practice to prune the global protein–protein interactome to a particular condition on the basis of transcriptomic (or ideally proteomic) information about which genes/proteins are expressed or not. The idea behind this is that proteins that are not present in a particular cell type cannot interact. But, of course, we also need to determine which protein isoform is expressed in that tissue or cell type and whether the isoforms are subject to PTMs in that particular condition.

Another fundamental challenge is to properly address protein interactions involving fully or partially disordered proteins. Such interactions are widespread, e.g. among RNA-binding proteins (RNAPs), but not well amenable to structural biology (X-ray and cryoEM) or to molecular dynamics simulations. Here, we point to a recent perspective article that addressed the potential of machine learning approaches in this respect [9].

What is currently also lacking is to better connect the different layers of interactomes. Proteins that interact with DNA, chromatin, and RNA are well-known to interact also with other proteins and with small molecules acting as effectors. But can proteins that interact with DNA also bind to RNA? Can they bind to membranes or to the cytoskeleton?

18.4 What Picture(s) May Evolve?

It is difficult to estimate to what fraction the protein interactome of an organism such as human or of a model plant such as *Arabidopsis thaliana* is known to date. Based on a Bayesian scheme for data integration, the developers of the repository PrePPI estimated in 2016 that there are between 127 000 and about 500 000 direct physical interactions of pairs of human proteins [10]. Yet, the question is also where we draw the line what contacts are considered as direct physical interactions. Chapter 12 presented the intriguing picture whereby practically all synthesized RNA molecules of a cell are constantly covered by many types of proteins during their entire lifetime. Hundreds of RNAPs can bind to individual mRNAs, often at many different places, and only concerted binding of several RNAPs can elicit the desired biological effect, such as, for example, mRNA processing. Can we expect the same multitude also with respect to interactions of proteins with DNA, membranes, and the cytoskeleton? In a particular human cell, a few hundreds of thousands of specific protein–protein contacts may exist with life times ranging from short-lived transient, yet specific interactions of redox proteins or between proteins of a signal transduction pathways, up to permanent assemblies such as ribosome or proteasome. Can we neglect that millions to billions of unspecific protein contacts transiently form in parallel to these specific assemblies? Do they give rise to particular phenomena due to their sheer number such as membrane rafts or molecular sponges around RNAs? Only time will tell us.

From a bioinformatics perspective, there will likely be plenty of work for us in the coming years. First of all, the experimental techniques are maturing so that the accumulated gold standard data sets are growing in size, coverage, and in quality. Hence, the relevance of data integration and machine learning is steadily rising. Although it may sometimes appear as if "the more we know, the more we don't know," looking at the current COVID-19 pandemic should provide us with good faith in the power of modern life sciences. Never before have scientists been able to respond so quickly to a deadly disease by mapping out the cellular and molecular consequences of a viral infection, characterizing the molecular interactions of the virus with host proteins on the surfaces of cells, and eventually even inventing novel vaccines. What helped,

in the end, was to understand how the spike protein of the envelope of the SARS-2 virus binds to the human ACE2 receptor. Do we need to say anything more why THIS BOOK addresses a timely topic?

References

1 Valouev, A., Johnson, D., Sundquist, A. et al. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* 5: 829–834. https://doi.org/10.1038/nmeth.1246.

2 Van Nostrand, E.L., Freese, P., Pratt, G.A. et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature* 583: 711–719. https://doi.org/10.1038/s41586-020-2077-3.

3 Alfaro, J.A., Bohländer, P., Dai, M. et al. (2021). The emerging landscape of single-molecule protein sequencing technologies. *Nat. Methods* 18: 604–617. https://doi.org/10.1038/s41592-021-01143-1.

4 Ouldali, H., Sarthak, K., Ensslen, T. et al. (2020). Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat. Biotechnol.* 38: 176–181. https://doi.org/10.1038/s41587-019-0345-2.

5 Baek, M., DiMaio, F., Anishchenko, I. et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373: 871–876. https://doi.org/10.1126/science.abj8754.

6 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589. https://doi.org/10.1038/s41586-021-03819-2.

7 Noé, F., Olsson, S., Köhler, J., and Wu, H. (2019). Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. *Science* 365: 6457.

8 Peck, S.A., Hughes, K.D., Victorino, J.F., and Mosley, A.L. (2019). Writing a wrong: coupled RNA polymerase II transcription and RNA quality control. *Wiley Interdiscip. Rev.: RNA* 10: e1529.

9 Lindorff-Larsen, K. and Kragelund, B.B. (2021). On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *J. Mol. Biol.* 433: 167196.

10 Garzón, J.I., Deng, L., Murray, D. et al. (2016). https://doi.org/10.7554/eLife.18715). A computational interactome and functional annotation for the human proteome. *eLife* 5: e18715.

Index

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

d

e

f