Learning Materials in Biosciences

Henrik Christensen *Editor*

Introduction to Bioinformatics in Microbiology

Springer

Learning Materials in Biosciences

Henrik Christensen
Editor

Introduction to Bioinformatics in Microbiology

Springer

Editor
Henrik Christensen
Department of Veterinary Animal Sciences
University of Copenhagen
Copenhagen, Denmark

ISSN 2509-6125 ISSN 2509-6133 (electronic)
Learning Materials in Biosciences
ISBN 978-3-319-99279-2 ISBN 978-3-319-99280-8 (eBook)
https://doi.org/10.1007/978-3-319-99280-8

Library of Congress Control Number: 2018961429

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Bioinformatics is the convergence of two trends in biological research since the 1960s: storage of molecular sequences in computer databases and application of computational algorithms to the analysis of DNA and protein sequences. Both sequence information accumulated in databases and computer capacities have been growing exponentially which explain why bioinformatics has been developing so fast as a scientific field.

Bioinformatics is integrating many scientific fields such as mathematics, computer science, and statistics. Other disciplines involve biochemistry, molecular biology, and physics. The microbiologists are the most important since they formulate the questions and hypothesis to be investigated. The bioinformaticians integrate all information, and bioinformatics has matured during the past two decades to become a scientific field of its own and so has the integration of bioinformatics with microbiology. Related to small and uniform size of most microorganisms, bioinformatics has revolutionized microbiological research by providing new insight into all aspects of microbiology. The analysis of all genomes (microbiome) from all organisms in a sample (microbiota) is designated metagenomics and has not only revolutionized the investigation of microbiology but also human medicine (human microbiome).

The aim of this book is to teach bioinformatics to microbiologists working within biotechnology, biology, veterinary medicine, clinical medicine and with environmental microbiology at graduate as well as postgraduate levels and equivalent continuous education. The book can be used for novices as well as scientists with some experience in the field. Some basic concepts have been explained very detailed to provide beginners with the proper understanding which allow the more complicated problems to be understood and the right conclusions drawn.

The book is based on material and experience gained from teaching the PhD course "Bioinformatics for Microbiology" during the past 15 years. Teaching has been performed by introductory lectures followed by computer exercises where participants have worked on cases including their own data. With a background in the course, they have worked out independent bioinformatics projects with realistic cases at the level of scientific publications.

The idea behind the course and also this book is to use the "freeware" computer programs without prepaid licenses. Readers will be able to start their own bioinformatics investigations just with a laptop computer and an internet connection. The book will for most users only be an introduction, and they will have to seek further teaching including self-teaching to continue with the more advanced bioinformatics pipelines as well as writing of bioinformatics programs which will not be covered in this book.

I acknowledge my colleagues at the Danish Informatics Network in Agriculture who took the initiative to the first bioinformatics PhD course level training in 1999. I also acknowledge the more than 200 PhD level students who have attended the course "Bioinformatics for Microbiology." They have contributed to a continuous development of

the course with their personal experience from active research projects, curiosity, and enthusiasm. Some former participants have become coauthors on this book, and they in particular are thanked for their work. Finally I would like to acknowledge editor Silvia Herold and project coordinator Srinivasan Manavalan at Springer Nature for inviting me to write the book and to provide optimal support during the process.

Henrik Christensen
Copenhagen, Denmark

Contents

Contributors

Anna Jasmine Andersson
HovedOrtoCentret, Øjenklinikken,
Øjenforskningen
Glostrup, Denmark
anna.jasmine.andersson.01@regionh.dk

Henrik Christensen
Department of Veterinary Animal Sciences
University of Copenhagen
Copenhagen, Denmark
hech@sund.ku.dk

Lisbeth E. de Vries
Københavns Professionshøjskole
Copenhagen, Denmark
ledv90@hotmail.com

Steffen Lynge Jørgensen
Department of Veterinary Animal Sciences
University of Copenhagen
Copenhagen, Denmark
Steffen@sund.ku.dk

Arshnee Moodley
Department of Veterinary Animal Sciences
University of Copenhagen
Copenhagen, Denmark
asm@sund.ku.dk

John Elmerdahl Olsen
Department of Veterinary Animal Sciences
University of Copenhagen
Copenhagen, Denmark
jeo@sund.ku.dk

Rikke Heidemann Olsen
Department of Veterinary Animal Sciences
University of Copenhagen
Copenhagen, Denmark
cava@sund.ku.dk

Josef Korbinian Vogt
National Food Institute, Technical
University of Denmark
Lyngby, Denmark
jkovo@food.dtu.dk

Abbreviations

ANI	Average nucleotide identity
ASN.1	Abstract Syntax Notation One
BAC	Bacterial artificial chromosome
BChl	Bacteriochlorophyll
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks substitution matrix
CAZyme	Carbohydrate-active enzymes
CC	Clonal complex
CDT	Cytolethal distending toxin
DDBJ	DNA Data Bank of Japan
dDDN	Digital DDN
DDN	DNA-DNA reassociation
ddNTPs	Dideoxynucleotides
DE	Differentially expressed
DLV	Double-locus variants
dN	Non-synonymous nucleotide substitutions per non-synonymous site
DRISEE	Duplicate read inferred sequencing error estimation
dS	Synonymous nucleotide substitutions per synonymous site
EBI	European Bioinformatics Institute
ENA	European Nucleotide Archive
FPKM	Fragments Per Kilobase Million
GGDC	Genome-to-genome distance calculator
HGT	Horizontal gene transfer
HSP	High-scoring segment pairs
INSD	International Nucleotide Sequence Databases
INSDC	International Nucleotide Sequence Database Collaboration
ISFET	Ion-sensitive field-effect transistor
MALDI-TOFMS	Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry
MS	Mass spectrometry
MIGS	Minimum information about a genome sequence
MLSA	Multilocus sequence analysis
MLST	Multilocus sequence typing
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
NIG	National Institute of Genetics
ORF	Open reading frames
OTU	Operational taxonomic units
PAM	Percent accepted mutations
PCA	Principal component analysis
PCoA	Principal coordinates analysis
PERMANOVA	Permutational multivariate analysis of variance
PDB	Protein Data Bank
PIR	Protein Information Resource
qPCR	Quantitative PCR
RAST	Rapid Annotation using Subsystem Technology
RBH	Reciprocal best hits
RDP	Ribosomal Database Project
RPKM	Reads Per Kilobase Million
SGML	Standard Generalized Markup Language
SLV	Single-locus variants
SMART	Simple Modular Architecture Research Tool
SNP	Single-nucleotide polymorphism

SRA	Sequence Read Archive
ST	Sequence type
TLV	Triple-locus variants
TPA	Third party annotation
TPM	Transcripts Per Million
USS	Uptake signal sequences
wgMLST	Whole-genome MLST
XML	Extensible Markup Language
ZMW	Zero-mode waveguides

Introduction

What the Book Will Teach You

Henrik Christensen

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_1

1

What You Will Learn in This Chapter

You will learn why bioinformatics became an independent scientific discipline and why bioinformatics is of particular relevance to microbiology. You will then be provided with the background of this book as well as its major outline. At the end, you will get information about computers, operating systems, and computer programs required to carry out bioinformatics. The book has many practical relevant examples.

1.1 Basics and History of Bioinformatics

Bioinformatics has become a major scientific field within biological sciences. This includes microbiology in particular related to the small and often uniform size of the organisms investigated which promoted molecular investigation early on. The use of molecular biology including DNA sequencing has been essential to microbiological investigations. The first literature references to bioinformatics date back to the early 1990s, and the term came into common use in the late 1990s. Bioinformatics is mainly the convergence of two trends in biological research: storage of molecular sequences in computer databases and application of computational algorithms to the analysis of DNA and protein sequences (Bogusky 1998). Now it has become a routine in most labs not only to sequence single genes but also whole genomes and in many labs to use metagenomics based on 16S rRNA amplicon sequencing. Scientists and students are automatically assumed to know how to analyze this immense information. The book will teach you how to handle this problem top-down. The bipartition of bioinformatics between the databases and the computers explains why bioinformatics has grown and is growing as a scientific field – both sequence information accumulated in databases and computer capacities have been and are still growing exponentially.

1.1.1 Bioinformatics Is Integrating Many Scientific Fields

Bioinformatics is integrating many scientific fields (Fig. 1.1), and a major task of the "information" part of bioinformatics relates to the understanding and translation of terms and concepts between these scientific disciplines. It all starts with a biological problem to

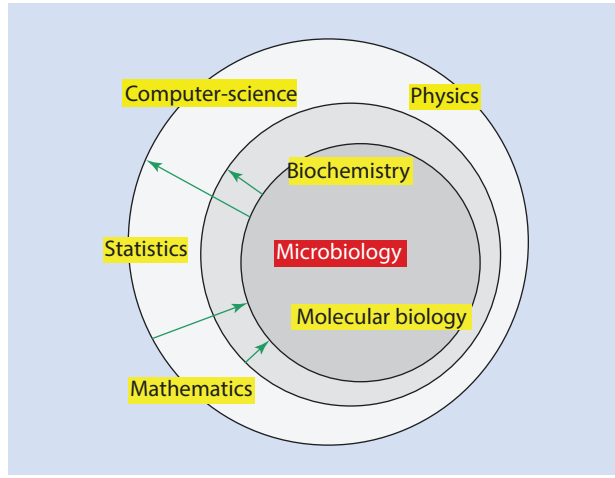


Fig. 1.1 Good communication between the different scientific disciplines is essential for bioinformaticians to do research

be solved. In the context of this book, the microbiologist formulates the biological question and hypothesis. For instance, if a microbiologist wants to use an enzyme from seawater for the production of "plastic" (synthetic polymers), a biochemist may need to set up tests to identify the synthetic polymers which are related to the organism or sample. Computer programs are used which include algorithms formulated by mathematicians and implemented by computer scientists to analyze the data and identify relevant protein sequences that have significant similarity to enzymes already known to be involved in the specific polymerization. The interaction with molecular biologists and physicists is needed to further interpret and annotate the information about the enzyme in the databases. The bioinformatician will have to integrate all information into one concept and know of all scientific disciplines needed. Bioinformaticians often have their background in one of the traditional disciplines such as computer science, biochemistry, and physics – or microbiology.

1.1.2 Homology

A key biological concept behind bioinformatics is homology. Homologous structures have the same ancestor or are predicted to have the same ancestor. The different bones in a wing of a bird and in a human arm can be compared since they are thought to belong to common structures in a common ancestor even though this common structure is several 100 millions of age. DNA and protein sequences also only can be compared in a meaningful way if they are homologs, i.e., if they once belonged to a common ancestor. The criteria of morphological homology were defined by Remane (1952). It is not needed to read this German text unless you love comparative morphology since the comparison of molecular sequences has provided us with tools to model ancestry which are many times stronger and faster than the comparison of morphologic characteristics. One reason for the success of bioinformatics has been that we are able to model evolution and test homology. The homology concept can be subdivided into orthology and paralogy. These concepts were defined by Walter M. Fitch and have become key terms used in bioinformatics (Fitch 2000). Ortholog genes or proteins are homologs that diverged following speciation events, whereas paralogs (para for parallel) are homologs which diverged as a consequence of gene duplication. In prokaryotes, multiple copies of a gene with the same function are usually not tolerated, and as a consequence, paralogs in the same species will have different functions.

An example is shown in ◘ Fig. 1.2 of the phylogeny of the proteins in the toxin cytolethal distending toxin (CDT). CDT is produced by some bacterial pathogens like species of *Campylobacter* and *Haemophilus*. CDT has not been found in Gram-positive bacteria. CDT can induce G2/M cell cycle arrest, chromatin fragmentation, cell distention, and nucleus enlargement of eukaryotic cells. CDT is a heterotrimer consisting of CDTA, CDTB, and CDTC protein chains. The three proteins are homologs but distantly related. From the figure is seen that all the CDTA proteins are related in the different species compared, and they are therefore orthologs. Also the CDTB proteins are related in different species, and they are also orthologs according to the definition. However, the CDTA and CDTB proteins are more distantly related, and they are paralogs since they are present in the same species but with different functions. The phylogenetic relationships are explained in more detail in ▶ Chap. 6.

Another example of the use of phylogeny is for the classification of all prokaryotes based on phylogenetic analysis of the 16S rRNA gene sequences. The 16S rRNA-based

1

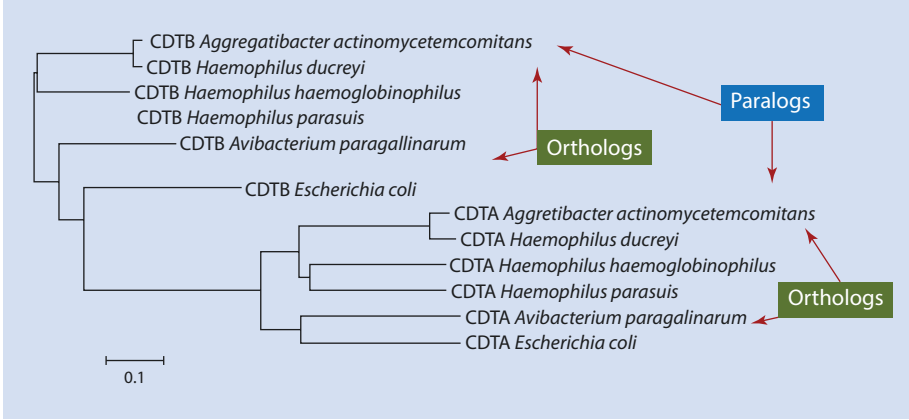


Fig. 1.2 The concepts of orthology and paralogy illustrated by phylogenetic relationships between protein sequences A and B of the cytolethal distending toxin (CDT)

phylogeny has changed the classification of all prokaryotes dramatically during the past two decades and as a consequence also how identification of prokaryotes is performed by sequencing (▶ Chaps. 7 and 8).

1.1.3 Evolution Is Related to Sequence Diversity Caused by Mutations

The observed differences between related DNA sequences are related to mutations. A mutation is a stable inheritable alteration of the genome. There are two types of mutations, point mutations and mutations involving larger fragments of DNA. Point mutations involve only one nucleotide. Point mutations are rare events related to mistakes in DNA replication. If such a mutation is changing the codon of an amino acid (non-synonymous substitution) of a protein, it may lead to a change in the function of the organism or be fatal for the organisms. Other point mutation may not result in the change of an amino acid (synonymous substitution); however, they can of course still be registered at the level of DNA. Point mutations are distinguished from changes of larger fragments of DNA that may have been exchanged between different organisms, and such changes are referred to as horizontal gene transfer (▶ Chap. 11).

1.2 The Aim, Structure, and Outline of the Book

The aim of the book is to provide readers with a background to carry out bioinformatics investigations within microbiology which will provide sufficient scientific background to publish their investigations in scientific peer-reviewed journals. The weight on bioinformatics teaching in this book has been understanding, introduction, use of proper controls to confirm results, evaluating the quality of analysis, and graduating the precision of results required. In some cases, only round figures are needed, for example, when the similarity between CDTA and CDTB proteins is compared to each other in Fig. 1.2 since they are so distantly related. In other cases, results need several digits to be significant,

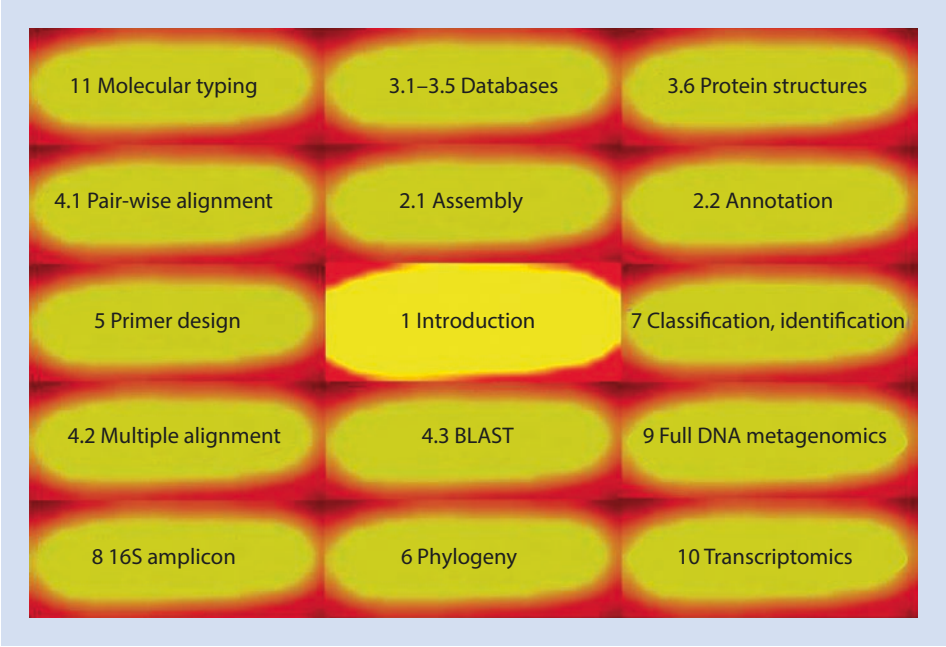


Fig. 1.3 The structure of the book. The topics are illustrated as bricks which can be used and combined to construct different types of analysis. The highlighted bricks are most closely related to the actual chapter shown with yellow

for instance, when 16S rRNA gene sequences of bacterial species are compared for the purpose of identification (▶ Chap. 7). Members of the prokaryotes have the main focus in the book since we have mainly worked with this group including the examples provided. The book is still relevant for microbiologists investigating fungi and virus.

Figure 1.3 shows the main topics of bioinformatics treated in this book. This structure has been used in many bioinformatics textbooks. After the current introductory chapter, the book starts with sequence assembly and annotation. Annotation is actually a consequence of the databases covered in ▶ Chap. 3. ▶ Chapter 4 introduces substitution matrices and pair-wise and multiple alignments used for comparing sequences. ▶ Chapter 4 also introduces BLAST which is used to compare sequences in the databases described in ▶ Chap. 3. In ▶ Chap. 5, primer design is described. This subject is important both for the generation of new sequences (▶ Chap. 2) and for diagnostic purposes. The chapter includes a distinction between primers used for PCR and oligonucleotides probes used for hybridization proposes. In ▶ Chap. 6, phylogeny is introduced which requires ▶ Chap. 4 about the multiple alignment as background. In ▶ Chap. 7, sequence-based classification and identification is introduced which includes backgrounds from all former chapters. The identification of prokaryotes has been dominated by 16S rRNA gene sequence-based identification for the past two decades. 16S rRNA gene sequence-based identification and classification is the background for 16S rRNA amplicon sequencing which currently is the most frequently used method to characterize prokaryotic communities (▶ Chap. 8). Full DNA metagenomics also called shotgun metagenomics is described in ▶ Chap. 9. In ▶ Chap. 10, RNA-based transcriptomics is described which allows an estimation of the relative transcriptional level of different genes. This heavily relies on the databases in

1

► Chap. 3. Finally, ► Chap. 11 describes molecular typing which is relying both on multilocus gene sequence typing (MLST) and whole genome analysis either by use of the single nucleotide approach (SNP) or by whole genomic MLST (wgMLST).

The aim is to provide readers with more understanding and inspiration of bioinformatics by reading this book which they may then use as a stepping stone toward further and deeper bioinformatics investigation.

Readers will have to be introduced to microbiology in other relevant textbooks such as *Brock Biology of Microorganisms* edited by Michael Madigan and coworkers (Madigan et al. 2019). Some readers may also need to consult basic textbooks about biochemistry and molecular biology.

1.3 Computers and Operating Systems Required for Bioinformatics

Bioinformatics is practiced with a computer. If you have not worked with bioinformatics before, you need to choose the optimal computer to solve a given problem. For a start, your favorite laptop will probably do the job. The computer needs to have at least 50 Gbyte free space on the hard disk and a reasonable amount of RAM (ready access memory) where 8 Gbyte is minimum and 16 Gbyte preferred – the more RAM, the better. If you do a lot of bioinformatics, it will be beneficial with a desktop computer with 32 Gbyte RAM. For beginners, operating systems can both be of the "Mac" type provided by Apple® and "Windows" type by Microsoft® (◘ Fig. 1.4). Most of the basics bioinformatics programs

◘ **Fig. 1.4** A collection of operating systems and computers. SuperUsers is acknowledged for this historical photo (► www.superusers.dk)

will be available for both types. Such programs will also usually run most smoothly on the Linux type of operating systems such as Ubuntu. However, it is only recommended to start on Ubuntu if you have some experience with this operating systems already or else you risk to use most of the time learning Linux instead of learning bioinformatics. Many bioinformatics tools can be performed for free on servers which can be accessed through the Internet. The same holds true for the databases, and a good Internet connection is mandatory. Readers of this book will be able to start their own bioinformatics analysis just with a laptop computer and an Internet connection.

1.4 Computer Programs and Pipelines

The majority of the best computer programs for bioinformatics are noncommercial in the way that they are free to use for nonprofit purposes. The weight will therefore be on free programs in this book. Unfortunately, most free programs are with command line interface which is unfamiliar for most beginners in this field. Especially for sequence assembly, it has been difficult to suggest free programs with a graphical user interface, and in this case, suggestions to use programs with a user fee have been made.

1.5 Activity

How much RAM is in your computer?

Windows 7: Left click on **Start**, right click on **Computer**, and select **Properties**.

Windows 10: Left click on the **Windows start icon**, left click on **Computer**, right click on **Computer**, and left click on **Properties**.

Mac: Click on the **Apple icon** in top left corner and click **About this Mac**.

Ubuntu: Left click at the settings **icon (cogwheel)** and left click at **Details (cogwheel)** under System.

Take-Home Messages

- Bioinformatics developed as the convergence of two trends in biological research: storage of molecular sequences in computer databases and application of computational algorithms to the analysis of DNA and protein sequences.
- Bioinformatics is integrating most of the scientific disciplines in natural sciences, and a major task of the "information" part of bioinformatics relates to the understanding and translation of terms and concepts between these scientific fields.
- A major reason for the success of bioinformatics is that we are able to model evolution and test homology.
- The majority of the best computer programs for bioinformatics are noncommercial in the way that they are free to use for nonprofit purposes, and many bioinformatics tools can be used for free on servers which can be accessed through the Internet.

References

Bogusky, M. S. 1998. Bioinformatics - a new era. Trends Guide to Bioinformatics. Trends Supplement. Elsevier, p. 1–3.

Fitch, W.M. 2000. Homology a personal view on some of the problems. Trends Genet. 16, 227–231.

Madigan, M., Bender, K., Buckley, D., Sattley, W. M. & Stahl, D. 2019. Brock biology of microorganisms. 15th ed. Pearson, New York.

Remane, A. 1952. Die Grundlagen de natürlichen Systems, der vergleichenden Anatomie und er Phylogenetik. Akad. Verlagsges. Geest & Portig, Leipzig.

Further Reading

Basic Bioinformatics Text Books

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. 1999. Biological sequence analysis. Cambridge Univ. Press.

Nei, M. & Kumar, S. 2000. Molecular evolution and phylogenetics. Oxford University Press.

Zvelebbil, M. & Baum, J. O. 2008. Understanding bioinformatics. Garland Science, New York.

DNA Sequence Assembly and Annotation of Genes

How to Generate the DNA Sequence and to Predict the Function of Genes

Henrik Christensen and Arshnee Moodley

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_2

2

What You Will Learn in This Chapter

In this chapter, you will learn about the different sequencing strategies currently available and the pros and cons of the different strategies to help you select the optimal DNA sequencing strategy for your research question. Thereafter, you are introduced to the challenges associated with assembly of DNA sequence data and the different quality criteria needed in the process including how genomes can be closed. The simple FASTA format used for representing DNA sequence is demonstrated, and annotation of the DNA sequence is then introduced.

2.1 DNA Sequencing

DNA sequencing is the determination of the order of nucleotides of parts or whole chromosomes of organisms and virus. DNA sequencing can be done for a single gene or a whole genome or many genomes at a time such as in metagenomics. The first decision to be taken is the selection of the optimal DNA sequencing strategy for your research question in your project. In ◘ Tables 2.1 and 2.2, we summarize the different technologies and what they could be used for and their benchmarks. Often laboratories outsource the sequencing to companies or institutions if they do not have their own in house sequencer, resulting in a longer turnaround time and higher sequencing costs. Therefore, research labs are now investing in smaller, more affordable benchtop next-generation sequencers for more cost-efficient sequencing (◘ Fig. 2.1).

The development of DNA sequencing technology has a long history with multiple leaps occurring within a few decades. In 1973, Maxim and Gilbert published the first

◘ **Table 2.1** Selection of DNA sequencing strategies and methods in relation to capacity and cost

Problem	Method	Read length (nt)	Capacity	Cost (related to capacity)
1. Single genes, sequencing of BAC, FOS[a] or plasmid inserts (cloning)	Sanger sequencing of PCR amplicons	500–1000	Low	High
2. Genomes and 16S metagenomics	Illumina	150–600	High	Low
	Ion torrent	400–600	High	Low
3. Single-cell sequencing	Single-cell dissection, DNA extraction, phi29 amplification, and Illumina sequencing	150–251	High	High
4. Closing of single genomes	PacBio® (Pacific Biosciences)	15,000–100,000	Medium	High
	Nanopore	1000,000	Medium	High

[a]*BAC* bacterial artificial chromosome, *FOS* fosmid, based on bacterial F-plasmid. Both are used for cloning of large fragments of DNA in the order of 50 kbases

Table 2.2 Comparison of different commonly used sequenced platforms

Platform	Read length	Amount of data produced	Number of reads	Run time	Accuracy	Cost of instrument (US$)	Cost per Gb (approx. US$)
Illumina MiSeq v3 600 bp	300 bp PE	13.2–15 Gb	44–50 million PE	56 h	99.9%	99,000	110
Illumina HiSeq 2500 v2	250 bp PE	125–150 Gb	600 million PE	60 h	99.9%	690,000	45
Ion Torrent Ion PGM 318	400 bp SE	1–2 Gb	4–5.5 million	7.3 h	98–99%	49,000	450–800
PacBio RS II	Approx. 20 Kb	500 Mb–1Gb	55,000	4 h	87%	695,000	1000
MinIon	Variable up to 900 Kb	5 Gb	Up to 1 million	Up to 48 h	88%	1000	500–900

Table derived from Goodwin et al. (2016) and ► https://blog.genohub.com/2017/06/16/pacbio-vs-oxford-nanopore-sequencing/
PE paired end, *SE* single end, *Gb* gigabase, *Mb* megabase, *Kb* kilobase

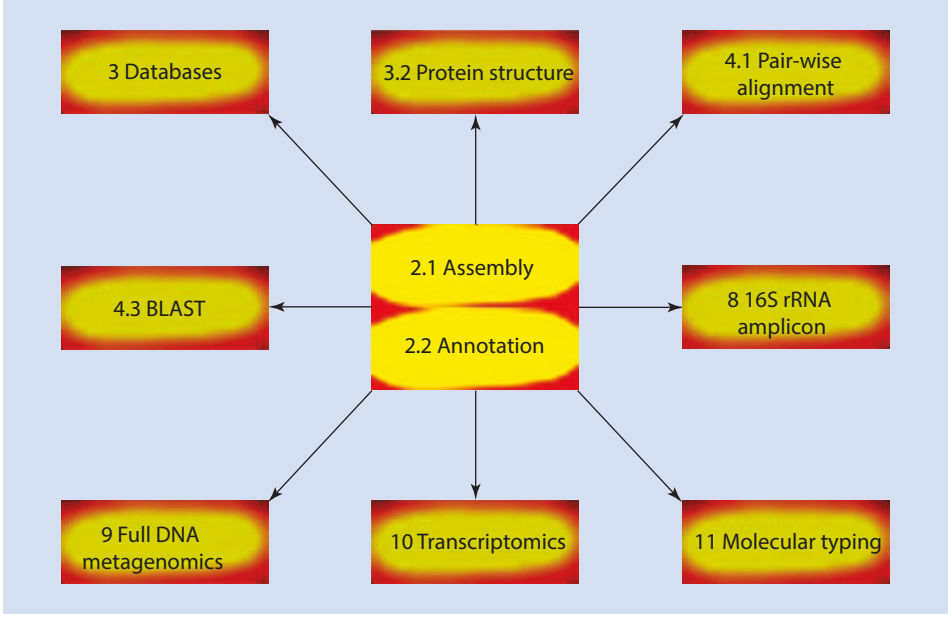


Fig. 2.1 DNA sequence assembly and annotation are related to most other chapters in this book by providing the raw DNA data and information about the location of genes on chromosomes and their functional and taxonomic relationships

ever DNA sequence, 24 bp of the *lacR* binding site. It took 2 years, meaning one base per month. In 1976, Sanger and Coulson described the chain-terminator method, which is based on the inclusion of one of the four radioactive-labeled dideoxynucleotides in a reaction and when incorporated would stop strand elongation (chain termination), producing fragments of different lengths (▶ Sect. 2.1.1). All four reactions would then be resolved individually by electrophoresis on a polyacrylamide gel (one lane, one dideoxynucleotide base). The X-ray gel picture would resemble a ladder from which the sequence could be read off. Traditional Sanger sequencing has the limitation in that sequencing larger genomes, e.g., a 3300 Mb human genome, is rather challenging leading to the development of next-generation DNA sequencing (NGS) or massive parallel sequencing of small DNA fragments. The key difference between Sanger sequencing and NGS is the ability to multiplex.

The first high-throughput sequencing platform that became available was 454 GS20 pyrosequencer marketed by Roche in 2005 (later replaced by model GSFLX). 454 sequencing is no longer widely used as the company has stopped sales of this instrument. However, data generated by this method are available in databases, and from a bioinformatical perspective, it is relevant to know about the platform. This is similar for data derived from an Ion Torrent system, which will be explained further down. When we are dealing with handling of the data, the 454/Ion Torrent-derived data is very similar to Illumina data since sequencing is done by replication of DNA and millions of short DNA reads are generated which need to be assembled to the final DNA sequence.

2.1.1 Sanger Sequencing

This sequencing method was developed in the 1970s (Sanger et al. 1977) and was the predominant method until the mid-2000s. Now it is mainly used as a method for sequencing single genes or fragments of DNA. The principle of the method is based on sequencing by replication of DNA and the incorporation of dideoxynucleotides (ddNTPs) which will stop replication randomly when one of the four dideoxynucleotides, ddATP, ddCTP, ddGTP, or ddTTP, is incorporated. The position of the stops is determined based on the specific radioactive or fluorescent label of each ddNTP to determine the DNA sequence. As mentioned, the initial method used radioactive labeling and polyacrylamide gel electrophoresis which was very time demanding. By the mid-1980s, automated fluorescence-based Sanger sequencing machines were developed, e.g., those by Applied Biosystems. Current methods are based on fluorescent labeling and capillary gel electrophoresis in automatic instruments. Further information of this method can be found in textbooks (Madigan et al. 2019).

2.1.2 Massive Parallel, Short-Read Sequencing

2.1.2.1 Illumina (Sequencing by Synthesis)

More space will be used to introduce Illumina sequencing since it is the most frequently used platform today. Illumina acquired Solexa GA in 2007 to develop and commercialize genome sequencing technology. The company currently has a number of different sequencers depending on the sequencing capacity needed, ranging from benchtop sequencers to production-scale sequencers. Microbiology research labs would invest in

one of the benchtop sequencers, in which at the moment the most common platform is the MiSeq (Fig. 2.2). Smaller systems such as the MiniSeq and iSeq 100 are also available. These are cheaper than the MiSeq but have a limited sequencing capacity. More information can be found on the company website: ▶ https://www.illumina.com/systems/sequencing-platforms.html. We will only focus on the MiSeq, but the principle of sequencing is the same across the machines.

The MiSeq is capable of doing small whole-genome sequencing, transcriptomics, 16S metagenomics, and DNA-protein interaction analysis (ChIP-Seq). It is possible to multiplex by using unique combinations of specific barcodes and indexes (see ▶ Chap. 8 for the use of barcodes and indexes). One of the first steps of sequencing is the fragmentation of DNA. Depending on the library preparation kit used, the DNA can be fragmented manually, e.g., using a Covaris instrument or an enzymatic fragmentation, e.g., transposon-mediated fragmentation as in the Nextera XT kit. This process is called tagmentation. For sequencing of small genomes, the Nextera XT kit is widely used since the library preparation protocol is simple. DNA is tagmentation which used two transposons that randomly cuts the DNA. During this tagmentation process, the transposon attaches adaptor sequences to the ends of the fragments. These sequences are then used as primer binding sites for the inclusion of indexes and the adaptors that are needed to anchor the DNA to a solid support which is the surface of the flow cell (Fig. 2.3).

From each template of DNA, "spots" or "cluster" is generated that is complementary to the original DNA template sequence. This is done by "bridge amplification" or "fold-back" step. This step is crucial to generate sufficient signal to be detected in the sequencing process. Sequencing is then done by the addition of fluorescently labeled nucleotides in the presence of DNA polymerase in cycles. This way, a complementary strand is generated, and for each cycle, which includes a wash step, the flow cell is imaged, and the emission from each cluster is recorded. The wavelength and intensity are used to identify the base. This is repeated a number of cycles depending on the flow cell used, e.g., V3 600 bp (600 cycles or 2 × 300 cycles for paired end sequencing). The final DNA sequence is then generated by comparing all images as a stack.

Aside from Illumina sequencers, there is the Ion Torrent system by Thermo Fisher Scientific. This is also a benchtop sequencer released in 2010. Similar to Illumina, the technology is based on sequencing by synthesis, but the detection method is different.

Fig. 2.2 A picture of our Miseq instrument (Illumina)

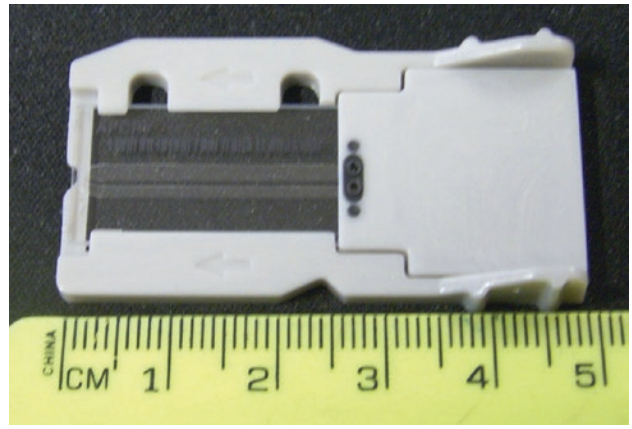


Fig. 2.3 The flow cell from Miseq (Illumina)

The Ion Torrent system uses ion semiconductor sequencing technology, which detects hydrogen ions released during DNA polymerization. After DNA amplification, each microwell contains a single species of template DNA. Unmodified nucleotides are washed into the microwells one dNTP species at a time. When a nucleotide is incorporated into a complementary strand, a covalent bond is formed, and a phosphate and hydrogen ion are released. This release of a hydrogen ion changes the pH, and this change is detected by an ion-sensitive field-effect transistor (ISFET) ion sensor on the chip. The major benefit of the Ion Torrent is the low cost of the instrument since unmodified nucleotides are used and no optics needed. One of the major limitations to this method is the correct identification of homopolymers, i.e., stretches of repeating nucleotides. If homopolymers are present, this means multiple nucleotides are incorporated and hence the release of more hydrogen ions in a single cycle. This results in a greater pH change and a greater electronic signal detected. The system is not able to determine exactly how many nucleotides were included.

2.1.3 DNA Sequencing in Metagenomic and for Single-Cell Sequencing

The principle of sequencing DNA extracted for metagenomics and for single-cell sequencing is the same as described above for single genomes. Further explanation is given in ► Chaps. 8 and 9 including the use of index primers to perform Illumina sequencing with multiple samples.

2.1.4 Real-Time, Single-Molecule Sequencing

All of the platforms that are capable of massive parallel sequencing (► Sect. 2.1.2) require fragmentation and amplification of the DNA. These steps can introduce errors, and there are sequence-dependent biases. Furthermore, these steps add to the overall turnaround time for a sequencing run. Real-time, single-molecule sequencing allows for sequencing of the native DNA, resulting in significantly longer read lengths and sequence information available when the bases are incorporated, i.e., information available in real time. The first

of this type of technology was developed by Pacific Biosciences, and the PacBio RS system was made available in 2011. The second approach is nanopore sequencing by Oxford Nanopore Technologies, who released the MinIon in 2014.

2.1.4.1 PacBio®

PacBio is built on the ability to optically observe polymerase-mediated synthesis in real time. On a SMRT® cell, a single polymerase is fixed at the bottom of a microwell (also called a zero-mode waveguides, ZMW) with a single DNA template. Four fluorescently labeled nucleotides are introduced. When a labeled nucleotide is incorporated, the fluorophore is cleaved, and a light pulse corresponding to the incorporated base is produced. Each pulse has its own color intensity, and hence the type of base is identified. This technology produces long read lengths, on average >15Kb, with some reads >100Kb. These long reads can be used to assist with closing small genomes. Furthermore, PacBio sequencing can be used to span long, repetitive regions and can also be used to determine the methylome of microorganisms (methylated nucleotides) used for so-called epigenetics. The major drawbacks to the technology are the high sequencing costs and high error rate.

2.1.4.2 Nanopore Sequencing

Nanopores are small holes created by pore-forming proteins in a membrane. The principle of nanopore sequencing is the direct detection of the nucleotide sequence when a single-stranded DNA molecule passes through the nanopore and a current is generated by a flow of ions across the flow. When the template is threaded through the pore, a voltage block occurs that affects the current passing through the pore. These changes are characteristic of a particular nucleotide base. A MinIon has 512 nanopores, with each pore capable of sequencing approx. 70 bp/s. The read lengths produced are comparable to PacBio. Nanopore sequencing has been brought to the market by Oxford Nanopore Technologies who released the MinIon in 2014. The uniqueness of this system is the extreme portability of the device. The MinIon is a USB sequencer that plugs directly to a computer. Sequencing is cheaper compared to PacBio but had initially the same high error probability drawback.

2.2 DNA Sequence Assembly

Base calling is the first step in sequencing where the electronic signal generated in the sequencing machine is separated from random noise and converted to nucleotide information. Then the nucleotide information needs to be assembled to DNA sequences which resemble the original DNA sequenced as best as possible. This can either be done de novo without a reference or with a reference if the genome of the organism or virus is well known.

2.2.1 Base Calling and Trimming

Base calling is the conversion of electronic signal generated in the sequencing machine to the sequence of nucleotides in the DNA sequence. First, the signal needs to be quality controlled. Only a certain threshold passes the quality control. The *phred* index (*q*, Q, or qual) defined by Ewing et al. (1998) and for ABI data manipulation (automatic Sanger) is one way of expressing the quality of signal and now has become a standard. Another way

is the error probability (p) of the signal that each nucleotide has to be determined with. The two, q and p, are related by

$$q = -10 \times \log 10(p)$$

An error probability of 0.01 equals *phred* 20 (Q10 = 10%, Q20 = 1%, etc.). In the CLC Genomic Workbench program, p is used. The quality thresholds are used to filter the output from the sequencing machine and are referred to as trimming. Stand-alone trimming programs are available such as trimmomatic (Bolger et al. 2014).

2.2.2 Assembly of DNA Sequences

The output from the DNA sequencing machine is short single "reads," a few hundred nucleotides in length that need to be combined to mimic the original DNA that was subjected to sequencing which is often several millions in length. After the reads have been trimmed as described above, they need to be assembled. At least an overlap of forward and reverse reads needs to be generated and assembled with Sanger sequencing compared to the original DNA sequence (Table 2.3).

With high-throughput Illumina sequencing, many times more reads need to be combined compared to the original DNA being sequenced. This relates to short read length and to high error rates with individual nucleotides of the reads.

During assembly, the short DNA reads from the sequencing machine are combined to long DNA strands which equal the original DNA which was sequenced and which can be used for gene prediction (annotation below). The goal is always to assemble the longest DNA strand as possible, and such a strand is called a contig (a sequenced contiguous region of DNA). Ideally a contig should span the entire chromosome.

Table 2.3 Assembly programs

Program	Principle	Use	Reference
CLC Main Workbench[a]	Overlap consensus	Single genes	► https://www.qiagenbioinformatics.com/products/clc-main-workbench/
Velvet	k-mer and de Bruijn graph	Genomes	Zerbino and Birney (2008)
SPAdes	k-mer and de Bruijn graph	Genomes and single-cell sequencing	Nurk et al. (2013)
DTU server assembly	Velvet, SPAdes	Genomes	Larsen et al. (2012) ► https://cge.cbs.dtu.dk/services/Assembler/
CLC Genomics Workbench[a]	NI	Genomes and 16S rRNA amplicons	► https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/

[a]Requires fee

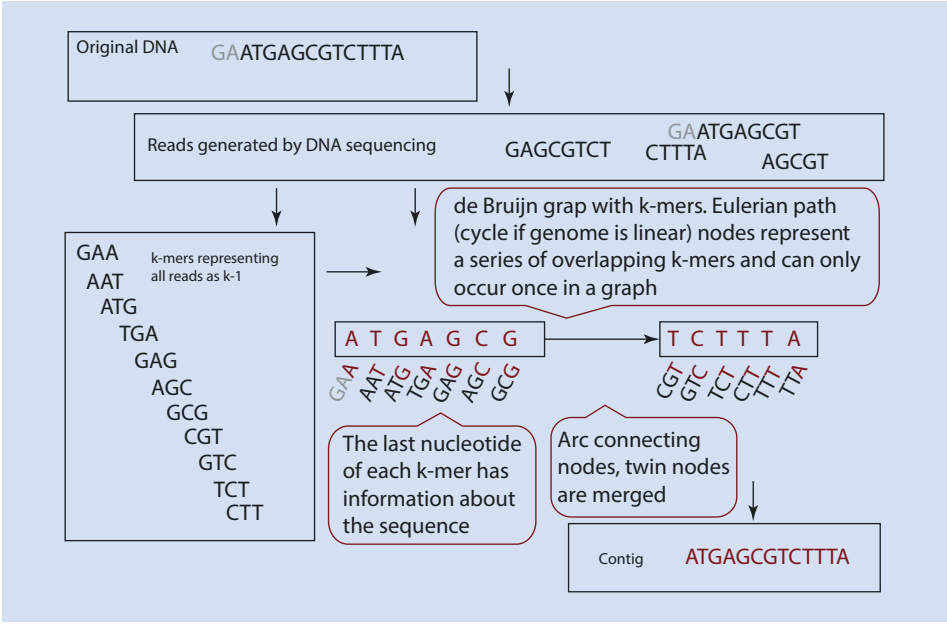
2.2.2.1 Assembly by Overlap Consensus Methods

Overlap and consensus methods are used for assembly of sequences generated by Sanger sequencing. Sequence reads are compared and matched by principles which equal pairwise alignments (▶ Chap. 4). Unfortunately assembly programs are not freely available for Windows or Apple operating systems. Reads generated by Sanger sequencing can be assembled by CLC Main Workbench which requires a minor fee.

2.2.2.2 Assembly by k-mer Strategy

Short "words" of DNA sequence called k-mers (length k) are used as an intermedial tool to combine the single reads from the sequencing machine to the DNA sequence of the organisms or virus originally sequenced. This method is most relevant to whole genomic sequences and to metagenomic datasets. The assembly of such datasets would not be possible by use of the "overlap" strategy above since it would be too demanding for the computers. This method originated in attempts to sequence by hybridization where the k-mers were used to link specific sequence motifs to hybridization probes (Idury and Waterman 1995). Sequence assembly is based on a so-called sequence graph (◘ Fig. 2.4). The k-mers will cover the DNA sequences observed in all of the reads generated by the sequencing machine. All k-mers are shifted one nucleotide to the right (1-tuples) so that their overlap is k-1. A popular k-mer size has been 31 since it fits with 4 base representation as a 64 bit integer. The idea is then to label a read on the k-mers and to visit all edges only once. This way, contigs are formed and shorter contigs are joined together (◘ Fig. 2.4).

Next to the handling of large datasets, the problems of repeats and sequencing errors can also be handled by use of Euler's algorithm and de Bruijn graphs. Errors are handled as



◘ **Fig. 2.4** Principles of assembly strategies based on Euler principle using k-mers resulting in a so-called de Bruijn graph. Only short fragments are shown to illustrate the principle. Only assembly of the forward DNA sequence is shown. Linear genomes need to be assembled by finding an Eulerian path. (Modified from Compeau et al. 2011)

"bulges" in the de Bruijn graph (Pevzner et al. 2001). The time to run computer implementation of Euler's algorithm is roughly proportional to the number of edges in the de Bruijn graph construction for the particular problem. The de novo assembly can be extended by scaffolding where the information from the paired end reads is used to link contigs. In ◘ Table 2.3, a comparison of different assembly programs has been made.

2.2.2.3 Quality Criteria of the Final Assembled DNA Sequence

The minimal requirements to assemblies are summarized from Chun et al. (2018).

The first important parameter to consider is the coverage = (number of reads × average read length)/genome size

The genome size will often be provided as the sum of contigs from assembly programs. For taxonomy, more than 50 times coverage is recommended (Chun et al. 2018). For 16S rRNA amplicon sequencing (► Chap. 8), Q20 or higher is preferred.

Another important parameter is N_{50} which is obtained by listing all contigs from the longest to the shortest. The length is then accumulated from the longest to the shortest contig, and when half the genome size is reached, the length of that particular contig is read as N_{50} (◘ Fig. 2.5).

Comparison of different assemblies can be made with Quast (Gurevich et al. 2013). The "minimum information about a genome sequence" (MIGS) specification (Field et al.

◘ **Fig. 2.5** Illustration of the calculation of N_{50}. First, the contigs are arranged in decreasing order from longest to shortest. The length of sequence is accumulated, and when the accumulation reaches 50%, the specific contig is identified and its length recorded. The length of this contig is defined as N_{50}

Fig. 2.6 The original reads of DNA sequences from the sequencing machine are mapped to the contigs assembled from the reads (consensus in the figure) and to another related genome (de novo assembly) in the figure. Green reads were determined by forward sequencing primers and red reads by reverse sequencing primers. Reads in blue are paired read of both forward and reverse reads. (Output from CLC Genomics Workbench (► https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/))

2008) provides an exhaustive list of the information required for genomic sequences including demands to metadata. The reads can be compared to the assembled contigs by mapping (Fig. 2.6). This way, local coverage can be evaluated.

2.2.2.4 De Novo or with Reference

The assembly process can be made easier if an existing genome of a related strain is used in the process. It is used as a reference. This approach is only possible for well-known species where fully closed genomes are available.

2.3 Closing of Genomes

If repeats in the DNA sequences to be assembled are longer than the read length, the DNA sequences become harder to assemble. With prokaryotes, the problem is often ignored. If the researcher is interested in coding genes to predict the function of the proteins, genome closing is not critical since the majority of functions can be scored without full assembly. Also whole genomic multilocus sequence typing (wgsMLST) can be done on contigs which are not fully assembled to a full chromosome. The rRNA operons (5–7 kilo bases) and tRNA genes are major reasons for the lack of full assembly since they are longer or at the same size

Table 2.4 The FASTQ format

Format	Example
@ (title line)	@SEQ_ID
DNA sequence	GATTTGGGGT
+ shows end of sequence and start of quality	+
Quality lines	"*(+)%0.1–5

have most read lengths. Fully closed genomes are preferred as references in single-nucleotide polymorphism (SNP) finding further described in ▶ Chap. 11. To combine contigs, short paired reads can be combined with long paired reads. Long reads can be obtained by the PacBio method with a significant fraction of reads longer than 7 kb (Koren et al. 2013).

2.4 DNA Sequence Formats

The most simple and frequently used format to represent DNA sequences is the FASTA format named after the sequence search program FASTA (Pearson and Lipman 1988). It is a text format with the first line starting with ">" and some information about the sequence. On the following lines follow the DNA sequence written as the order of nucleotides. Files can include one or more sequences. FASTA format files often have the extension .fa, .fna, .fsa, or .fasta. However, the extension is only required for certain programs to recognize the format, and the files can be used without an extension or with .txt as extension if used for text editing. An extension to the FASTA format is the FASTQ, the format where the "Q" is short for quality (Table 2.4). Such files include both DNA sequences and quality scores, and they are suited for high-throughput sequencing (Cock et al. 2010). A quality score is included for each nucleotide in the DNA sequence. The quality score is in the form *phred*: QUAL format. Scores are based on ASCII codes 33–126 for *phred* qualities 0–93 (1 wrong base to $10^{-9.3}$ is extremely accurate).

2.5 Annotation

Genome annotation is the identification and labeling of all the relevant features of the genomic sequence. At first this includes the coordinates provided as nucleotide positions in where coding regions are predicted. It is mainly a prediction of coding genes; however, other structural genes such as rRNA are also identified.

First, the coordinates of candidate genes open reading frames (ORF) are predicted. ORFs are stretches of DNA that can translate to amino acids without stop codons. There are three reading frames of the forward (sense) strand of DNA and three reading frames on the reverse (antisense) (Appendix). Normally only one out of six will be without stop codons over hundreds of nucleotides, and such a reading frame is a candidate for a coding gene. Further, to predict the function of ORFs, the gene sequences are compared to different databases by a variant of BLAST. The databases include UniProt, RefSeq at NCBI, and

Table 2.5 Annotation programs

Program	Database	Principle	Reference
BlastKOALA	NCBI RefSeq, GenBank	Single genomes	▶ http://www.kegg.jp/blastkoala/ Kanehisa et al. (2016)
Ghost-KOALA	NCBI RefSeq, GenBank	Metagenomes	▶ http://www.kegg.jp/ghostkoala/
RAST	SEED, many databases	Subsystem	▶ http://rast.nmpdr.org/ Aziz et al. (2008), Overbeek et al. (2014)
MG-RAST		Subsystem	Glass et al. (2010)
Prokka[a]	RefSeq, UniProt, Pfam	Suite of existing programs	Seemann (2014)

[a]Runs on Unix/Linux

domain databases such as Pfam (▶ Chap. 3). For some ORFs, a functional prediction cannot be obtained, and they are designated as hypothetical proteins (no match in databases).

The most frequent way to annotate prokaryotic genomes is during the deposition of the sequence with NCBI. Here whole genomic sequences will be automatically annotated. To obtain an independent annotation, the server RAST (rapid annotation with subsystem technology) can be used (Aziz et al. 2008; Overbeek et al. 2014) (Table 2.5). In some cases the NCBI and RAST annotation will be different for the same ORF. It is related to different priorities for best hits in the databases. To investigate such conflicts further, the researcher needs to compare the specific sequence to different databases (▶ Chap. 3) by the use of BLAST or similar search tool (▶ Chap. 4). Prokka is a suite of software tools used for annotation (Seemann 2014) (Table 2.5).

2.6 Activities

2.6.1 Assembly with DTU Server

First, we need to get some DNA sequence reads to assemble. They can be downloaded from SRA at NCBIl. At NCBI, select SRA (▶ https://www.ncbi.nlm.nih.gov/sra/) and write the SRR number (SRR639898). You will see a list with the different sequence runs from the project. Use SRR639898 as a link. A new window opens with the experiment number (SRX145167). Click the upper panel Download bottom and select FASTA/FASTQ and write the experiment number from above in the window. Mark the run SRR639898 and to the right select FASTQ and download. You will now download a .gz format file and save it on your computer.

Open ▶ https://cge.cbs.dtu.dk/services/Assembler/.

Select Illumina **paired end reads** and upload the file just downloaded.

2.6.2 Annotation at RAST

You need to register to get an account ► http://rast.nmpdr.org/. The online program both performs ORF finding and annotation and returns GenBank formatted files as well as other outputs. The optimal function is with fully assembled genomes, but it can also work with unassembled contigs where you just include all contigs in one file. You can use the contigs assembled in ► Activity 2.6.1 above.

When logged on RAST, select **Your jobs | Upload new job** and select the file with the contigs in FASTA. Select genetic code "Bacteria" and "11" since the genome was from an enterococcus and *Enterococcus* is a Prokaryote.

Leave the next page with default settings. **Press submit.** Now you can follow the job at "Your jobs." It usually takes overnight to finish. The three main output files are the **GenBank** with the annotations and the **Amino-Acid FASTA** and the **Nucleic-Acid FASTA** with amino acid and nucleotide sequences of all predicted genes, respectively.

Take-Home Messages

- DNA sequencing is the determination of the order of nucleotides of parts or of whole chromosomes of organisms and virus.
- DNA sequencing can be done for a single gene or a whole genome or many genomes at a time such as in metagenomics.
- One of the most popular high throughput sequencing machines is the MiSeq from Illumina which is capable of doing whole genome sequencing of smaller genomes, transcriptomics and 16S rRNA metagenomics.
- Multiple samples can be sequenced by high throughput sequencing by multiplexing using unique combinations of specific barcodes and indexes.
- Real time, single molecule sequencing allows for sequencing of the native DNA, resulting in significantly longer read lengths compared to Illumina sequencing.
- Base calling is the first step in sequencing where the electronic signal generated in the sequencing machine is separated from random noise and converted to nucleotide information.
- The nucleotide information needs to be assembled to DNA sequences which resemble the original DNA sequence as best as possible.
- The most important quality parameter to consider in the final assembled DNA sequence (contig) is the coverage indicating how many reads that can be aligned with the assembled DNA at a given position of the DNA sequence.
- The 'minimum in formation about a genome sequence' (MIGS) specification provides an exhaustive list of the information required for genomic sequences including demands to metadata.
- Genome annotation is the identification and labelling of all the relevant features of the genomic sequence and it includes the coordinates provided as nucleotide positions where coding regions are predicted as well as structural genes such as rRNA.

References

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, Rooney AP, Yi H, Xu XW, De Meyer S, Trujillo ME. 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol. 68, 461–466.

Cock et al. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38, 1767–1771

Compeau PE, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 29:987–91.

Ewing B, Hillier L, Wend MC, & Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome research 8, 175–185.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G, Wipat A. 2008. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 26, 541–7.

Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F. 2010. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb Protoc.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 17:333–51.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29,1072–5.

Idury RM, Waterman MS. 1995. A new algorithm for DNA sequence assembly. J Comput Biol. 1995 Summer;2(2):291–306.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acid Res. 44(D1):D457–62.

Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM. 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biology 14: R101.

Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol. 50, 1355–61.

Madigan M, Bender KS, Buckley DH, Sattley WM, & Stahl D. 2019. Brock biology of Microorganisms. Pearson, Harlow UK.

Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, Prjibelski AD, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, Clingenpeel SR, Woyke T, McLean JS, Lasken R, Tesler G, Alekseyev MA, Pevzner PA. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol. 20, 714–37.

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 42(Database issue):D206–14.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 85, 2444–8.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 98, 9748–53.
Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74, 5463–7.
Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–9.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–9.

Further Reading

Loosdrecht, M. C. M. van, Nielsen, P. H., Lopez Vazquez, C. M. and Brdjanovic, D. 2016. Experimental methods in wastewater treatment. IWA publishing, London, UK

Databases and Protein Structures

Henrik Christensen and Lisbeth E. de Vries

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_3

What You Will Learn in This Chapter

Bioinformatics databases store biological information from life science research. It is an important topic for understanding of the majority of the other chapters in this book. In this chapter you will learn how to access the bioinformatics databases and extract information from them. You will be guided for best choices of databases and how to make sequences and related molecular material accessible to the scientific community by deposition of such material with the databases. You will be introduced to tools to visualize protein structures and to compare protein structures. Finally, a brief introduction to proteomics is provided.

3.1 Introduction to Bioinformatics Databases

Bioinformatics databases contain biological data from scientific experiments most important, DNA and protein sequences and protein structures. In addition to this core level, databases of published literature, computational analysis of primary data, and metadata are also important. Bioinformatics databases are usually affiliated with query facilities as well as data analysis tools. Bioinformatics databases are important for most other fields of bioinformatics and relate to most other chapters of the book (Fig. 3.1).

We expect the databases to represent high diversity of information as well as data formats that can be handled by most computers. The databases need to be reliable and updated and the documentation for the databases easily accessed. Documentation for the databases can be found in the journal, *Nucleic Acids Research* (▶ https://academic.oup.com/nar/issue/46/D1), where the first issue each year publishes papers dealing with new developments in the bioinformatics databases. We will refer to these papers and recommend the readers to refer to the papers when they include bioinformatics databases in their scientific publications.

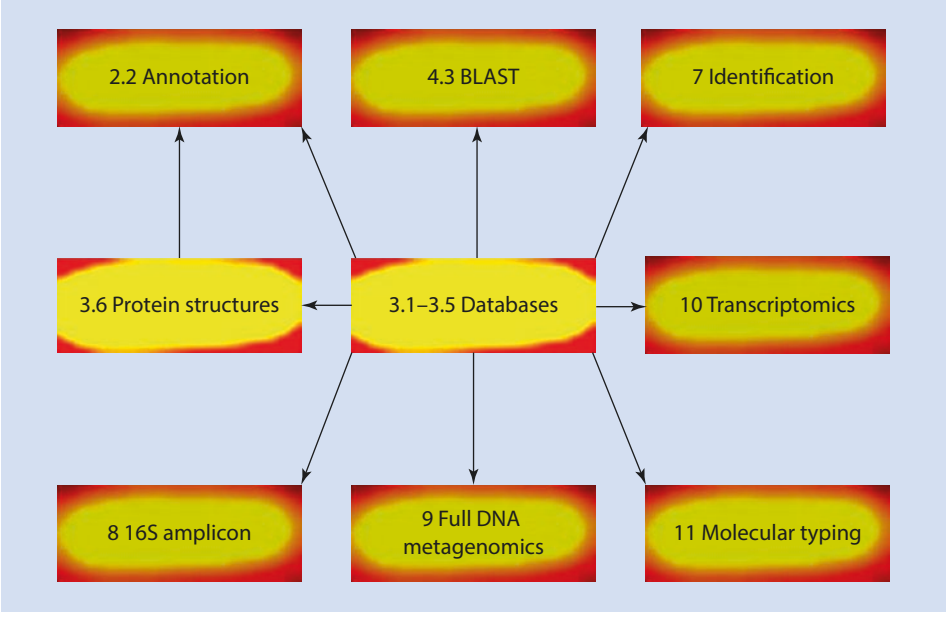


Fig. 3.1 Relation of this chapter to the other chapters in the book

Table 3.1 Major bioinformatics institutions

Abbreviation	Institution and activity	URL and reference
NCBI	National Center for Biotechnology Information Maintains the most important and frequently used bioinformatics databases such as GenBank and PubMed	► https://www.ncbi.nlm.nih.gov/ (NCBI 2018)
UniProt	Universal Protein Resource Maintains secondary protein databases, most importantly Swiss-Prot	► http://www.uniprot.org/ (UniProt 2018)
EMBL-EBI	European Bioinformatics Institute Bioinformatics databases, research, and training	► https://www.ebi.ac.uk/ (Cook et al. 2018)
NIG	National Institute of Genetics hosts DDBJ (DNA databank of Japan). DDBJ provides databases and analysis services from life science researches and advances science	► https://www.nig.ac.jp/nig/ (Kodama et al. 2018)

Some databases present parallel information, and each user has a special preference for a certain set of databases, and in this respect the weight and priority in this chapter reflect the author's personal choices and not any ranking of databases.

Bioinformatics databases are hosted by bioinformatics institutions who provide the resources to maintain and curate the databases (Table 3.1). The major bioinformatics databases are available on the Internet with free access from bioinformatics institutes. For example, GenBank is hosted with NCBI (NCBI 2018) and ENA with EBI-EMBL (Cook et al. 2018) (Tables 3.1 and 3.2). The bioinformatics institutions are providing the curation of the databases and maintaining the servers needed to host the large datasets and to make data accessible for the scientific society. NCBI is located and maintained from the USA, whereas EBI-EMBL is located in the UK and maintained mainly by European funding. DDBJ is located in Japan and maintained by NIG (Kodama et al. 2018).

The first bioinformatics database was established in 1965 (PIR) (Barker et al. 1993), and PIR later became associated with UniProt (Uniprot 2018). The majority of other databases were established during the 1980s (Table 3.2).

3.1.1 Data Formats Used with Bioinformatics Databases

From the information technical point of view, databases can be classified based on the database file structure, for example, as flat file, object-oriented, and relational databases. A flat file database is an ordered collection of similar files in some standard format. A flat file database is made useful by ordering the files and indexing them, which makes them searchable (Gibas and Jambeck 2001). These files are in ASCII text file format (► https://www.asciitable.com/). The ASCII format translates characters to numbers that can be read into computers. The ASCII format can be read both by humans and computers. Many popular sequence databases, including GenBank, are flat file databases.

For relational databases, information is stored in a collection of tables, which makes it possible to extract specific type of information across the database content. For example,

Table 3.2 Bioinformatics sequence and protein structure databases

Name	Type[a]	Content	Established and hosted at	URL
GenBank	P	DNA sequences[b]	Established in 1982 and hosted at NCBI since 1992	► https://www.ncbi.nlm.nih.gov/ (Benson et al. 2018)
ENA (European Nucleotide Archive)	P	DNA sequences[b]	Established in 1982 and hosted at EMBL-EBI since 1994	► https://www.ebi.ac.uk/ena
DDBJ (DNA Data Bank Japan)	P	DNA sequences[b]	Hosted at NIG since 1987	► http://www.ddbj.nig.ac.jp/ (Kodama et al. 2018)
SRA (Sequence Read Archive)	P	DNA sequences	NCBI	► https://www.ncbi.nlm.nih.gov/ (Haft et al. 2018)
RefSeq	S	DNA sequences	NCBI	► https://www.ncbi.nlm.nih.gov/ (NCBI 2018)
Swiss-Prot	S	Protein sequences	Established in 1986 and since 2003 included with UniProt	► http://www.uniprot.org/ (UniProt 2018)
PDB (Protein Data Bank)	PS	3D structures of proteins	Established in 1971 and hosted at Research Collaboratory for Structural Bioinformatics (RCSB)	► https://www.rcsb.org/ (Rose et al. 2017)

[a]*P*, primary; *S*, secondary databases
[b]and their protein translations

from the PBD database, you can extract only information about the secondary structure for a specific type of proteins (Gibas and Jambeck 2001).

Object-oriented databases are more complex. They handle data as object (instead of tables), which enables the storage of various information types in different formats, from simple text formats to images and videos (Gibas and Jambeck 2001).

3.2 Organization of Databases and Bioinformatics Institutions

The International Nucleotide Sequence Databases (INSD), DDBJ, EMBL, and GenBank in short DDBJ/EMBL/GenBank are organized in INSDC (International Nucleotide Sequence Database Collaboration) (► http://www.insdc.org.). The three databases collaborate to enable access to DNA data in standardized formats for the scientific community worldwide (Karsch-Mizrachi et al. 2018). The online content of the databases is exchanged on a daily basis. To do that the officially supported XML format of INSDC, the INSDSeq, is used (see ► Sect. 3.3.8).

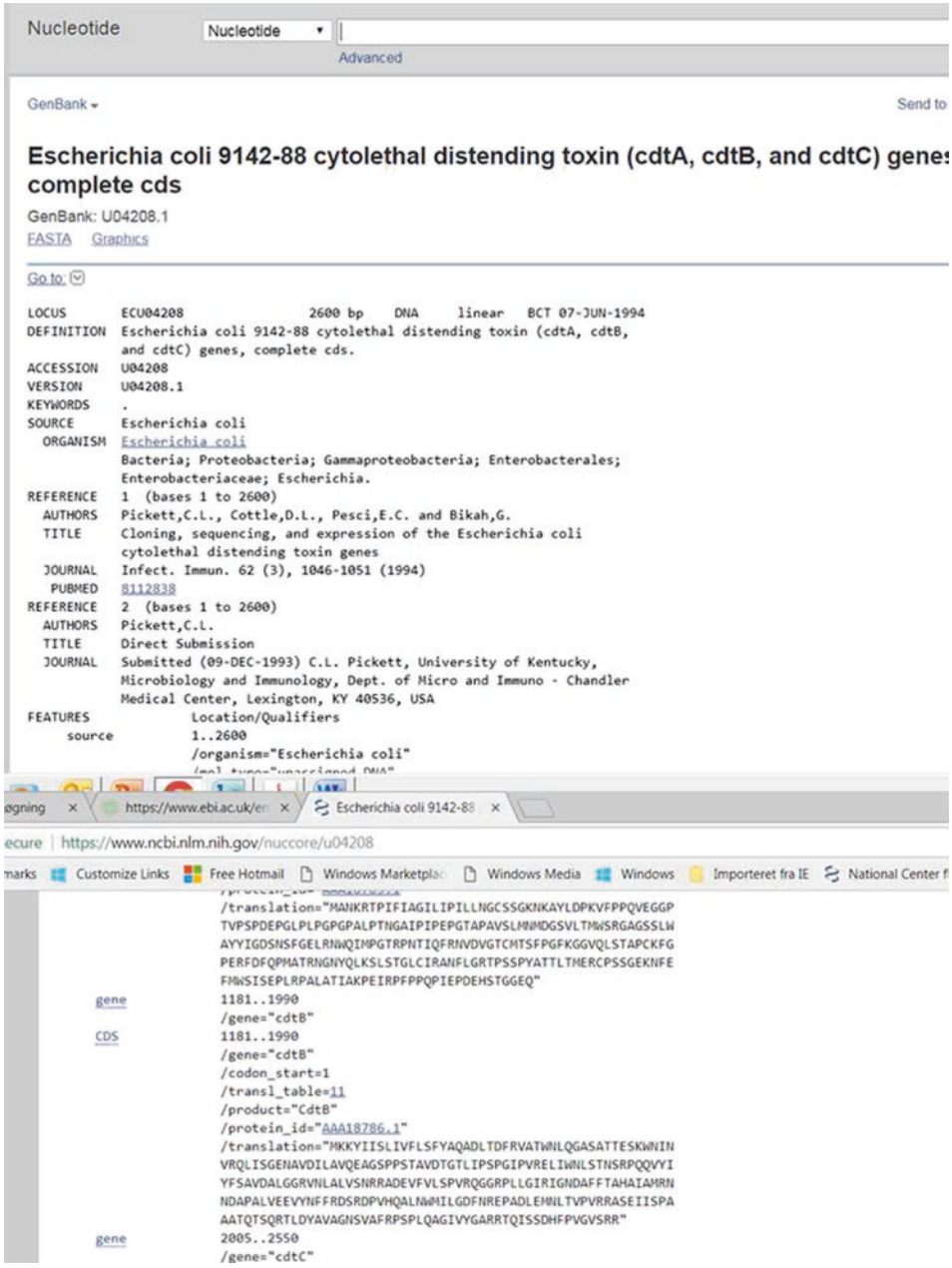


Fig. 3.2 The GenBank format. Note only the start of the entry and a part of the protein section are shown. Readers can easily look at the full record at ► https://www.ncbi.nlm.nih.gov/nuccore/u04208

The nucleotide databases DDBJ, EMBL, and GenBank are automatically translated to the protein level if the DNA sequences are coding. These protein translations are available along with the DNA sequences from the databases (Fig. 3.2).

3.3 Major Bioinformatics Databases

3.3.1 GenBank

The best known bioinformatics database is GenBank (Fig. 3.2). This is a primary database for DNA sequences. Primary DNA sequence databases are sometimes called "nucleotide" databases and even shortened as "nt." GenBank is distributed from NCBI as flat file and ASN.1 (Abstract Syntax Notation One) file formats (► https://www.ncbi.nlm.nih.gov/genbank/) (see ► Sect. 3.3.8).

From GenBank all translations to protein are in the GenPept database. The well-recognized GenBank flat file format is shown in Fig. 3.2. Each GenBank record contains three parts: the first part includes information about the record (e.g., type of molecule, submission date, the unique accession number, keywords, source, and references) and the second part contains sequence annotations such as biological features, e.g., genes, CDS (coding sequences), base counts, and origin. The main characteristics of this format are that it is easy to read both for humans and computers and information fields are easy to understand (Fig. 3.2).

GenBank can be searched by keyword including accession number via text-based query directly at the root of NCBI (► https://www.ncbi.nlm.nih.gov/) or at the Nucleotide section (GenBank) where annotations associated with a sequence entry are searched (► https://www.ncbi.nlm.nih.gov/nucleotide/).

Sequence similarity search can also be carried out via the algorithm BLAST ► https://blast.ncbi.nlm.nih.gov/Blast.cgi, where a query sequence (DNA or protein) is compared to the sequence database (see ► Chap. 4 for more details). Data can be retrieved from GenBank in several file formats including the GenBank file record and FASTA file format (see also ► Activities 3.8.1 and 3.8.2).

3.3.2 The European Nucleotide Archive (ENA)

EMBL was the original primary database for DNA sequences hosted with EBI-EMBL. Now EMBL is included with ENA (Fig. 3.3). The translations of DNA sequences in EMBL to proteins were called TrEMBL (Cook et al. 2018). TrEMBL was developed specifically for comparison to the Swiss-Prot database (see below). The original EMBL format was using two-letter codes for all information fields (Fig. 3.4) which was harder to read than the GenBank format. The original two-letter format is sometimes used. However, the information fields in ENA are now also self-explanatory like GenBank (Silvester et al. 2018).

3.3.3 Swiss-Prot and UniProt

The aim of Swiss-Prot was to include all information from one protein from one species in one entry. Swiss-Prot was included in UniProt in 2002 (in full UniProt Knowledge database) (UniProt 2018), but the context of the Swiss-Prot database has survived.

PIR (Protein Information Resource) was established in 1965 by Margaret O. Dayhoff. The first IT version became available in 1972. The database has been included with UniProt since 2003. PIR is currently focusing on protein information comparisons (► http://pir.georgetown.edu/).

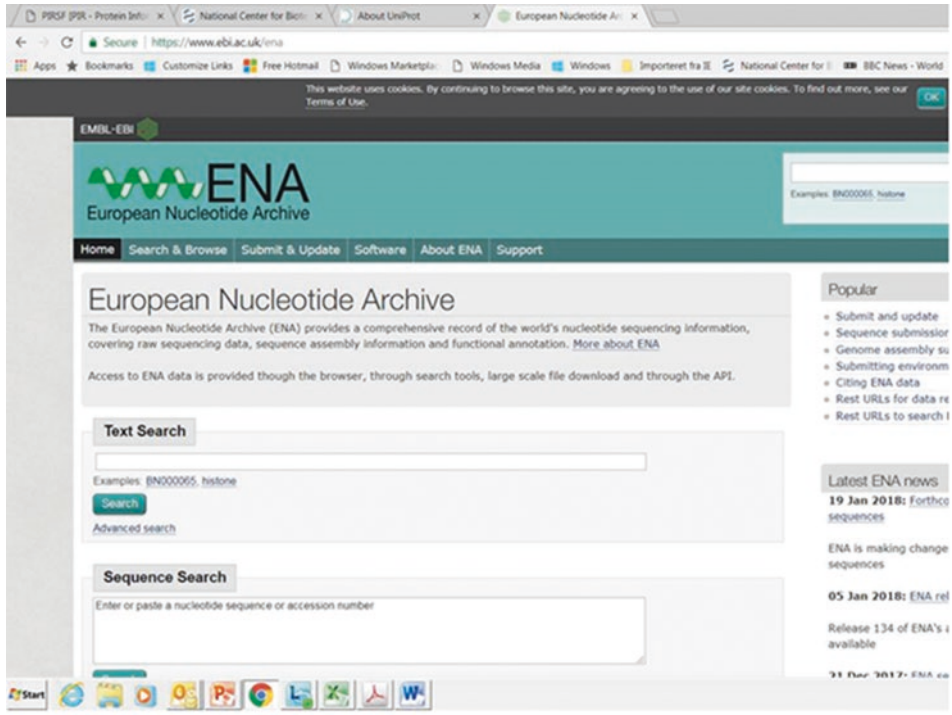


Fig. 3.3 The homepage of the European Nucleotide Archive (ENA) homepage. Text search by keyword including accession number can be carried out

Really well-known and a well-annotated proteins are referred to as "Swiss-Prot" in the UniProt database, and they are also recognized as "reviewer" and labeled with a golden star in UniProt. Sometimes the Swiss-Prot section is labeled "sp" for short, and this is used at NCBI where proteins from Swiss-Prot also can be looked up. Whenever possible one should look for information about proteins in the Swiss-Prot section of UniProt (Fig. 3.5).

The information about a protein in UniProt can be a bit overwhelming at first sight. This is because *all* information is actually here gathered about the protein. To guide readers, the information fields are shown in Table 3.3.

3.3.4 Genomics Databases

NCBI registers genomic sequences from single isolates according to biosample and bioproject in the format SAMN00000000 and PRJNA000000, respectively. A genome which is not fully assembled will get an accession number in the form AAAA00000000. For fully closed genomic sequences, a GenBank format DNA sequence accession number will be provided in the form AB123456 (see also ▶ Activity 3.8.4.2).

```
ID 004208; SV 1; linear; unassiqned DNA; STD; PRO; 2600 BP.
AC 004208;
DT 31-DEC-1993 (Rel. 38, Created)
DT 04-MAR-2000 (Rel. 63, Lase updated, Version 4)
DE Escherichia coli 9142-88 cytolethal distending toxin (cdtA, cdtB, and cdtC)
DE qenes, complete cds.
OS Escherichia coli
OC Bacteria; Proteobaceeria; Gammaproteobaceeria; Enterobacterales;
oc Enterobacteriaceae; Escherichia.
RN [1] RP 1-2600
RX PUBMED; 8112838.
RA Pickett C.L., Cottle D.L., Pesci E.C., Bikah G.;
RT "Cloning, sequencing, and expression of the Escherichia coli cytolethal
RT distendinq toxin qenes"; RL Infect Immun 62(3) :10 46-1051(1994).
RN [2] RP 1-2600 RA Pickett C.L.;
RT ;
RL Submitted (09-DEC-1993) to the INSDC. RL C.L. Pickett, University of
Kentucky, Microbiology and Immunoloqy, Dept. of RL Micro and Immune - Chandler
Medical Center, Lexington,
KY 40536, USA
FH Key Location/Qualifiers
FH FT source: 1..2 600
FT /organism="Escherichia coli"
FT /strain• "9142-88" FT /mol_eype•"unassigned DNA"
FT /db_xref="taxon:562"
FT CDS 408..1184
FT /codon start=1
FT /trans1_ table=11
FT /gene="cdtA"
FT /product="CdtA"
FT /db xref="UniProtKB/Swiss-Prot:Q46668"
FT /protein_id="AAA18785 .1"
FT /translation="MANKRTPIFIAGILIPILLNGCSSGKNKAYLDPKVFPPQVEGGPT
FT VPSPDEPGLPLPGPGPALPTNGAIPIPEPGTAPAVSLMNMDGSVLTMWSRGAGSSLWAY
FT YIGDSNSFGELRNWQIMPGTRPNTIQFRNVDVGTCMISFPGFKGGVQLSTAPCKFGPER
FT FDFQPMATRNGNYQLKSLSTGLCIRANFLGRTPSSPYATTLTMERCPSSGEKNFEFMWS
FT ISEPLRPALATIAKPEIRPFPPQPIEPDEHSTGGEQ" FT CDS 1181..1990
FT I codon_start=1
FT /transl_table=ll FT /gene="cdtB" FT / product="CdtB"
FT /protein_id="AAA18786 .1"
FT /translation="MKKYIISLIVFLSFYAQADLTDFRVATWNLQGASATTESKWNINV
FT RQLISGENAVDILAVQEAGSPPSTAVDTGTLIPSPGIPVRELIWNLSTNSRPQQVYIYF
FT SAVDALGGRVNLALVSNRRADEVFVLSPVRQGGRPLLGIRIGNDAFITAHAIAMRNNDA
FT PALVEEVYNFFRDSRDPVHQALNWMILGDFNREPADLEMNLTVPVRRASEIISPAAATQ
FT TSQRTLDYAVAGNSVAFRPSPLQAGIVYGARRTQISSDHFPVGVSRR"
FT CDS 2005..2550
FT I codon_start
IT /transl_table=11
FT /gene-"cdeC" FT/produce-"CdtC"
IT /protein id="AAA18787.1"
FT /translation="MKKLAIVFTMLLIAGCSSSQDSANNQIDELGKENNSLFTFRNIQS
FT GLMIHNGLHQHGRETIGWEIVPVKTPEEALVTDQSGWIMIRTPNTDQCLGTPDGRNLLK
FT MTCNSTAKKILFSLIPSTTGAVQIKSVLSGLCFLDSKNSGLSFETGKCIADFKKPFEVV
FT PQSHLWMLNPLNTESPII" XX SQ Sequence 2600 BP; 834 A; 555 C; 508 G; 703 T;
0 other;
attaacaaat tacaacaaaq atcacattaa ataaaaattq acaattqacc tqtaqttcat 60
caeeegeaaa eecaegaeee aeatcaatca cgceeegege ecgqageaag ceeaeaaaet 120
```

Fig. 3.4 EMBL format for acc. U04208 (► https://www.ebi.ac.uk/ena/data/view/U04208&display=text). To simplify, some of the reference, cross-reference, and sequence sections have been omitted on the figure

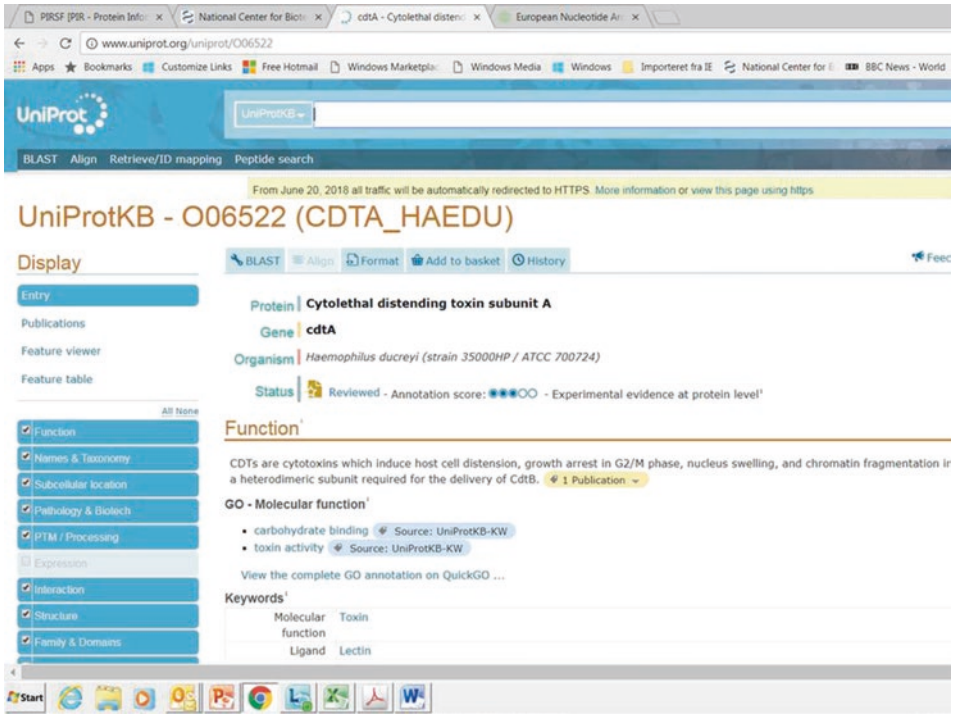


Fig. 3.5 An entry in the UniProt database showing that the protein is in Swiss-Prot (reviewed, note the golden star). The full entry can easily be viewed from ▶ http://www.uniprot.org/uniprot/O06522 (UniProt 2018)

3.3.5 Raw Sequence Read Datasets

SRA (Sequence Read Archive) (▶ https://www.ncbi.nlm.nih.gov/) (Haft et al. 2018) (Table 3.2) is a primary database for raw sequence reads hosted with NCBI. Here whole raw datasets from Illumina sequencing of whole genomes (▶ Chap. 2) and 16S rRNA metagenomes (▶ Chap. 8) and full DNA metagenomes (▶ Chap. 9) can be stored and downloaded. A similar facility is provided by ENA (European Nucleotide Archive) (Table 3.2).

3.3.6 Other Databases

The following databases are all more specialized (Table 3.4). CAZyme is a database of carbohydrate active enzymes (André et al. 2014). This database separates enzymes in GH (glycoside hydrolase) families. It is possible to annotate protein sequences of whole genomes according to CAZy by the use of dbCAN (▶ http://csbl.bmb.uga.edu/dbCAN/).

KEGG (Kyoto Encyclopedia of Genes and Genomes) has the main objective to establish links from collective sets of genes in the genome to high-level function of the cell and organism (Kanehisa et al. 2016). Annotation can be performed in KEGG (▶ Chap. 2). For a long time KEGG GENES database was created from NCBI's RefSeq database. From 2014 on, KEGG GENES also includes genomes from GenBank. KEGG has a section devoted to

Table 3.3 Major information fields used in a UniProt protein entry. The CdtB protein has been used as example (accession number Q46669 from UniProt). Note that you can make a custom arrangement of the information fields at UniProt. For instance, if your main interest is *structure*, this field can be shown on the top

Information field	Content	Example
1. Function	Short summary of biological function	"CdtB exhibits a DNA-nicking endonuclease activity"
2. Names and taxonomy	Protein names, gene names, organism names	*Escherichia coli*
3. Subcellular location	Localized to the membrane or cytosol	"Secreted"
4. Pathology and biotech	Effect of mutagenesis on function	Mutagenesis H154A will not cause cell cycle arrest in HeLa cells
5. PTM/ processing	How the molecule is processed	Signal peptide at position 1–18
6. Structure	Graphical representation of the secondary structure. Links to 3D models. Note, only models with PDB numbers have been experimentally determined	2F1N link
7. Sequence	The protein sequence	MKKYIISLIV FLSF
8. Family and domains	Links to family and domain databases	"View protein in Pfam"
9. Interaction	Other proteins that the protein is associated with in vivo	Heterotrimer of three subunits, CdtA, CdtB, and CdtC
10. Entry information	Information about the background information for the record	O32586 a similar protein

Table 3.4 Specialized databases

Name	Content and use	URL
KEGG	Metabolic pathway database	► http://www.genome.jp/kegg/ (Kanehisa et al. 2016)
CAZyme	Database of carbohydrate active enzymes	(► www.cazy.org) (André et al. 2014)
Essential genes	Genes which are indispensable for the survival of organisms	► http://www.essentialgene.org/ (Gao et al. 2015)
Silva	rRNA gene sequence database all three domains of life	► https://www.arb-silva.de/ (Yilmaz et al. 2014)
RDP	16S rRNA gene sequence database of prokaryotes and fungi	(Cole et al. 2014)
GreenGenes	16S rRNA gene database and quality control	(DeSantis et al. 2006; McDonald et al. 2012)

antimicrobial resistance. The benefit of KEGG is that it focuses on modules and pathways directed against functionality, and the drawback is that it excludes poorly annotated functions. The 16S rRNA gene sequence databases are further used in ▶ Chap. 8.

3.3.7 Primary and Secondary Bioinformatics Databases

From the biological point of view, the most important categorization of databases is into primary and secondary databases, which refers to the type and source of stored data. Primary databases, such as GenBank and ENA, are also called archives or repositories, and they take information directly from the individual researcher, and data are owned by the submitter with privileges to change data. The benefit is fast publication of new entries in the databases and high diversity. The drawback of this system is that information occasionally can be wrong, for instance, if the researcher should accidentally have used the wrong coding table (▶ Chap. 2) or wrong names for organisms (▶ Chap. 7) or wrong information fields.

Users should therefore always be aware of errors in the databases. It is recommended to contact the authors listed with the specific entry of database linked to the dataset if errors are suspected. Such errors will be passed on to all other databases that are linked to the dataset if they are not corrected. Another problem with the primary databases is the tendency for redundancy (repeating information) since two scientists could in principle submit the same sequence for the same gene for the same organism independently of each other. When searching primary databases, users should always be critical of the information, especially if the data is not associated with a published reference.

The secondary databases (e.g., Swiss-Prot and PDB) are curated, and they perform a quality control and sorting of information before the information is made accessible to the public. These databases have better chances of reducing redundancy. They can also bypass the submitters of entries in the primary databases which are no longer updated. This is important in case the submitter changes affiliation or no longer is active in science.

A variant of the secondary database is the third party annotation (TPA). This is a secondary database mainly of nucleotide sequences. To avoid shortcomings with primary data and bypass submitters' privileges, the database has released the submitters' burdens of updating data. The best-known example is RefSeq at NCBI which is used a lot with genomic sequences, providing a backbone to a range of DNA metagenomics studies (▶ Chap. 9).

3.3.8 Data Formats in Bioinformatics Databases

The most simple format is the FASTA format introduced in ▶ Chap. 2 (▶ Sect. 2.4). This format is not very suitable for arranging information in the bioinformatics databases since the information next to the sequence is not structured and can only be very brief. Therefore more special data formats are used.

ASN.1 (Abstract Syntax Notation One) is a format suitable for transferring data between different computers and computer system. If you select this format for a sequence in GenBank (try the link: ▶ https://www.ncbi.nlm.nih.gov/nuccore/U04208.1?report=asn1&log$=seqview), it can be read but will not give much meaning to the human reader but to the computer.

A very precise way of representing database information is the XML format. XML (Extensible Markup Language) is derived from Standard Generalized Markup Language (SGML). All information of a document is annotated in the format <tag>text</tag>. The INSDC variant of XML is used. For instance, the sequence in a GenBank file in INSDSq format is between tags <Seq-data> and <\Seq-data>, and the accession number in a GenBank file in INSDSq format is between tags <Seq-id> and <\Seq-id>.

3.4 Accession Numbers

All sequences in the databases are recognized by accession numbers. They are the unique identifiers for all information. Already the format of an accession number tells about its origin (Table 3.5). Nucleotide sequences in primary DNA database like GenBank have the format of two letters and six numbers. Their translated protein sequences have the format of three letters and five numbers. All GenBank entries had a GI number (GenBank unique number for each sequence 1234567); however, this system was stopped in September 2016, and now only older sequences in GenBank can be recognized by this number. New WP numbers were introduced from 2013 for proteins. Non-redundant protein sequences will only be given one WP number (Clark et al. 2016). To get information about proteins included in WP numbers, use the number as link, and see *identical proteins* where all the accession numbers included with the WP number are listed.

Table 3.5 Accession numbers

Database	Section	Accession number format	Example
International Nucleotide Sequence Database Collaboration (GenBank, EMBL, DDBJ)	GenBank	Common to DNA X12345 (until 1999) (1 + 5), AB123456 (1999 -) (2 + 6)	U04208
	Protein	P12345 (until 1999) (1 + 5), ABC12345 (1999 -) (3 + 5)	O06522
	RefSeq	Accession no. 2 + 6 with underscore NT_123456 genomic DNA NM_123456 mRNAs NP_123456 protein	
	Swiss-Prot	Primary accession number on the form P12345 but also six-digit combinations of numbers and letters: NNNN_YYYYY The Ns represent the protein name and the Ys the organism	O06522 CDTA_HAEDU
PDB		1N1N (capital letters and numbers without fixed order)	2F1N
NCBI	Genomes	Incomplete NNNN12345678 Complete AB123456 (Or A12345 old number)	LXWV01000000 L42024

3.5 Protein Structure Databases and Predictions

This section provides an introduction to databases of protein structures and how structures can be downloaded and visualized from the databases. Prediction of structures based on a query protein sequences will be described.

One of the beautiful synergistic achievements in bioinformatics is that the function of a protein can be predicted by a rather low identity of proteins over rather short length of comparison and rather low similarity to protein structures. As a rule of thumb, the function of a protein can be predicted if it is at least 40% similar in its primary structure to known proteins in the database. In the range of 20–40% similarity, only folds can be predicted but not function. Homology modeling based on 3D structure performed with Swiss-Prot or similar (■ Table 3.6) can predict 3D structures for proteins if they are at least 25% identical in a pairwise alignment of at least 100 amino acid (Petsko and Ringe 2004).

With respect to the prokaryotes, we are interested in predicting if a protein is secreted to the extracellular space, if it is part of the cell wall, or if it is located in the cytoplasm of the prokaryote. A special group of very important proteins form pores in the prokaryotic cell and occupy important positions in the periplasmic space of bacteria with a Gram-negative cell wall structure. Both function and location of uncharacterized proteins can be predicted bioinformatically by comparison to specialized protein structure databases.

■ **Table 3.6** Prediction and comparison tools for protein structures

Tool	Function	URL
PRED	Predicting and discriminating beta-barrel outer membrane proteins	► http://bioinformatics.biol.uoa.gr/PRED-TMBB/ (Bagos et al. 2004)
Dali	Comparison of structures	► http://ekhidna2.biocenter.helsinki.fi/dali (Holm and Laakso 2016)
Phyre2	For comparison of models. Visualization in CLC Main Workbench using superimpose function	► http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index (Kelley et al. 2015)
SignalP	Determination of signal peptide cleavage sites	► http://www.cbs.dtu.dk/services/SignalP/ (Petersen et al. 2011)
PSORT	Protein localization sites	► https://psort.hgc.jp/(Nakai and Horton 1999)
PHOBIUS	Transmembrane topology and signal peptide predictor	► http://phobius.sbc.su.se/ (Käll et al. 2004)
OMPdb	Beta-barrel membrane proteins located in the outer membrane of bacteria with a Gram-negative cell wall structure	► http://aias.biol.uoa.gr/OMPdb/ (Tsirigos et al. 2011)
SWISS-MODEL	Homology modeling of protein structures	► https://swissmodel.expasy.org/ (Waterhouse et al. 2018)
PyMOL	The ultimate tool for protein modeling	► https://pymol.org/2/

3.5.1 Primary and Secondary Structures

The primary structure of a protein is simply the amino acid sequence. The secondary structure can mainly be separated in α-helices and β-sheets (Fig. 3.6); however, also the "loop" and "turn" regions linking α-helices and β-sheets are important. Wool contains α-helices and silk β-sheets. In the prokaryotic world, proteins rich in α-helices may be transmembrane; however, the beta-barrel structure predominated by β-sheets may also be transmembrane. An example of a protein with β-sheets is the cytolethal distending toxin (Cdt) (Fig. 3.6). It is a tripartite complex (A, B, C) that is required for the CDT activity (Pickett and Whitehouse 1999). The holoprotein (consisting of proteins A, B, and C) induces G2/M cell cycle arrest, chromatin fragmentation, cell distention, and nucleus enlargement. CDTA and CDTC probably form a heterodimeric subunit required for the delivery of CDTB. CDTB exhibits a DNA-nicking endonuclease activity and very probably causes DNA damage in intoxicated cells.

The M-protein is used as an example of a protein with mainly α-helix secondary structure (Ghosh 2018). The α-helices are here exposed on the surface of the streptococci, and the variable N-terminal may interact with the immune system of the host. Two α-helices are held wound together to coiled coil holoproteins. A coil-coil structure is formed as a super twisted helix. The two α-helices will have a periodicity of seven amino acids with a certain combination of hydrophobic amino acids allowing two adjacent hydrophobic amino acids to bury their hydrophobic nature from the surrounding solvent.

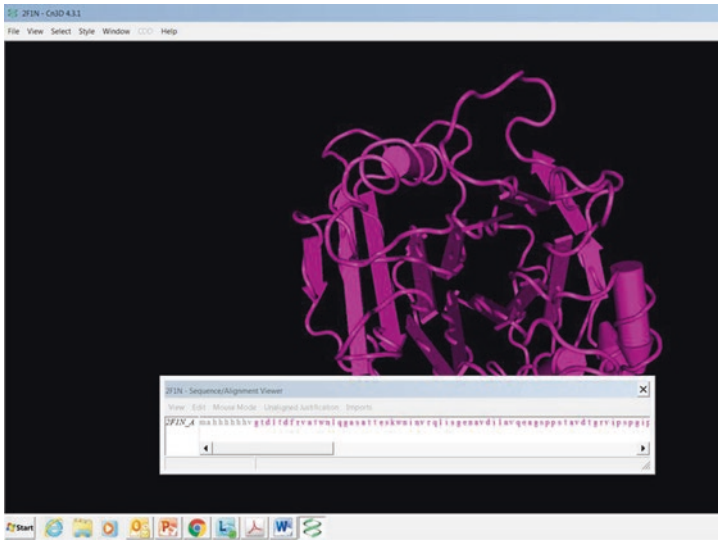


Fig. 3.6 The 3D structure of the CdtB protein (acc. no. 2F1N) visualized with Cn3D viewer (► Activity 3.8.3). The flat arrows illustrate regions with β-sheets and the "pencil" regions with α-helices. (STRUCTURE [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2018 04 17]. Available from: ► https://www.ncbi.nlm.nih.gov/structure/)

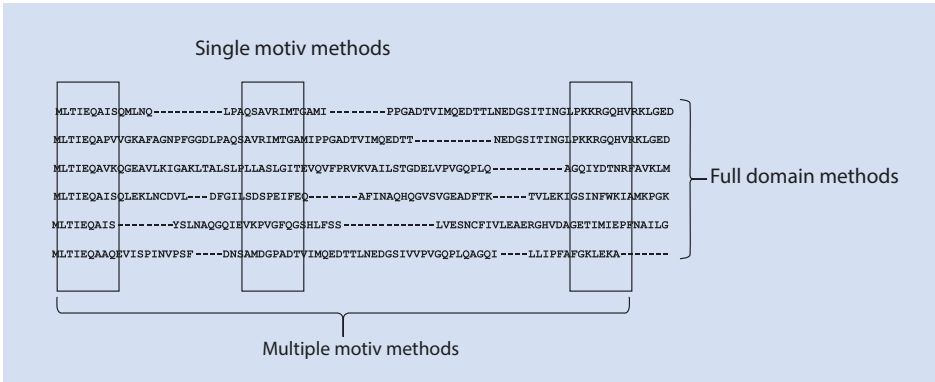


Fig. 3.7 Databases and prediction tool of protein motifs, domains, and families. (Figure redrawn from Higgs and Attwood (2005))

3.5.2 Domain Prediction and Databases

Domains are compact units of proteins that may behave independently and be associated with certain functions. Motifs are conserved regions of proteins which may be part of domains (Fig. 3.7). On the figure, the three types of predictors are illustrated.

3.5.2.1 Single Motif

PROSITE (▶ https://prosite.expasy.org/) is based on the so-called regular expressions to define biologically significant protein patterns and profiles. A regular expression is a sequence of characters that define a *search pattern*.

A regular expression is, for instance:

Y-x-[NQH]-K-[DE]-[IVA]-F-[LM]-R-[ED] which is the heat shock hsp90 proteins' family signature. The code is the single symbol code for amino acids, and positions where variation is allowed are labeled with "x," and positions where only a subset of amino acids are possible are labeled within brackets. In the PROSITE database, an entry will look like this (note the EMBL coding; see Fig. 3.4).

```
ID CUTINASE_1; PATTERN.
AC PS00155;
DT APR-1990 (CREATED); NOV-1997 (DATA UPDATE); MAR-2005 (INFO UPDATE).
DE Cutinase, serine active site.
PA P-x-[STA]-x-[LIV]-[IVT]-x-[GS]-G-Y-S-[QL]-G.
```

Epitope prediction is an important task for the prediction of interactions of proteins with the immune system. The paper by Soria-Guerra et al. (2015) compares different tools. Surface-exposed epitopes are often associated with loops and turns located between α-helices and β-sheets.

3.5.2.2 Multiple Motifs

BLOCKS (Henikoff and Henikoff 1992) is based on multiple alignments of conserved regions. PRINTS is parallel to BLOCKS based on OWL (Attwood et al. 1997).

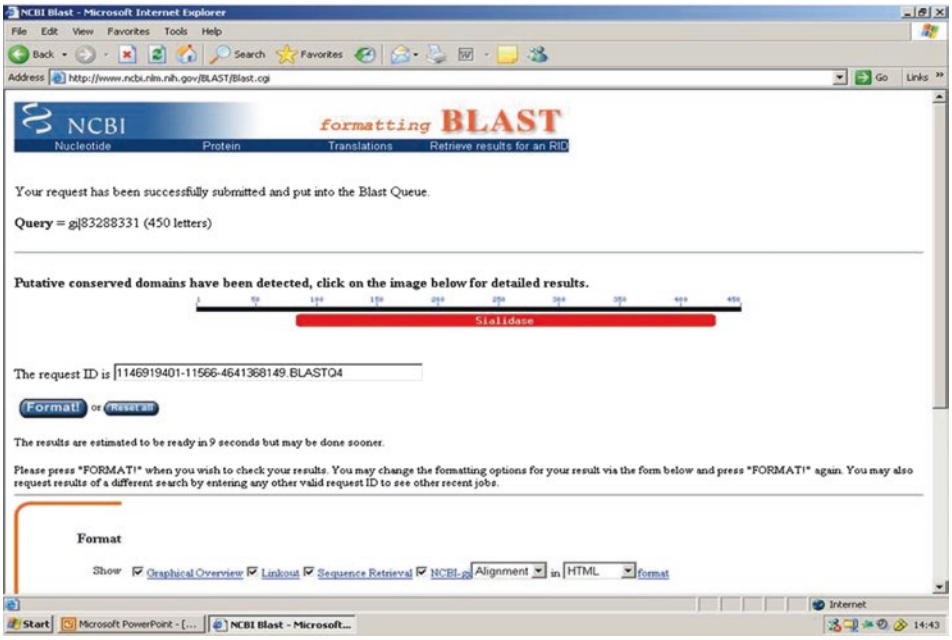


Fig. 3.8 CDD domain prediction invoked in BLASTp. The search resulted in a sialidase domain being predicted. (BLAST [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004. Available from: ► https://www.ncbi.nlm.nih.gov/Blast.cgi)

3.5.2.3 Full Domain

Pfam is primarily based on UniProt reference proteomes (Finn et al. 2016). A representative subset of matching sequences are aligned to make a seed alignment. It is used to construct a profile hidden Markov model (HMM) using the HMMER software (► http://hmmer.org).

SMART (simple modular architecture research tool) ► http://smart.embl-heidelberg.de/ is for identification and annotation of protein domains based on manually curated models. Prediction is also here based on hidden Markov modeling (HMM) (Letunic and Bork 2018).

ProDOM (automatic created blocks, France) (Sonnhammer and Kahn 1994) (► http://prodom.prabi.fr/prodom/current/html/home.php) is a comprehensive set of protein domain families automatically generated from the UniProt Knowledge database.

3.5.2.4 Mixing Different Methods

InterPro (Finn et al. 2017) is linking information in PRINTS, PROSITE, ProDOM, CDD, Pfam, and seven additional domain databases. InterProScan is the underlying software that allows users to search with protein and DNA sequences against InterPro's predictive models.

CDD (Conserved Domain Database) is linking information from Pfam and SMART (Fig. 3.8). CDD is automatically invoked when a protein BLAST search is activated at NCBI (► Chap. 4).

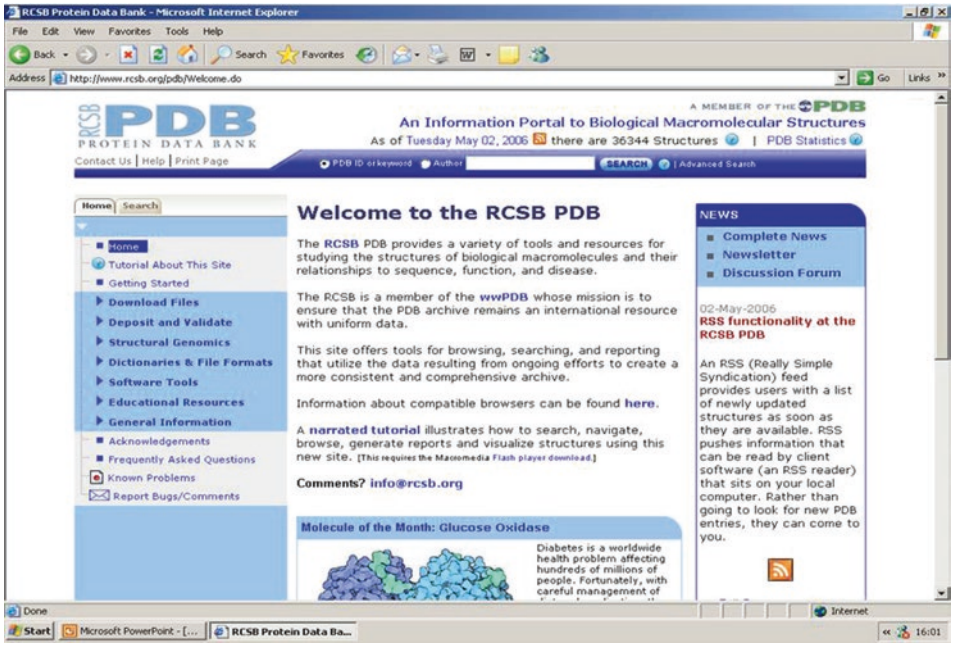


Fig. 3.9 The portal of PDB (Protein Data Bank) hosted by RCSB (Research Collaboratory for Structural Bioinformatics). PDB is a worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids (► https://www.rcsb.org/)

3.5.3 Protein 3D Structure

The 3D structure of a protein is representing the full folding pattern of a protein. Three-dimensional structures are needed to analyze the location of active sites where amino acid substitutions will have an effect on the function of the protein. The 3D structure is also needed to investigate interactions between proteins and other molecules.

The 3D structure of a protein is an exact measurement based on X-ray crystallography or NMR spectroscopy. X-ray crystallography can only be performed on crystals of proteins. For proteins to form crystals, small ions are sometimes needed (ligands). 3D structures for proteins that cannot be crystallized need to be determined by NMR. The database for the 3D structures is PDB (Table 3.2) (Fig. 3.9).

An example of a 3D structure has been given for a protein with predominant β-sheets (CDTB) (Fig. 3.8), and the M-protein from *Staphylococcus aureus* has been used to represent a structure predominantly with α-helices.

The resolution of the 3D structure is very important. High resolution is best. With 1 Å resolution, the atoms are visible; however, 30 times more date are needed compared to 3 Å resolution. With 3 Å resolution, folds can be seen, and with 2 Å, side chains are resolved.

Comparison of 3D structures can be done in Phyre2 or SWISS-MODEL (Table 3.6). The folding pattern is theoretically determined by the phi (φ) and psi (ψ) angles. The peptide bound between amino acids in protein chains is planar (180°) (rarely 0°), and the φ (N-C) and ψ (C-C) angles determine the folding. The deviation between expected and observed folding can be visualized in a Ramachandran plot (Fig. 3.10). The upper left field represents φ and ψ angles of β-sheets, whereas the middle left section represents α-helices. The variants observed in the right quartets may be the amino acid glycine which

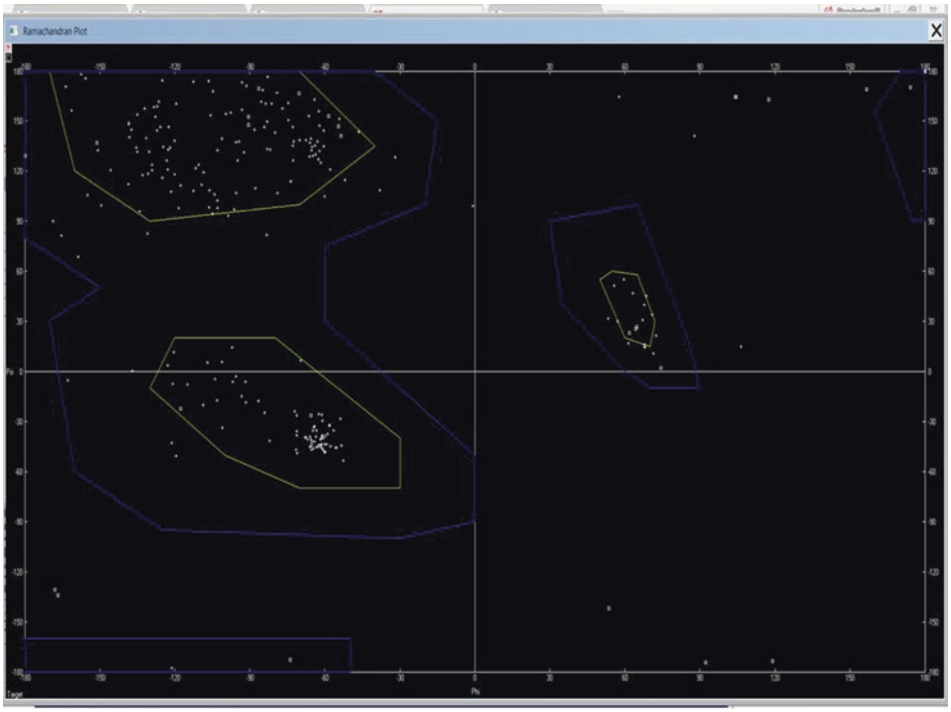


Fig. 3.10 Ramachandran plot is a comparison between observed and expected phi and psi angles. (The plot was obtained from SPDB (► Activity 3.8.5) (► https://spdbv.vital-it.ch/))

is not associated with any particular folding pattern. Amino acids at other positions in the plot represent observed variants not fitting with the theoretical expectations.

3.6 Overview of Proteomics Databases and Servers

Proteomics is dealing with the prediction of proteins based on measurements of mass-to-charge ratios. Proteins need to be degraded to peptides before they can be analyzed by mass spectrometry (MS). Trypsin is the preferred enzyme to degrade proteins to peptides resulting in molecular weights of 300–1500 daltons. Liquid chromatography (LS) is used to fractionate the protein extract for the analysis. The double MS (MS/MS) is needed to obtain sufficient resolution for precise determination of mass-to-charge ratios (m/z). Disulfide bridges are normally ignored in proteomics. The majority of proteins in cells are conserved between types and tissues. Only a minor fraction of proteins varies and changes with disease. The membrane-bound proteins are the most difficult to identify since they cannot be extracted free from their matrix (the membrane). The drawbacks with proteomics are that all proteins cannot be identified by proteomics. One group of proteins cannot be degraded to peptides. Another drawback is that the state of the art requires highly advanced hardware and bioinformatics that can acquire full images of samples and full storage of the image that can afterward be analyzed in different ways. For these reasons proteomics is even more complicated than "genomics." The most simple prediction of proteins is done with Mascot (Fig. 3.11). In this program the m/z coordinates from an analysis are held up against the reviewed part of UniProt (► Sect. 3.3.3). MaxQuant is a free program used for metaproteomics (Fig. 3.12).

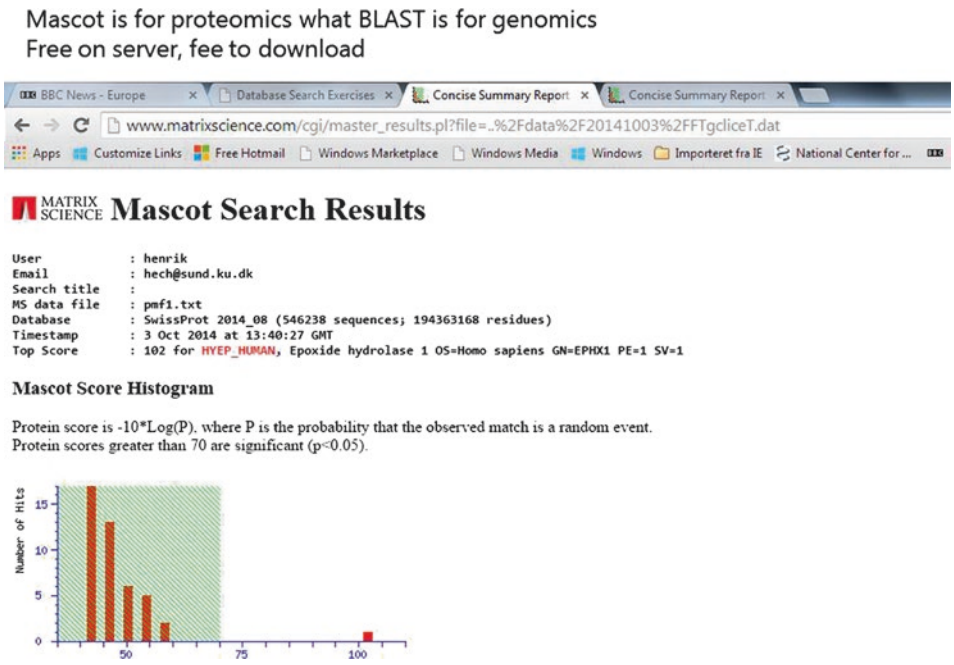


Fig. 3.11 Mascot proteomics search tool (► http://www.matrixscience.com/)

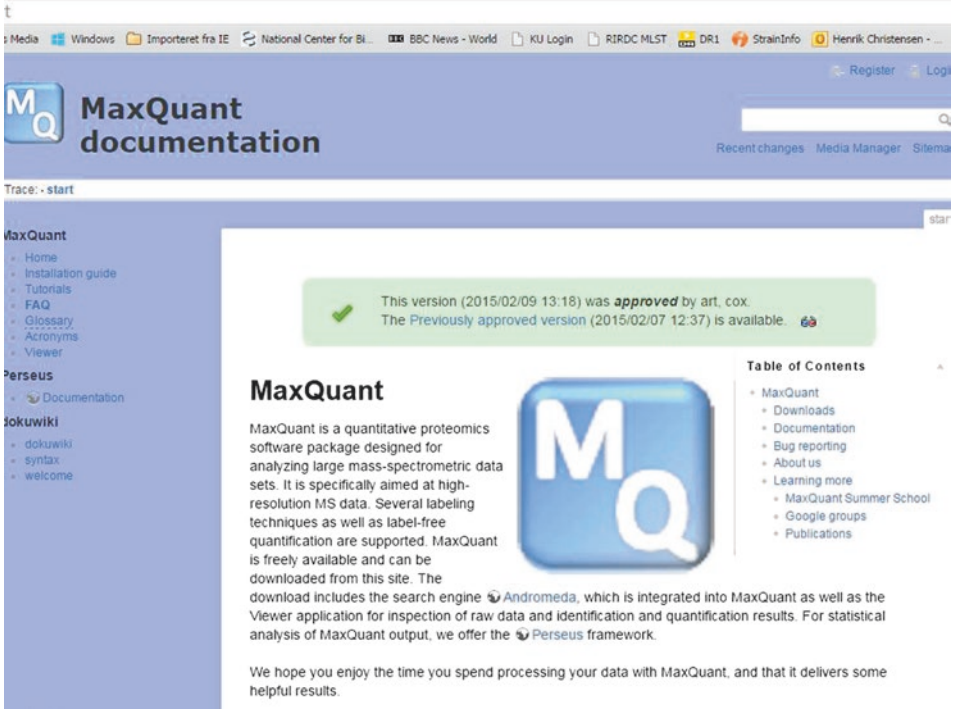


Fig. 3.12 MaxQuant (Cox). Freeware used for shotgun proteomics. Includes Perseus module for statistics and Andromeda search engine based on Mascot (Fig. 3.11) (► http://www.biochem.mpg.de/5111795/maxquant)

Table 3.7 Help to databases

Database	Subject	URL
Instructions to Swiss-Prot	The Swiss-Prot manual, here you will find all documentation	► http://www.expasy.org/sprot/userman.html
Instructions to NCBI	The NCBI handbook	► https://www.ncbi.nlm.nih.gov/books/NBK143764/
Overview of all databases	First issue of *Nucleic Acid Research* each year	► https://academic.oup.com/nar/issue/45/D1

3.7 Help to Databases

The major bioinformatics databases maintain updated online "handbooks" (Table 3.7).

3.8 Activities

3.8.1 Download a Sequence from NCBI

We will download the DNA sequence for the *cdtB* gene encoding the CdtB protein mentioned in ► Sect. 3.5.1.

Open GenBank (NCBI) from your browser with the URL: ► http://www.ncbi.nlm.nih.gov/.

Choose "Nucleotide" in the **Search window**, and write the accession number **U04208** in the window after "for" and press **Search**. As a standard, the result is shown in GenBank format. The view should be the same as Fig. 3.2.

Convert the format to FASTA by selecting FASTA (left top corner).

Save the file on your computer by clicking **Send to:** (top right corner), select **File** in the **Choose Destination** window, and select **FASTA** in the **Format window** and **Create File**.

Open the file by **WordPad,** check that it is in FASTA format, and save it on the computer with the acc. no. as name and filename extension, fasta or fna (e.g., U04208.fasta).

You actually downloaded the DNA sequence of *cdtA*, *cdtB*, and *cdtC* genes since they are located in polycistrons. If you only want to download the DNA sequence of the CDS of one gene, then press the CDS link when you are in the GenBank format, and then press the FASTA link in the lower right corner, and then only the DNA sequence of the CDS will be downloaded.

■ Batch Download

If you need many DNA sequences to download, all acc. no. can be written in the search field at NCBI just separated by space, and all can be downloaded at once. However, in this case you will get the full sequence of all entries and not only the CDSs.

3.8.2 Download a Genome from NCBI

To access genomic sequences of prokaryotes at NCBI,

► http://www.ncbi.nlm.nih.gov, select **Genome** in right panel and then **Microbe** in the middle column, and then **Browse microbial genomes** to the left.

Insert the organisms' name, *Haemophilus influenzae*, under **Genome information by organism** and press **Search**.

Use the first strain (**Rd KW20**) at **Organisms name** as link, and get access to a sub-list for that species.

The INSDC or RefSeq columns show the acc. numbers that can be used as links to the sequences. Use L42023 as link and open the GenBank record. This information can be downloaded as just described at ▶ Sect. 3.8.1 for a record with a few genes.

Try to open the file with a text editor to control that is what you expected.

You should see a typical GenBank format just with much more information than the one on ◘ Fig. 3.2.

At **Genome Region** the **Graphics** provides representation of annotated genes by arrows. You will have to zoom in. You can select "tracts" like for eukaryotic browsers. Right click in the window will allow you to save a pdf format of the view.

3.8.3 Cn3D Viewer for Protein Structures

Install the Cn3D 4.1 program from NCBI by downloading the program from ▶ ftp://ftp.ncbi.nih.gov/cn3d/, and save it on your computer. Install Cn3D:

- For Window users download and run ▶ Cn3D-4.3.1_setup.exe file.
- Mac users: see the instruction on NCBI to install X11, and then it should work.

When the Cn3D viewer is installed, you can open the structures from NCBI at ▶ http://www.ncbi.nlm.nih.gov/ and select **Structure**. Write the acc. no. (2F1N) of the structure in the **Search** window and press **Search**. Go to the section right to the graphic model and press **Download.** Here you can download the structure in **MMDB** format on your computer.

You can now view protein structures stored on your computer by either activating the **Cn3D** program or activating the program and select the stored structure with **Open**.

The viewer will show a 3D model in "Worm" format where the string is the protein sequence and the "pencils" and "arrows" illustrate α-helices and β-sheets, respectively. The output should be like ◘ Fig. 3.6**.** The position in the primary protein sequence is seen in the "**Sequence/Alignment**" viewer. The amino acid pos. is shown below the sequence, and by "mousing over" the sequence, it is marked in the 3D structure. The whole primary structure may not have been listed if it is not included in the structure. Different graphical outputs can be selected at **Style** | **Rendering Shortcuts** (tubes, wire, ball and stick, space fill).

3.8.4 Deposition of Sequence with GenBank

3.8.4.1 Procedure for Single DNA Sequences

Before you deposit a DNA sequence, you need to be 100% certain about the unbiased nature of the sequence and have all metadata available for the sequence. The tool BankIt will be most easy to start with. You need to register as user by following the link: ▶ http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank.

Sequences are submitted by interactive online approach.

Do not upload the translation independently. The program will automatically translate the DNA sequence. For partial sequences mark 5′ partial and 3′ partial.

3.8.4.2 Genomic Sequence

The following instruction is based on one bacterial strain with the genome assembled into contigs with CLC or similar program listed in ▶ Table 2.2 (▶ Chap. 2).

You should prepare the contigs in a text file and control that the sequences are not including Ns. This means that they should be assembled without ambiguities.

You edit all description lines (first line of FASTA format with the > sign) to >contx [organism = NNN] [strain = XY] where x is the contig number, NNN the species, and XY the strain number.

The replace function in WordPad can be used for that. Alternatively, if your genome has many contigs or you have many genomes, you can use a format function in CLC Genomics Workbench for this: In **CLC** select **Help | Batch rename.**

You then upload from ▶ https://submit.ncbi.nlm.nih.gov/subs/genome/.

The NCBI-specific identifiers **Bioproject** and **Biosample** numbers are registered with NCBI during upload. At the end you will get a genome accession number in the format NNNN12345678 (◘ Table 3.5).

3.8.5 Protein Structure Prediction with Swiss Model and SPDBV

To predict the structure of a protein, we can use SWISS-MODEL found at ▶ http://swissmodel.expasy.org/ (Arnold et al. 2006).

Start modeling by uploading the protein sequence. Try the protein with acc. no. OOF68275. See ▶ Activity 3.8.1 of how to download. Select **Build Model**. When the analysis is done, usually the top result is the best prediction. However, sometimes a protein can be split up on many 3D models. This can, for instance, happen with a long autotransporter protein. The output(s) show the best hit to a structure already known.

The bars show how good the prediction of structure has been through the sequence based on different tools. This information can be used if you are interested in specific regions of the sequence, for example, conformations of regions involved in substrate binding.

You can download the predicted model (sequences with .pdb extension) by a **Model** selecting **SPDBv format** or **pdb format,** and save the model.

Use Vast to convert pdb to a structure that can be visualized in Cn3D (▶ Activity 3.8.3) (▶ https://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html).

As an alternative, you can **search for template** if you have no clue about the structure, or specify a model if you want to test the similarity to that one.

The models can be opened with the **SPDBV_4.10_PC**. You need to install the program on the PC first.

▶ http://swiss-pdb-viewer.software.informer.com/download/

This happens sooner than expected, and the panel is rather narrow and the file opened from **File | Open**. Tools to move, turn, and zoom in on the structure from the narrow panel are available.

It is possible to look at the **Ramachandran plot** from **Select | All | Wind** (◘ Fig. 3.10).

Take-Home Messages

- Bioinformatics databases store biological information from life science research, most importantly DNA and protein sequences and protein structures.
- Primary databases, such as GenBank, DDBJ, and EMBL-ENA, are also called archives or repositories, and they take information directly from the individual researcher which "owns" the data with privileges to change them.
- The secondary databases like Swiss-Prot and PDB are curated, and they perform a quality control and sorting of information (non-redundancy) before the information is made accessible to the public.
- Domains are compact units of proteins that may behave independently of the proteins and can be associated with specific functions.
- Motifs are conserved regions of proteins which may be part of domains.
- Protein structure databases and prediction tools refer to secondary structures, motifs, domains, as well as protein 3D structures.
- The function of a protein can be predicted even based on a rather low identity to other known proteins and rather low similarity to protein structures.
- Proteomics is dealing with the prediction of function of proteins based on the measurements of the mass-to-charge ratio (m/z) and comparison to the Swiss-Prot database.

References

André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M. 2014. CAZyme discovery and design for sweet dreams. Curr. Opin. Chem. Biol. 19:17–24.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22, 195–201.

Attwood TK, Avison H, Beck ME, Bewley M, Bleasby AJ, Brewster F, Cooper P, Degtyarenko K, Geddes AJ, Flower DR, Kelly MP, Lott S, Measures KM, Parry-Smith DJ, Perkins DN, Scordis P, Scott D, Worledge C. 1997. The PRINTS database of protein fingerprints: a novel information resource for computational molecular biology. J Chem Inf Comput Sci. 37, 417–424.

Bagos PG, Liakopoulos TD, Spyropoulos IC, & Hamodrakas SJ. 2004. A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. BMC Bioinformatics 15;5:29.

Barker WC, George DG, Mewes HW, Pfeiffer F, & Tsugita A. 1993. The PIR-International databases. Nucleic Acids Res. 21:3089–92.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers EW. 2018. GenBank. Nucleic Acids Res. 46(D1):D41–D47.

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. Nucleic Acids Res. 44, D67–72.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42, D633–42.

Cook CE, Bergman MT, Cochrane G, Apweiler R, Birney E. 2018. The European Bioinformatics Institute in 2017: data coordination and integration. Nucleic Acids Res. 46:D21–D29.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 72, 5069–5072.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–85.

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang H-Y, Dosztányi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Xenarios, I, Yeh L-S, Young S-Y, & Mitchell AL. 2017. InterPro in 2017 — beyond protein family and domain annotations. *Nucleic Acids Res.* 45, D190–D199

Gao F, Luo H, Zhang CT, Zhang R. 2015. Gene essentiality analysis based on DEG 10, an updated database of essential genes. Methods Mol. Biol. 1279:219–33.

Ghosh P. 2018. Variation, Indispensability, and Masking in the M protein. Trends Microbiol. 26, 132–144.

Gibas C, & Jambeck P. 2001. Developing Bioinformatics Computer Skills An Introduction to Software Tools for Biological Applications. O'Reilly Media, Beijing.

Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu F, Marchler GH, Song JS, Thanki N, Yamashita RA, Zheng C, Thibaud-Nissen F, Geer LY, Marchler-Bauer A, Pruitt KD. 2018. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res. 46:D851–D860.

Henikoff S, & Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 89, 10915–10919.

Higgs PG, & Attwood TK. 2005. Bioinformatics and Molecular Evolution. Wiley.

Holm, L. & Laakso, L.M. 2016. Dali server update. Nucleic Acids Res. 8;44(W1):W351–5.

Käll L, Krogh A & Sonnhammer EL. 2004. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338:1027–36.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44(D1):D457–62.

Karsch-Mizrachi I, Takagi T, Cochrane G, International Nucleotide Sequence Database Collaboration. 2018. The international nucleotide sequence database collaboration. Nucleic Acids Res. 46(D1) :D48–D51.

Kelley LA, Mezulis S, Yates CM, Wass MN & Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 10:845–58.

Kodama Y, Mashima J, Kosuge T, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, Takagi T. 2018. DNA Data Bank of Japan: 30th anniversary. Nucleic Acids Res. 46(D1):D30–D35.

Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 2018 Jan 4;46(D1):D493–D496.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6, 610–618.

Nakai K, Horton P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci. 24:34–6.

NCBI. 2018. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 46, D8-D13.

Petersen TN, Brunak S, von Heijne G, & Nielsen H 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8:785–786.

Petsko GA & Ringe D 2004. Protein structure and function. Primers in Biology. New Science Press Ltd. London, UK.

Pickett CL & Whitehouse CA.1999. The cytolethal distending toxin family. Trends Microbiol. 7, 292–297.

Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, Green RK, Goodsell DS, Hudson B, Kalro T, Lowe R, Peisach E, Randle C, Rose AS, Shao C, Tao YP, Valasatava Y, Voigt M, Westbrook JD, Woo J, Yang H, Young JY, Zardecki C, Berman HM, Burley SK. 2017. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. 45:D271–D281.

Silvester N, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L, Cleland I, Harrison PW, Jayathilaka S, Kay S, Keane T, Leinonen R, Liu X, Martínez-Villacorta J, Menchi M, Reddy K, Pakseresht N, Rajan J, Rossello M, Smirnov D, Toribio AL, Vaughan D, Zalunin V, Cochrane G. 2018. The European Nucleotide Archive in 2017. Nucleic Acids Res. 46(D1):D36–D40.

Sonnhammer ELL & Kahn D. 1994. Modular arrangement of proteins as inferred from analysis of homology, Protein Sci. 3, 482–492

Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S. 2015. An overview of bioinformatics tools for epitope prediction: implications on vaccine development. J Biomed Inform. 53: 405–14.

Tsirigos KD, Bagos PG, Hamodrakas SJ. 2011. OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. Nucleic Acids Res. 39(Database issue):D324–31.

UniProt. 2018. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 46: 2699.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018 May 21. https://doi.org/10.1093/nar/gky427. [Epub ahead of print]

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 42(Database issue):D643–8.

Pairwise Alignment, Multiple Alignment, and BLAST

Henrik Christensen and John Elmerdahl Olsen

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_4

What You Will Learn in This Chapter

Pairwise alignments and multiple alignments are the basic tools to compare sequences. BLAST is the most frequently used bioinformatics program to compare your own sequence (query sequence) to all sequences in a database (subject sequences) based on local pairwise alignments. The outcome of the BLAST analysis provides qualitative information about homologous sequences and can quantify the identity of the query sequence to the sequences in the database. You will learn about multiple alignments and how to construct them. Multiple alignments are used for a range of applications described in the other chapters of the book.

4

4.1 The Pairwise Alignment Problem

A pairwise alignment is a model of the homology between two sequences considering all nucleotides or amino acids and all deletion and insertions.

Two sequences to be compared could, for instance, be the ones below:

- GCAGTAGCATGACGATAG
- GCGGTAGCATGATAC

A pairwise alignment requires a model for the evolutionary steps that led to the differences observed. First, we should test if they are homologous. Such a test can be done by BLAST (see ▶ Sect. 4.3), and it will tell if the sequences are from the same gene. Once this has been established, we need to model evolution to construct the pairwise alignment. The simplest assumptions are that identical nucleotide pairs and that identical amino acids pair and, further, that different amino acids with similar physiochemical properties form pairs.

For most comparisons it is preferred to base pairwise alignments on the amino acid sequences especially if the sequences are very divergent. Pairwise nucleotide comparisons are preferred for closely related sequences only.

Gaps are used to show that an amino acid or nucleotide is without a match in the other sequence and the gaps represent insertions or deletions in an evolutionary context. Most evolutionary modeling assumes that it is more difficult to form gaps than to form mismatches and that it is more difficult to form a gap than to extend one already formed. Scoring systems are used to separate matches (high score) from mismatches (low score) and gaps (very low score). For proteins, scoring matrices (see ▶ Sect. 4.1.2.1) are used. At the end, scores between amino acids or nucleotides are summarized over the whole pairwise alignment, and penalties for gaps are subtracted. The result is a unit score for the pairwise alignment given the parameters used. Given the same sequences compared, the pairwise alignment with the highest total score is preferred compared to one with a lower score.

When the pairwise alignment has been constructed, the identity and similarity can be quantified. The identity is the number of nucleotides or amino acids matching in two sequences compared at all positions between the two sequences. Similarity is a further comparison also considering different types of nucleotides or amino acids as well as the gaps.

An alignment of the sequences above, and a possible scenario leading to the differences between them, is shown in ◘ Fig. 4.1.

The principles of pairwise alignments are used for most multiple alignment programs and for BLAST. The three tools are fundamental to carry out most bioinformatics (◘ Fig. 4.2).

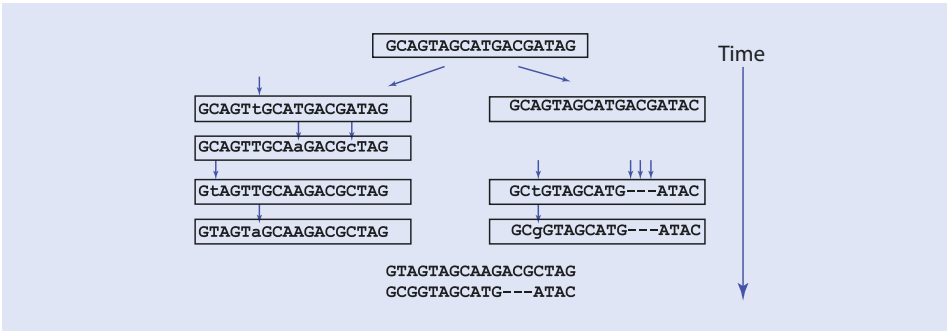


Fig. 4.1 Hypothetical example of a pairwise alignment of two sequences if their evolution was known. The first sequence is the ancestor sequence which then diverged into two new lineages. In the sequence to the left, five-point mutations occurred over time, and in the one to the right, two-point mutations plus a deletion of three nucleotides happened probably resulting in the loss of an amino acid. With the exact knowledge of the position of the gap, we can construct the pairwise alignment at the bottom without any assumptions needed

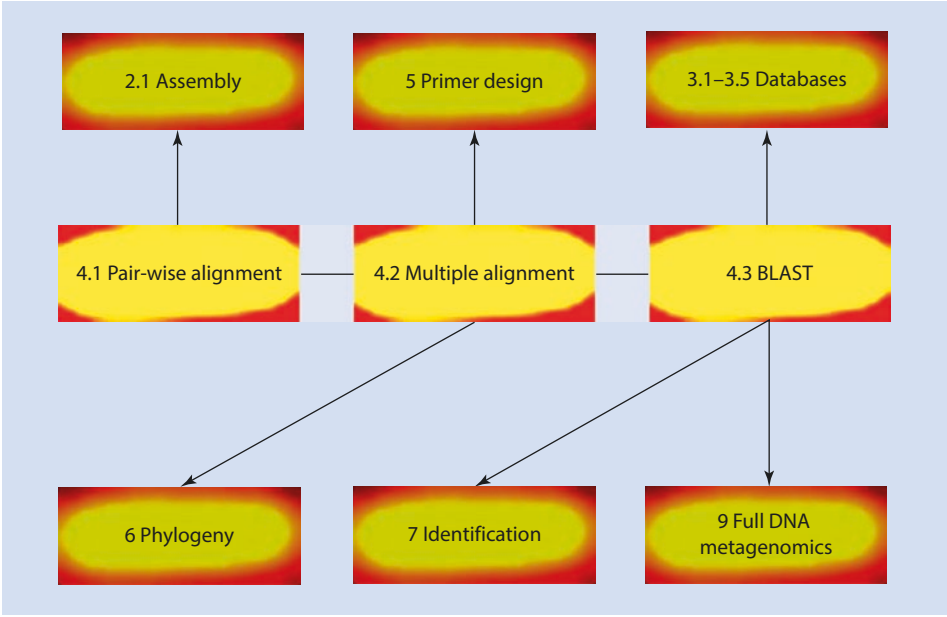


Fig. 4.2 Pairwise and multiple alignments as well as BLAST introduced in this chapter provide the background for most other chapters in this book

4.1.1 Global or Local Pairwise Alignments?

Before a pairwise alignment is constructed, it needs to be decided if it should be global or local. A pairwise alignment is global if it is know that the sequences are homologous in their full length. In this situation, it is sound to align both sequences in their full length. If preliminary evidence has showed that the genes are encoding for the same proteins and the full length of the gene has been sequenced, a global alignment can be constructed.

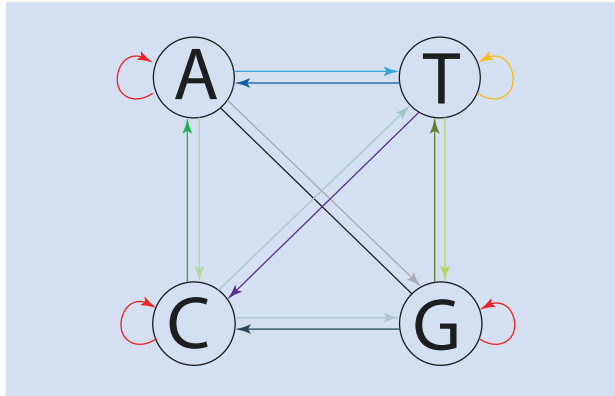


Fig. 4.3 Markov substitution model of nucleotides. The consequence of the Markov model is that it only depends on the observed nucleotide or amino acid and not on the past changes. The arrows with different colors indicate that potentially 16 different probabilities exist for changes (unrestricted model on Table 4.3). (Modified from Durbin et al. 1999)

A local alignment is needed if it is know that one sequence is shorter than the other and that it cannot be related to the other in its full length. This can happen if the DNA sequence of one gene has not been determined in the full length or if the domain structure of the proteins that the genes encode for are rather divergent.

4.1.2 Substitution Matrices

Substitution matrices are used to model the probability of mutual amino acid or nucleotide substitutions in sequences. Amino acids or nucleotides with a lower probability of changing are given more weight in the comparisons compared to those with a higher probability of changing. The changes are assumed to occur according to a Markov model where changes only depend on the current nucleotide or amino acid observed and not on past changes (Fig. 4.3).

4.1.2.1 Amino Acid Substitution Matrices

Amino acids have different biochemical and physical properties that influence their relative replaceability during evolution (Table 4.1). The probability of replacement has been related to the physiochemical properties of the amino acids in the way that hydrophobic amino acids are easier replaced by other hydrophobic amino acids again compared to other types such as charged amino acids. The substitution matrices reflect these properties as seen from Fig. 4.4. Here the score between the two hydrophobic amino acids isoleucine and histidine is 5, whereas the score between isoleucine and the positive charged arginine is only −4. Some amino acids belong to more groups, for instance, histidine which is hydrophobic, positively charged, polar, and aromatic.

The PAM (percent accepted mutations) matrices are based on global alignments of closely related proteins. The number at the end of a PAM matrix refers to the number of steps required for a given percent change. PAM1 corresponds to 99% identity (1% change) between the aligned sequences. This means that PAM112 will recognize more distantly related sequences (40% identity) than PAM23 (80% identity). PAM matrices are based on global alignments and therefore best suited for global pairwise alignments used in evolutionary studies (Table 4.2).

BLOSUM (blocks substitution matrix) matrices are based on local alignments. The number at the end of a BLOSUM matrix refers to the minimum percent identity allowed

Table 4.1 Physiochemical properties of amino acids

Amino acid	Hydrophobic	Positive	Negative	Polar	Charged	Small	Tiny	Aliphatic	Aromatic
Ile	+							+	
Leu	+							+	
Val	+					+		+	
Cys	+					+			
Ala	+					+	+		
Gly	+					+	+		
Met	+								
Phe	+								+
Tyr	+			+					+
Trp	+			+					+
His	+	+		+	+				+
Lys		+		+	+				
Arg		+		+	+				
Glu			+	+	+				
Gln				+					
Asp			+	+	+				
Asn				+		+			
Ser				+		+	+		
Thr	+			+		+			
Pro						+			

in the set of aligned sequences used to build the matrix. This means that BLOSUM50 will recognize more distantly related sequences than BLOSUM80 (Table 4.2).

4.1.2.2 Nucleotide Substitution Matrices

Substitution matrices for nucleotides give higher weight to transitions (G to A or C to T and vice versa) than to transversions (G to C, G to T, A to C, A to T, and vice versa) (Table 4.3). This is mainly related to the larger size of the purine compared to the pyrimidine nucleotides. The simplest model is the Jukes and Cantor model where equal probabilities for the substitution of all four nucleotides are assumed (Table 4.4). The transversion/transition bias is included in the models of HKY, Kimura, Tamura-Nei, Tamura, and the general reversible model. An additional account for the frequency of nucleotides is included in models of Tamura-Nei, the equal input, and the general reversible models. These frequencies are the observed frequencies of nucleotides under comparison. The unrestricted model allows different probabilities of changes for all 12 combinations of nucleotides (Table 4.4).

```
A  5
R -2  7
N -1 -1  7
D -2 -2  2  8
C -1 -4 -2 -4 13
Q -1  1  0  0 -3  7
E -1  0  0  2 -3  2  6
G  0 -3  0 -1 -3 -2 -3  8
H -2  0  1 -1 -3  1  0 -2 10
I -1 -4 -3 -4 -3 -3 -4 -4 -4  5
L -2 -3 -4 -4 -2 -2 -3 -4 -3  2  5
K -1  3  0 -1 -3  2 -1 -2  0 -3 -3  6
M -1 -2 -2 -4 -2  0 -2 -3 -1  2  3 -2  7
F -3 -3 -4 -5 -2 -4 -3 -4 -1  0  1 -4  0  8
P -1 -3 -2 -1 -4 -1 -1 -2 -2 -3 -4 -1  3 -4 10
S  1 -1  1  0 -1  0 -1  0 -1 -3 -3  0 -2 -3 -1 -5
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 -2  5
W -3 -3 -4 -5 -5 -1 -3 -3 -3 -3 -2 -3 -1  1 -4 -4 -3
Y -2 -1 -2 -3 -3 -2 -2 -3  2 -1 -1 -2  0  4 -3 -2 -2  2  8
V  0 -3 -3 -1 -1 -3 -3 -4 -4  4  1 -3  1 -1 -3 -2  0 -3 -1  5
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
```

Fig. 4.4 The BLOSUM50 scoring matrix. Note that low scores both can be 0 and with a minus sign. Low scoring pairs occur with high frequencies in sequences. (Jesper Larsen is acknowledged for the figure)

Table 4.2 Practical rules for use of PAM and BLOSUM amino acid substitution matrices

Application	Default	Related sequences	Distant sequences
Search	BLOSUM62	BLOSUM80	BLOSUM45
Evolution	PAM100	PAM50	PAM250

Table 4.3 The sixteen different types of nucleotide pairs that can be observed between two sequences

Class	Nucleotide	Pair		
Identical nucleotides	AA	TT	CC	GG
Transition-type pair	AG	GA	TC	CT
Transversion-type pair	AT	TA	AC	CA
	TG	GT	CG	GC

Modified from Nei and Kumar (2000)

Table 4.4 Models of nucleotide substitutions

	A	T	C	G	A	T	C	G
	Jukes and Cantor				HKY			
A	–	α	α	α	–	βg_T	βg_C	αg_G
T	α	–	α	α	βg_A	–	αg_C	βg_G
C	α	α	–	α	βg_A	αg_T	–	βg_G
G	α	α	α	–	αg_A	βg_T	βg_C	–
	Kimura				Tamura-Nei			
A	–	β	β	α	–	βg_T	βg_C	$\alpha_1 g_G$
T	β	–	α	β	βg_A	–	$\alpha_2 g_C$	βg_G
C	β	α	–	β	βg_A	$\alpha_2 g_T$	–	βg_G
G	α	β	β	–	$\alpha_1 g_A$	βg_T	βg_C	–
	Equal input				General reversible			
A	–	αg_T	αg_C	αg_G	–	$a g_T$	$b g_C$	$c g_G$
T	αg_A	–	αg_C	αg_G	$a g_A$	–	$d g_C$	$e g_G$
C	αg_A	αg_T	–	αg_G	$b g_A$	$d g_T$	–	$f g_G$
G	αg_A	αg_T	αg_C	–	$c g_A$	$e g_T$	$f g_C$	–
	Tamura				Unrestricted			
A	–	$\beta\Theta_2$	$\beta\Theta_1$	$\alpha\Theta_1$	–	a_{12}	a_{13}	a_{14}
T	$\beta\Theta_2$	–	$\alpha\Theta_1$	$\beta\Theta_1$	a_{21}	–	a_{23}	a_{24}
C	$\beta\Theta_2$	$\alpha\Theta_2$	–	$\beta\Theta_1$	a_{31}	a_{32}	–	a_{34}
G	$\alpha\Theta_2$	$\beta\Theta_2$	$\beta\Theta_1$	–	a_{41}	a_{42}	a_{43}	–

Modified from Nei and Kumar (2000)
g_A, g_T, g_C, and g_G are the nucleotide frequencies, $\Theta_1 = g_C + g_G$, $\Theta_2 = g_A + g_T$. The nucleotide frequencies either can be estimated from the data compared or set as fixed parameters
a_{ij} is the substitution rate from the nucleotide in the i-th row to the nucleotide in the j-th column

4.1.3 Gaps

Gaps are used to model insertions or deletions in sequences. If arbitrarily many gaps are inserted, this can lead to high-scoring alignments of nonhomologous sequences. To abolish this effect, gaps are penalized when alignments are constructed to obtain relatively few gaps and separate penalties used for gap opening and gap elongation. In the linear model, the penalty Υ is proportional to the number of gaps (g) given the penalty for one gap d:

Linear gap penalty score:

$$\Upsilon(g) = -gd$$

If the penalty of one gap is 8 and there are 3 gaps, the total penalty will be – 24.

To better model biological events where many gaps often occur in a row, the affine penalty score model is used where the cost is higher for inserting the first gap than to extent this gap to lengths of two and more:

$$\Upsilon(g) = -d - (g-1)e$$

where $\Upsilon(g)$ = gap penalty score of gap length g, d = gap opening penalty, and e = gap extension penalty. For a gap of length 3 with a penalty of opening the gap of 8 and of extending the gap with two more, the total penalty will be – 16.

4.1.4 Dynamic Programming

Dynamic programming is a way to compute the pairwise alignment? If all possibilities for the pairwise alignment of two sequences should be tested, there would be around 10^{59} possibilities for two nucleotide sequences of 100 in length. This number is approx. the same as the number of molecules in the Milky Way, and even the largest computer would not be able to handle the problem. There is also no need to consider possibilities without relevance, for instance, that one nucleotide in one sequence would pair only with gaps in the other.

To reduce the computing effort but still to perform a careful analysis, dynamic programming is used. The computer is only testing relevant comparisons (high scores) and keeping in memory the highest ones already calculated. The most relevant dynamic programming algorithms for comparison of global and local pairwise alignments were published be Needleman and Wunsch (1970) and Smith et al. (1981), respectively.

The Needleman and Wunsch algorithm is used when it is known that the sequences represent exactly a gene or protein in full length (few amino acids or nucleotide differences are tolerated) and it is used to build global pairwise alignments.

The Smith and Waterman algorithm is used with partial sequences or with sequences of unknown length to build local alignments. For both algorithms, the adjustment of gap penalty functions depends on previous knowledge about domain structures, repeats, and other properties. Both algorithms and their implementations in computer programs can be used for both amino acid and nucleotide sequences.

4.1.4.1 Needleman and Wunsch

The score function in this model is illustrated in Fig. 4.5. This function is used on all comparisons between the two sequences. Here we will use the example with the two sequences:

- Sequence 1: HEAGAWGHEE
- Sequence 2: PAWHEAE

Figure 4.6 shows one sequence (1) on the top of the table and sequence 2 as a vertical column. First an alignment path matrix is created. For each cell, $F(i, j)$ is calculated based on the score function in Fig. 4.5. This matrix allows a stepwise calculation of score values. The score of the best alignment is calculated between the initial segment $x_{1\ldots i}$ of x

$$F(i,j) = \max \begin{cases} F(i\text{-}1, j\text{-}1) + s(i,j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

Fig. 4.5 For the Needleman and Wunsch algorithm, the score function F is defined as the maximum of the three expressions. $F(i-1, j-1)$ is the score of the previous diagonal cell in the matrix which added the score for a match between an amino acid pair in the current cell. $F(i-1, j)$ is the score for the previous cell in the column subtracted the penalty of a gap (d), and $F(i, j-1)$ is the equivalent score for the previous cell in the row subtracted the gap penalty (see Fig. 4.7 for an example of this calculation)

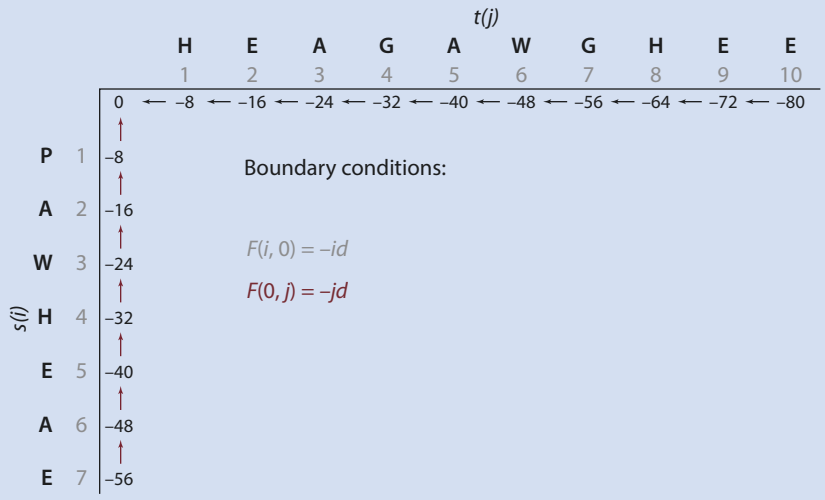


Fig. 4.6 Pairwise alignment with the Needleman and Wunsch algorithm showing the boundary conditions calculated. Here both of the sequences are paired only with gaps resulting in the lowest negative score. (Jesper Larsen is acknowledged for the figure)

up to x_i and the initial segment $y_{1...j}$ of y up to y_j. The algorithm function, $F(i, j)$, is built recursively beginning with F(0,0) = 0 (Figs. 4.6, 4.7, and 4.8). When the matrix is filled, backtracking is started (evaluation of the optimal path).

The scoring parameters are given by the BLOSUM 50 matrix (Fig. 4.4) and a linear gap penalty of $d = -8$. We will align the whole sequence length of the sequences meaning that we need to form a global alignment and use the Needleman and Wunsch algorithm. The procedure can be followed on Figs. 4.6, 4.7, 4.8, and 4.9.

A real example of the use of the Needleman and Wunsch algorithm is shown on Fig. 4.10. Here realistic long sequences are aligned by the algorithm implemented as the **needle** program in the EMBOSS package (Rice et al. 2000).

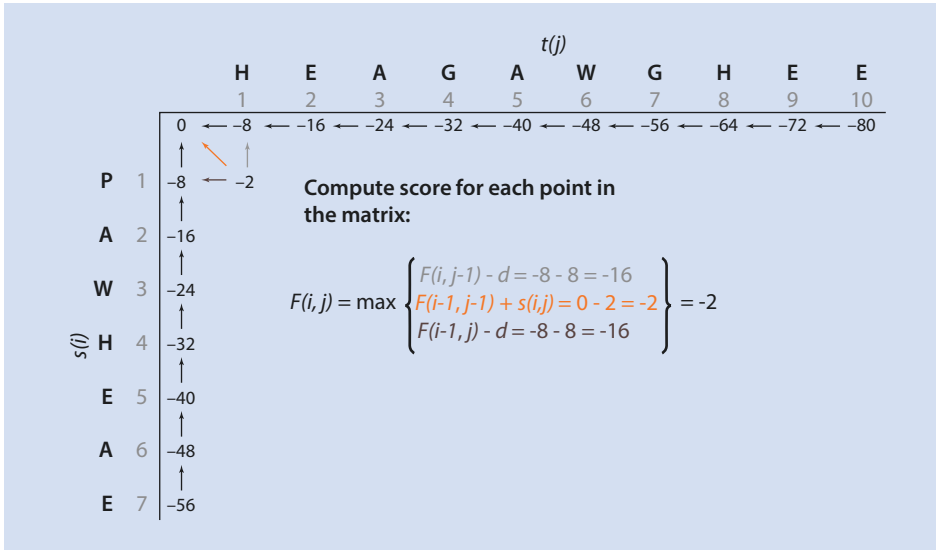


Fig. 4.7 Pairwise alignment with the Needleman and Wunsch algorithm. Here the first real position of the matrix is calculated. The three possibilities, H pairing with a gap, P pairing with a gap, or H pairing with P are considered. The one with the highest score: H pairing with P (−2) (see Fig. 4.4) gives the maximum score and is preferred. (Jesper Larsen is acknowledged for the figure)

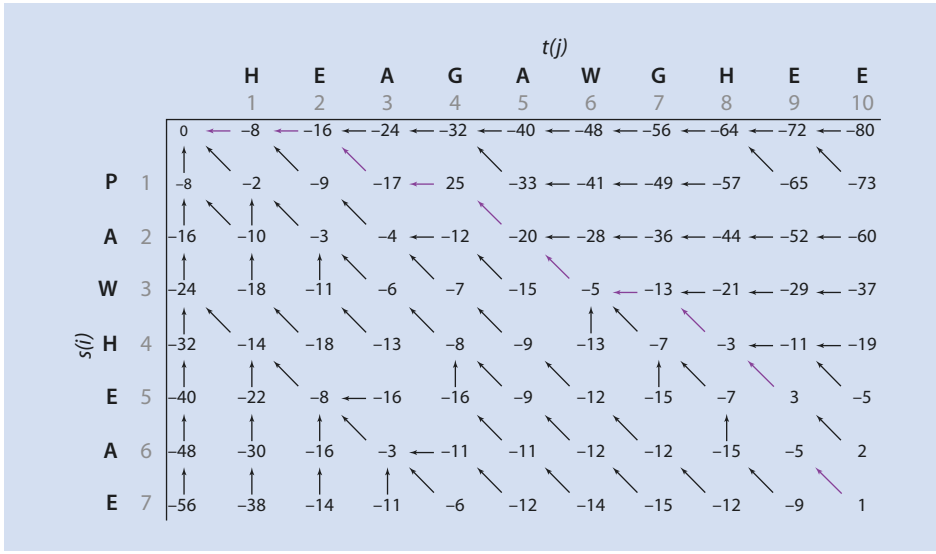


Fig. 4.8 Pairwise alignment with the Needleman and Wunsch algorithm. Here the maximum score has been inserted for all cells in the matrix, and backtracking from the bottom right corner has been performed following the path which maximizes the score. After the score of 1 for the cells to the lowest right, −5 is chosen since it was marked as a path to 1 in the forward tracking. The higher score of 2 cannot be selected since it was not marked in the forward tracking. (Jesper Larsen is acknowledged for the figure)

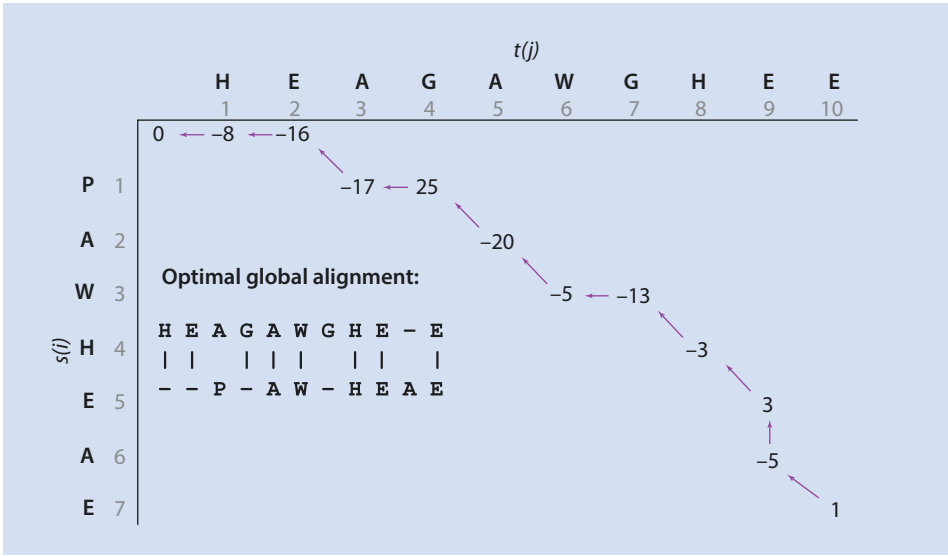


Fig. 4.9 Pairwise alignment with the Needleman and Wunsch algorithm. The optimal path for backtracking is shown again (Fig. 4.8), and the pairwise alignment has been built. Note the combination of H and P considered in Fig. 4.7 has not been selected since the backtrack path did not pass through this cell. The total score for the multiple alignments is 1 (bottom left corner). (Jesper Larsen is acknowledged for the figure)

4.1.4.2 Smith and Waterman

Now we will compute a pairwise alignment of the sequences using the Smith and Waterman algorithm. There are two differences in the algorithm compared to Needleman and Wunsch. The first is that the alignment can start/end anywhere in the matrix, and the other that backtracking is started at the highest value rather than in the lower right corner. Backtracking is terminated as soon as zero is encountered. The score function is shown in Fig. 4.11.

We will align the same two short model protein sequences with Smith and Waterman:

- Sequence 1: HEAGAWGHEE
- Sequence 2: PAWHEAE

And use the same parameters of BLOSUM50 and a linear gap penalty of $d = -8$.

Using the score function in Fig. 4.11 results in 0 in most cells when the matrix is filled in the forward direction (Fig. 4.12). The reason is that the three other possibilities in Fig. 4.11 result in negative values and we then have to use 0. Backtracking starts with the highest score (28) and terminates when 0 is reached. Note that an alternative path is also possible. However, the maximum score of this one is lower (21), and therefore the first is chosen to construct the final pairwise alignment (Fig. 4.13). Similar to the example with Needleman and Wunsch (Fig. 4.10), a more realistic example is shown in Fig. 4.14 with longer sequences. The sequences are the same that we used in Fig. 4.10, and we see little difference between the two pairwise alignments probably because they were of nearly the same length.

```
########################################
# Program: needle
# Rundate: Thu Feb 12 14:49:37 2004
# Align_format: srspair
# Report_file: outfile.align
########################################
#=======================================
#
# Aligned_sequences: 2
# 1: AAA85484.1
# 2: AAA85485.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 163
# Identity: 113/163 (69.35)
# Similarity: 137/163 (84.0%)
# Gaps: 1/163 (0.6%)
# Score: 607.0
#
#
#=======================================

AAA85484.1     1             MKFFAVLALCIVGALAHPLTSDEAALVKSSWAQVKHNEVDILYTBFKAYP 50
                             ||||||||||:|||||.||::||||:|||||.||||||||||..||.|||
AAA85485.1     1             MKFFAVLALCVVGALASPLSADEAAIVKSSWDQVKHNEVDILAAVFAAYP 50

AAA85484.1     51            DIQARFPQFAGKDLDTIKTSGQFATHATRIVSFLSELLALSGSESNLSAI 100
                             ||||:|||||||||.:||.:..||||||||||.:|:|:|||:::||||:
AAA85485.1     51            DIQAKFPQFAGKDLASIKDTAAFATHATRIVSFFTEVISLSGNQANLSAV 100

AAA85484.1     101           YGLISKMGTDHKNRGITQTQFNKFRTALVSYISSNVAWGDNVAAAWTHAL 150
                             |.|:||:|.|||.|||:..|| :||||||||.:::|:|||||||||.|||
AAA85485.1     101           YALVSKLGVDHKARGTSAAQFGEFRTALVSYLQAHVSWGDNVAAAWNHAL 150

AAA85484.1     151           DNVYTAVFQIVTA 163
                             ||.|....:.:.
AAA85485.1     151           DNTYAVALSKLE 162
```

Fig. 4.10 Pairwise alignment with the Needleman and Wunsch algorithm. A real example with sequences of realistic length has been computed using implementation of the algorithm in EMBOSS as the needle program. (see ► Activity 4.4.1 for details about this program)

Fig. 4.11 Score function for Smith and Waterman. An extra possibility of 0 can be selected in addition to the three in Fig. 4.5. The 0 is chosen if the other three are negative. This way cells are marked as 0 to account for different length of sequences in the local alignment (Fig. 4.12)

$$F(i,j) = \max \begin{cases} 0 \\ F(i-1, j-1) + s(i,j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

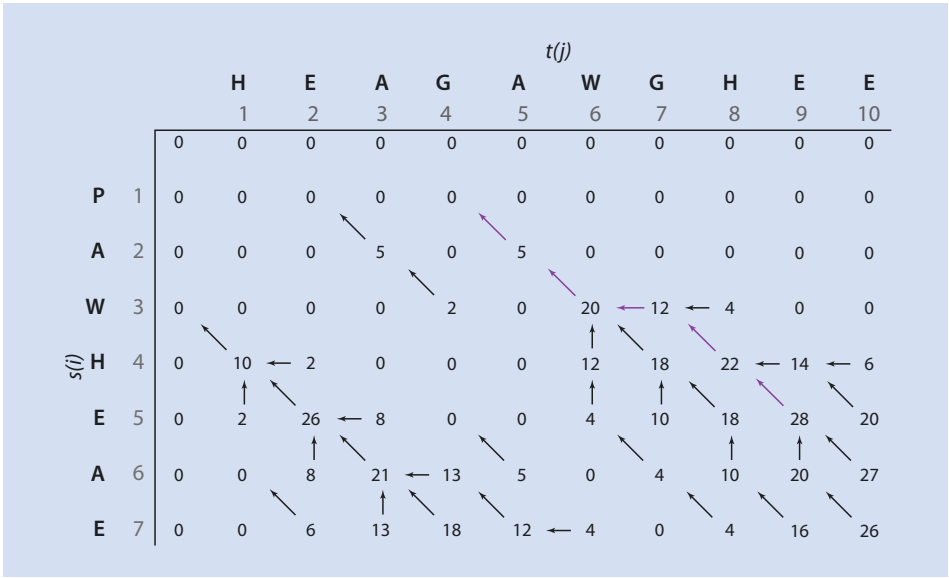


Fig. 4.12 Smith and Waterman alignment. Here the matrix has been filled with combinations based on the score function in Fig. 4.11. Backtracking started from the highest score in the matrix (28) and continued to join the highest possibilities until 0 was reached. (Jesper Larsen is acknowledged for the figure)

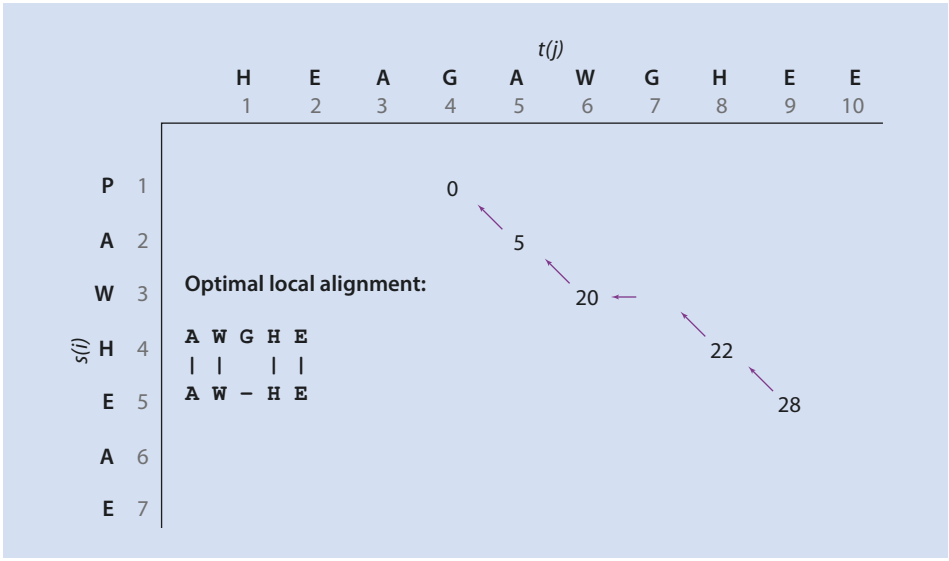


Fig. 4.13 Smith and Waterman algorithm. Here the backtrack path is shown and the resulting pairwise alignment has been constructed. Note the difference to global alignment generated by the Needleman and Wunsch algorithm in Fig. 4.9. The total score for the pairwise alignment is 28. (Jesper Larsen is acknowledged for the figure)

```
#############################################
# Program: water
# Rundate: Fri Feb 13 13:11:09 2004
# Align_format : srspair
# Report_file: outfile.align
#############################################
#=======================================
#
# Aligned_sequences: 2
# 1: AAA85484.1
# 2: AAA85485.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 159
# Identity: 113/169 (71.1%)
# Similarity: 136/159 (85.5%)
# Gaps: 0/159 (0.0%)
# Score: 609.0
#
#
#=======================================

AAA85484.1        1 MKFFAVLALCIVGALAHPLTSDEAALVKSSWAQVKHNEVDILYTBFKAYP 50
                    ||||||||||:|||||.||::||||:||||||.||||||||||..||.|||
AAA85485.1        1 MKFFAVLALCVVGALASPLSADEAAIVKSSWDQVKHNEVDILAAVFAAYP 50

AAA85484.1       51 DIQARFPQFAGKDLDTIKTSGQFATHATRIVSFLSELLALSGSESNLSAI 100
                    ||||:||||||||||.:||.:..||||||||||.:|:|:|:||:::||||:
AAA85485.1       51 DIQAKFPQFAGKDLASIKDTAAFATHATRIVSFFTEVISLSGNQANLSAV 100

AAA85484.1      101 YGLISKMGTDHKNRGITQTQFNKFRTALVSYISSNVAWGDNVAAAWTHAL 150
                    |.|:||:|.|||.|||:..||  :||||||||:.:|:|:|||||||||.|||
AAA85485.1      101 YALVSKLGVDHKARGISAAQFGEFRTALVSYLQAHVSWGDNVAAAWNHAL 150

AAA85484.1      151 DNVYTAVFQ 159
                    ||.|....:
AAA85485.1      151 DNTYAVALK 159


#---------------------------------------
#---------------------------------------
```

Fig. 4.14 Pairwise alignment with the Smith and Waterman algorithm. A real example with sequences of realistic length has been computed using the water program of EMBOSS which has implemented the Smith and Waterman algorithm (see ► Activity 4.4.2 for details about this program). Note that there are only small differences to the pairwise alignment generated by Needleman and Wunsch (Fig. 4.10). It is related to the comparable length of the two sequences

4.2 Multiple Alignment

A multiple alignment is the simultaneous alignment of three or more nucleic acid or amino acid sequences. The procedure involves the insertion of gaps in the sequences so as to maximize the overall similarity (Higgins and Sharp 1988). Multiple alignments are rarely used for their own sake but are usually created for another purpose – for instance, primer design (► Chap. 5) or analysis of phylogeny (► Chap. 6). Users select a favorite program or program package and try to optimize program settings for that.

To start the construction of a multiple alignment, we will use the same criteria as for the pairwise alignment problem. We will assume that they shared ancestors (homologous). We need to model evolution the way that every column represents only orthologous amino acids or nucleotides that have evolved from a common ancestor. The simplest assumptions are again that identical nucleotides or identical amino acids pair, that different amino acids with similar physiochemical properties pair, that it is more difficult to form a gap than to form a mismatch, and that it is more difficult to form a gap than to extend one already formed. The scoring between nucleotides and amino acids is based on the system above for pairwise alignment (▶ Sect. 4.1.2).

4.2.1 Clustal

To form a multiple alignment of the clustal type, first pairwise alignments between sequences 1–2, 1–3, 1–4, 1–5, 2–3, 2–4, 2–5…4–5 are formed (■ Fig. 4.15). This is the progressive alignment part of the process, and there are $(n(n-1))/2$ combinations where n is the number of sequences. In this example with five sequences, there are ten combinations. The next step is to construct a "guide tree" (■ Fig. 4.16) uniting the pairs with highest score. The final step is to construct the multiple alignments from the guide tree which is called profile alignment (■ Fig. 4.17). The clustal type of multiple alignments is therefore performing progressive profile alignment. Clustal has been called "a quick and dirty version of Feng and Dolittle (1987)" where "only residues that are part of the matches of a given length (k-tuple matches) are scored" (Higgins and Sharp 1988).

The developments in the clustal programs were started with Clustal (1988) (Higgins and Sharp 1988) and continued with ClustalV (five) (1992), ClustalW (weights) (1994), and ClustalX 2.0 (2007) (Larkin et al. 2007). ClustalX is the implementation of the program for PC (■ Figs. 4.18 and 4.19) and is further described in ▶ Activity 4.4.2. Clustal

```
 CLUSTAL W (1.81) Multiple Sequenc Alignments

Sequence format is Pearson
Sequence 1: 001            238 bp
Sequence 2: 002            228 bp
Sequence 3: 005            227 bp
Sequence 4: 018            227 bp
Sequence 5: 120            228 bp
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score: 98
Sequences (1:3) Aligned. Score: 86
Sequences (1:4) Aligned. Score: 75
Sequences (1:5) Aligned. Score: 80
Sequences (2:3) Aligned. Score: 86
Sequences (2:4) Aligned. Score: 76
Sequences (2:5) Aligned. Score: 81
Sequences (3:4) Aligned. Score: 76
Sequences (3:5) Aligned. Score: 81
Sequences (4:5) Aligned. Score: 76
Guide tree        file created:   [/ebi/extserv/clustalw-work/interactive/clustalw-20040304-09194253.dnd]
Start of Multiple Alignment
There are 4 groups
Aligning...
Group 1: Sequences:   2      Score:4197
Group 2: Sequences:   3      Score:3797
Group 3: Sequences:   4      Score:3662
Group 4: Sequences:   5      Score:3496
Alignment Score 10389
CLUSTAL-Alignment file created [/ebi/extserv/clustalw-work/interactive/clustalw-20040304-09194253.aln
```

■ **Fig. 4.15** Output from ClustalW showing the pairwise combinations and the grouping of sequences based on the guide tree (■ Fig. 4.16)

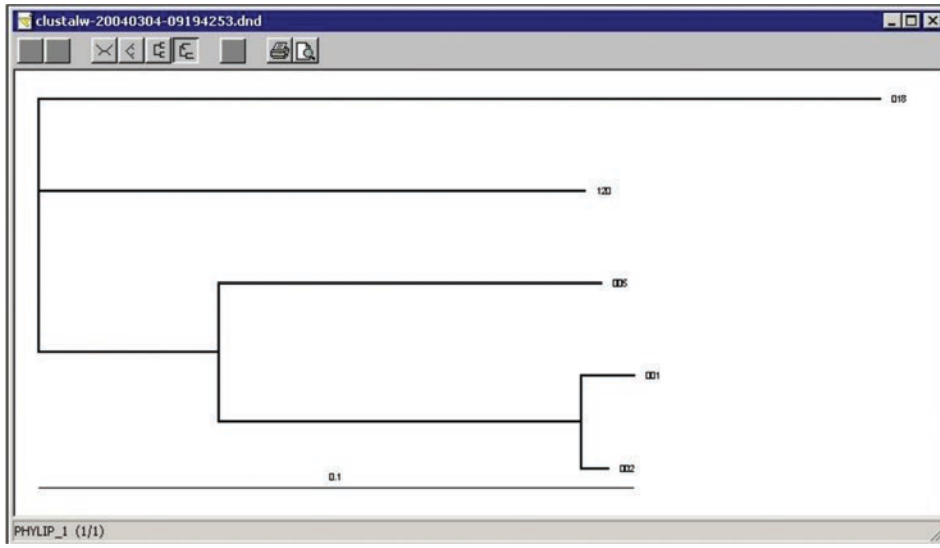


Fig. 4.16 The guide tree of ClustalW which is used by the program to perform the full multiple alignment

```
CLUSTAL W (1.82) multiple sequence alignment


001             AAAAAGAAAAGTATATGTTACATTAATATTACAATGTAATTATTTTGTTTAATTTCCCTA 60
002             AAAAAGAAAAGTATATGTTACATTAATATTACAATGTAATTATTTTGTTTAATTTCCCTA 60
005             AAAACAAAAAGTATATGTTACATTAATATTACAATGTAATTATTTTGTTTAATTTCCCTA 60
120             AAAATGAAAAGTATATGTTACATTGGTTTTACAATGTAATTATTTTGTTTAATTTCCCTA 60
018             AAAATGAAAAGTAAATATTACATTAATTTTACAATGTAATTATTTCGTTTATTTTCTCTA 60
                ****  ******* ** *******  * ***************** ***** **** ***

001             CATTTGATATAACTTTAAAACACTCCTTTTTCTCTTCTGATTATATAAAAGACAAAAAAT 120
002             CATTTTGTATAACTTTAAAACACTCCTTTTTCTCTTCTGATTATATAAAAGACAAAAAAT 120
005             CATTTTGTATAACTTTAAAACATTCAATTTTATCTTCTGGCTATACAAAAGACAAAAAAT 120
120             CATTTTGTATAACTTTAAAACATTTTATTTTCCAGTCTGTACACATAAAAGGCAAAAAAT 120
018             CATTTTGTGTAACTTTAAAGCATTTAATTTTACAACTATTTTATATAAAAGATAAAAAAT 120
                ******** ********** ** *   ****          * * *****  *******

001             ACAATTTAAGCTACAAAAAACAACAAAAAACAACAAAAAACACGACAATAAGATCGAGTA 180
002             ACAATTTAAGCTAAAAAAAAACAACAAAAAATA----------CGACAATAAGATCGAGTA 170
005             ATAATTAAAGCTACAAAAA-TAACAAAAACAA----------CTGCACTTATATCAAAAT 169
120             ATAATTAAAGCCACAAAAG-TAATAAAAAGTA----------CTGTGGTCTGATCATATA 169
018             TCAACTAAATCCATAAAAA-CAATAAAAAATC----------CTGTGGTAAGAGCAGTTG 169
                  ** * ** * * ****   ** *****           *     *   * *

001             ATGATTATATTATGTTATAATTTTTGACCTAATTTAGAATAATTATCGAGTGCAAATT 238
002             ATGATTATATTATGCTATAATTTTTGACCTAATTTAGAATAATTATCGAGTGCAAATT 227
005             GTGGTTATATTATGCTATAATTTTGGGTCTAATTTAGAATAATTATCGAGTGCAAATT 227
120             ATGAATACATTATGCTATAATCAGCCCCCTAATTTAGAATAATTATCGAGTGCAAAAT 227
018             GTTAATGTATTATGCTAAAATTTTTATTCTAATCTAGAATCATTATCGAGTGTGAATT 227
                 *   *  ****** ** ***       ***** ****** ***********  ** *
```

Fig. 4.17 The final multiple alignments made with ClustalW

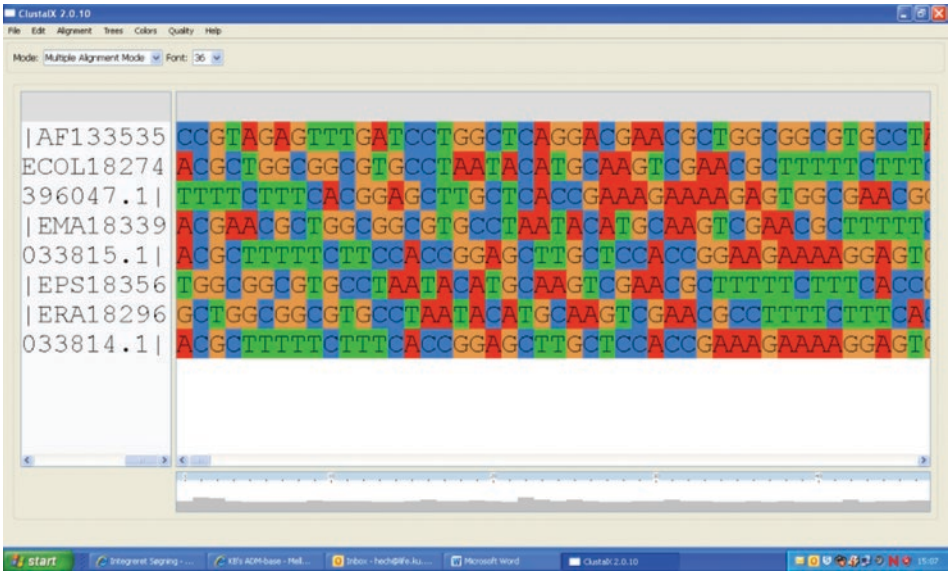


Fig. 4.18 CLUSTALX (Larkin et al. 2007) showing the sequences not yet aligned. This is seem from the unordered arrangement of the nucleotides in the columns

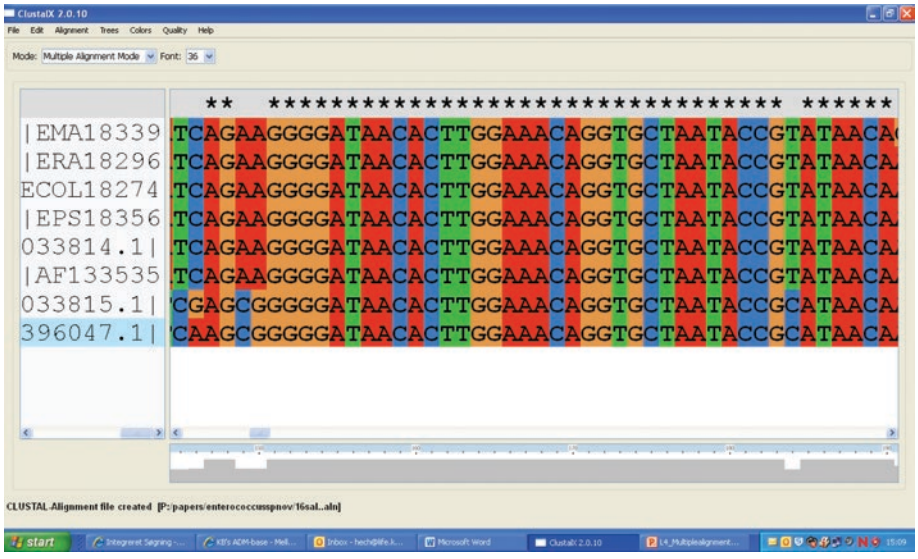


Fig. 4.19 CLUSTALX (Larkin et al. 2007) showing aligned sequences where most of the columns show only one type of nucleotide

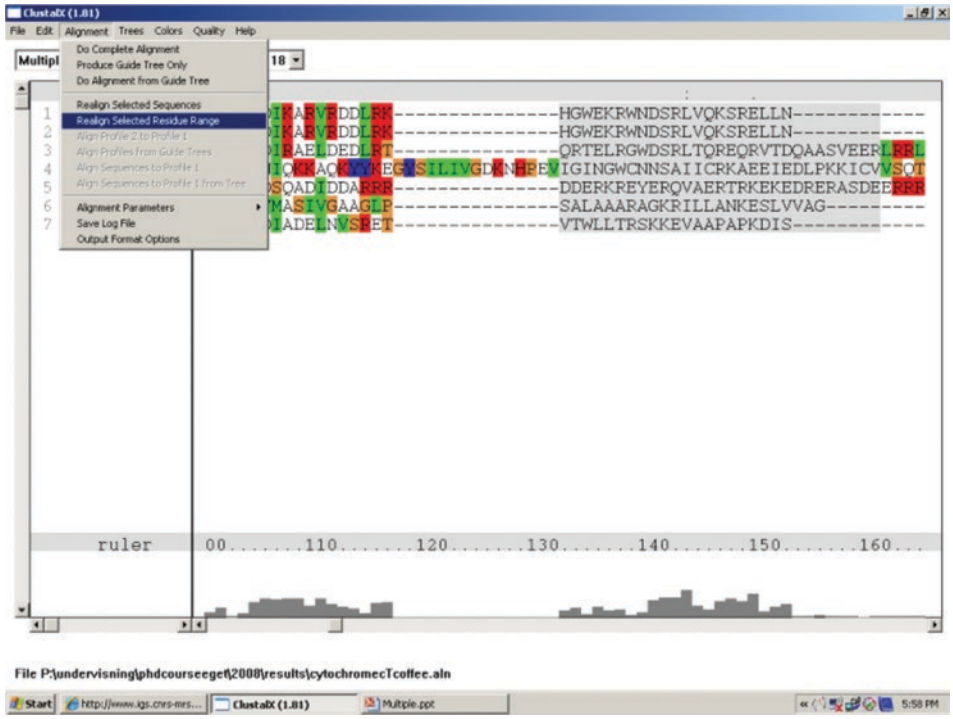


Fig. 4.20 CLUSTALX (Larkin et al. 2007) showing optimization by realignment of a region of the multiple alignment to improve local regions

Omega is the most recent version of the program for multiple alignments of very large protein datasets (Sievers et al. 2011).

ClustalX/W produces global alignments which include benefits and drawbacks inhered from the pairwise alignment part included in the construction of the alignment. Domain structures of proteins can be optimized by ClustalW/X (Fig. 4.20). ClustalW/X invokes different "hidden" functions during alignment and tends to be cluster gaps and take hydrophobic/hydrophobic properties into account. An inherent problem with the progressive approach used in ClustalX/W is that mistakes made in the initial alignment cannot be corrected later. To account for this and other problems, other programs have been constructed.

4.2.2 Other Multiple Alignment Programs

MUSCLE (Multiple Sequence Comparison by Log-Expectation) (Edgar 2004) (► https://www.drive5.com/muscle/downloads.htm) includes no guide tree. MUSCLE is claimed both to achieve better average accuracy and better speed than ClustalW2 or T-Coffee (Edgar 2004). The main focus is on protein multiple alignments, but it works also on DNA. The principle is based on K-mer comparison between sequences. K-mers are contiguous subsequence of short length (k-tuple). Related sequences tend to have more k-mers in common than expected by chance. The program can be used from MEGA7 (Tamura et al.

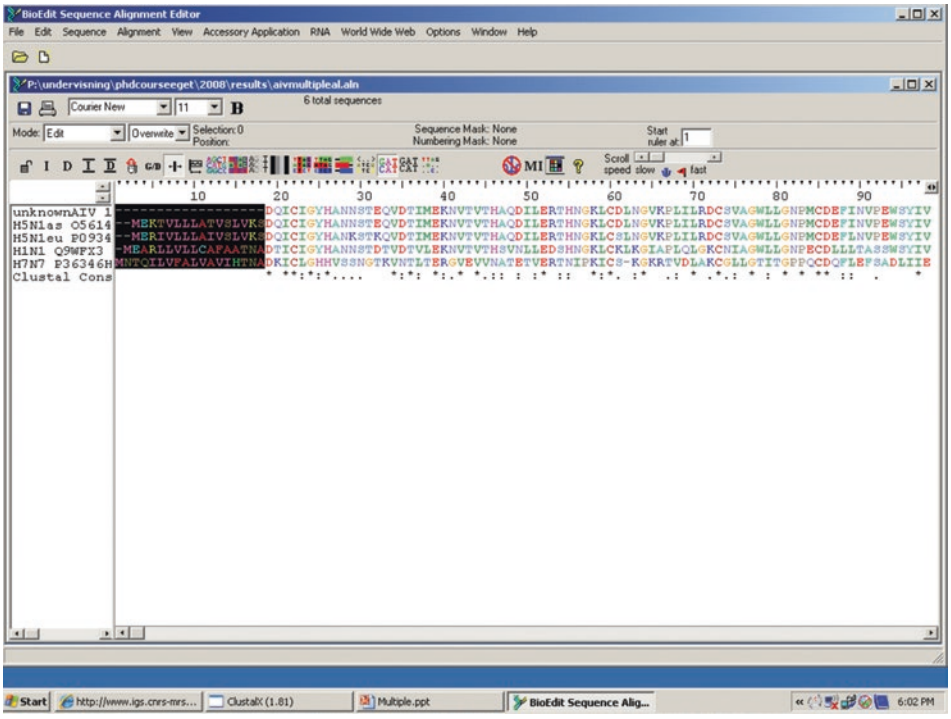


Fig. 4.21 BIOEDIT (Hall 1999) (► http://www.mbio.ncsu.edu/BioEdit/bioedit.html) used to trim a region of a multiple alignment with many gaps inserted as a consequence of lack of data (short sequences). The box marked in black was deleted

2011) (► https://www.megasoftware.net/) used in ► Chaps. 6 and 11. Like the following programs, T-Coffee and MAFFT have options to adjust alignment parameters for a range of applications.

T-Coffee (► http://tcoffee.crg.cat/) (Notredame et al. 2000) incorporates a "tree-based consistency objective function for alignment evaluation." The benefit should be high accuracy. The program comes in a range of flavors suitable for different applications.

MAFFT (Yamada et al. 2016 and other papers) (► https://mafft.cbrc.jp/alignment/software/) is also a multiple alignment package that includes applications for very large datasets.

Multiple alignment programs can be compared with sets of reference sequence (BAliBASE) (Thompson et al. 1999).

To account for the secondary structures in 16S rRNA sequence, the ARB package (Ludwig et al. 2004) (► http://www.arb-home.de/) has been developed. It has been used to construct precomputed multiple alignments in the SILVA database (► www.arb-silva-de) which is further considered in ► Chap. 8. Folding patterns in 16S rRNA genes can be predicted with Mfold (► http://unafold.rna.albany.edu/?q=mfold) (Zuker and Jacobson 1998) or similar program.

Trimming should only be done with limited datasets of closely related organisms. Trimming can be done in BioEdit (Fig. 4.21). For Mac users Jalview can be used. (► http://www.jalview.org) (Waterhouse et al. 2009)

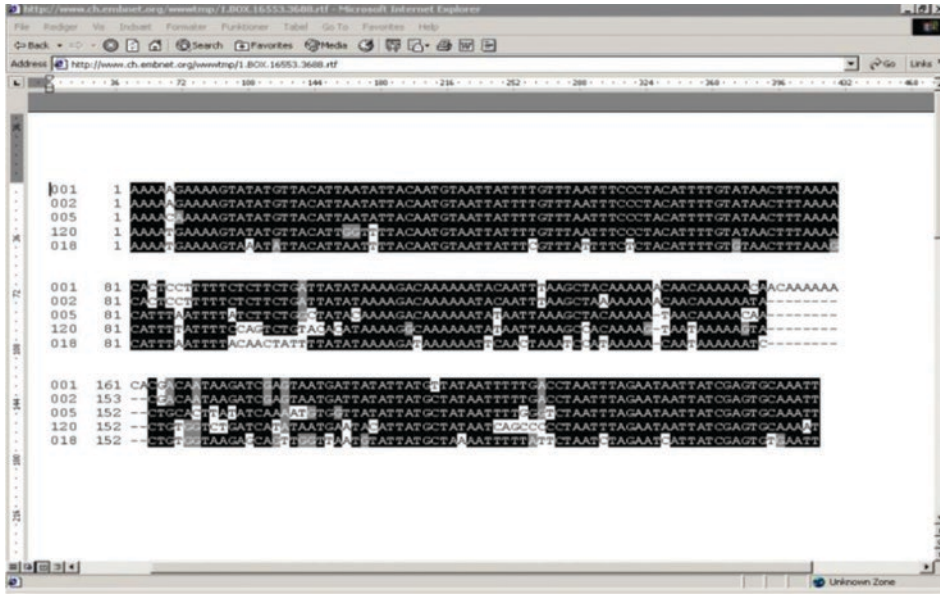


Fig. 4.22 Output from BoxShade (▶ http://www.ch.embnet.org/software/BOX_form.html). This output can be included for presentation in a publication

Graphic output of multiple alignments can be made in BoxShade (▶ http://www.ch.embnet.org/software/BOX_form.html) (Fig. 4.22).

4.3 BLAST

BLAST (Basic Local Alignment Search Tool) was originally described by Altschul et al. (1990), and this tool enabled fast search in the electronic nucleotide and protein databases during the 1990s when the Internet was invented. Later an improved algorithm was described allowing gaps to be inserted during the analysis (Altschul et al. 1997). Both versions of BLAST are in use as the "ungapped" (1990) and "gapped" (1997), respectively.

BLAST is based on a heuristics search algorithm which is able to search through the ever-growing sizes of databases. BLAST can be compared to the search tool Google in popularity within the bioinformatics community including the formation of the verb "to blast." BLAST is based on the initial identification of the users' query sequence by short pieces of sequence (words). These words are compared to the database with all the sequences (subjects). If the word is identified in a sequence of the database, the match of the word can be extended and will form a high-scoring pair (HSP). When no further extension is possible, the result is returned as a hit list to the used with indication of similarity between the query and the closest related subjects in the database. The "gapped" version of BLAST is most frequently used, and its "two-hit method" requires two non-overlapping "words" on the same diagonal with a distance of A before an extension is invoked. Gapped BLAST should only require 1/3 computer time compared to ungapped version because of less extensions of the words. BLAST uses dynamic programming to extend residues in both directions; BLAST is most suitable

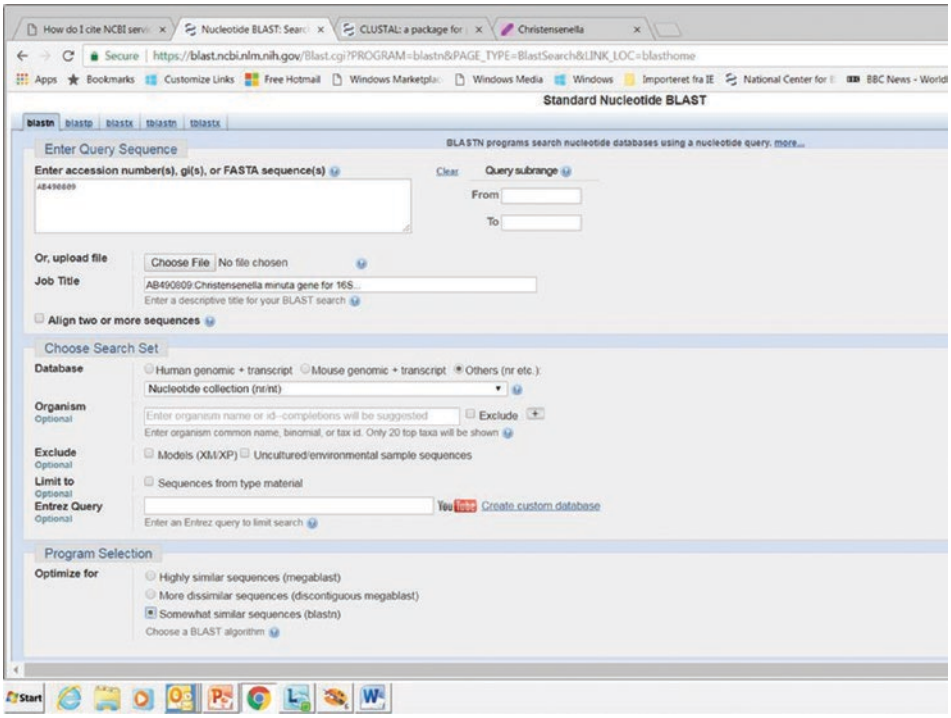


Fig. 4.23 BLAST search (Altschul et al. 1990, 1997). (BLAST [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2018 04 17]. Available from: ► https://www.ncbi.nlm.nih.gov/Blast.cgi)

for finding of subject DNA and protein sequences in databases that match with a reasonable similarity and a comparable length to the query. BLAST is good for sequences of at least 100 in length that are well represented in the databases. BLAST is not suitable to make really precise determination of similarity between sequences, and pairwise alignment programs should instead be used for that (► Sect. 4.1). BLAST is not suitable for sequence-based identification of bacterial species since there are simply too many sequences in GenBank and the type strains are not always clearly labeled (► Chap. 7). BLAST is not suitable for databases beyond a certain size and cannot be used to search very large metagenomics databases (► Chap. 9). In principle the database search of BLAST is based on local alignments. It needs to be local since we are never certain if our query sequence will be represented by a homolog in the database with similar length. BLAST is most often used on the NCBI server (► http://www.ncbi.nlm.nih.gov/BLAST/); however, the BLAST program can also be installed on your local computer and used with your own database.

4.3.1 NCBI BLAST

The most common way of performing a BLAST search is to access the program at NCBI and search in their databases. A search can be performed by a DNA or protein sequence in FASTA format, by uploading a sequence file in FASTA format or by inserting an accession number as AB490809 shown in Fig. 4.23 if the sequence is already included with

GenBank. In the example shown on Fig. 4.23, default setting has been selected for a BLAST search with a DNA sequence. Under **Program Selection** and **Optimize for different versions of BLAST** can be selected. In this case **megablast**, **discontigous megablast,** and **blastn** can be selected. The three versions differ by the scores given to matches and their way of penalizing gaps. For shorter sequences **blastn** should be used.

When the search is terminated, you will see the familiar graphical overview in Fig. 4.24 – the colored section with horizontal bars. You can see details of subject sequences by mousing over the bars in this section.

In the next section comes the descriptions with one line for each subject which gives a more detailed information about the subject sequences providing **Score**, **Query cover**, **E value,** and **accession number**. These parameters will be explained in the following.

The hit list is arranged by decreasing **Score**. In the example you see that hits are sorted by decreasing score. The identities are also decreasing but not systematically since the score is not always directly related to the identity. The longer the match between two sequences, the higher the score given the same identity. E values cannot be directly seen here since they are very small and they are just shown as "0.0."

In the alignments section, the pairwise alignments are shown. This is the most detailed information of comparison between your query sequence and the subjects in the database. The parameters from the description section are available, and additional information of gaps and visual location of both gaps and mismatches can be observed.

4.3.2 Ortholog Detection

Sequences are orthologs if they are homologs and have the same function in different species. The paralogs are also homologs but have been duplicated in evolution, and they have got divergent functions in the same species, and the sequences are often quite divergent. BLAST can be used to sort the orthologs from the paralogs. The principle of reciprocal best hits (RBH) is used as explained in Moreno-Hagelsieb and Latimer (2008) "Two genes residing in two different genomes are deemed orthologs if their protein products find each other as the best hit in the opposite genome" (Moreno-Hagelsieb and Latimer 2008). Further investigation of orthologs and paralogs can be done by phylogeny (▶ Chap. 6).

4.3.3 BLAST2 Sequences

BLAST2 sequences is a version of NCBI BLAST that allows pairwise comparisons on server or on stand-alone BLAST installed on your PC. The program is available on both DNA and protein level, and it can be used to build your own database for many purposes by including many sequences in the 2nd window (Fig. 4.25). For instance, if the query DNA or protein sequence is included with the 1st search window and a range of genomes included in the 2nd search, you can identify the query sequence in

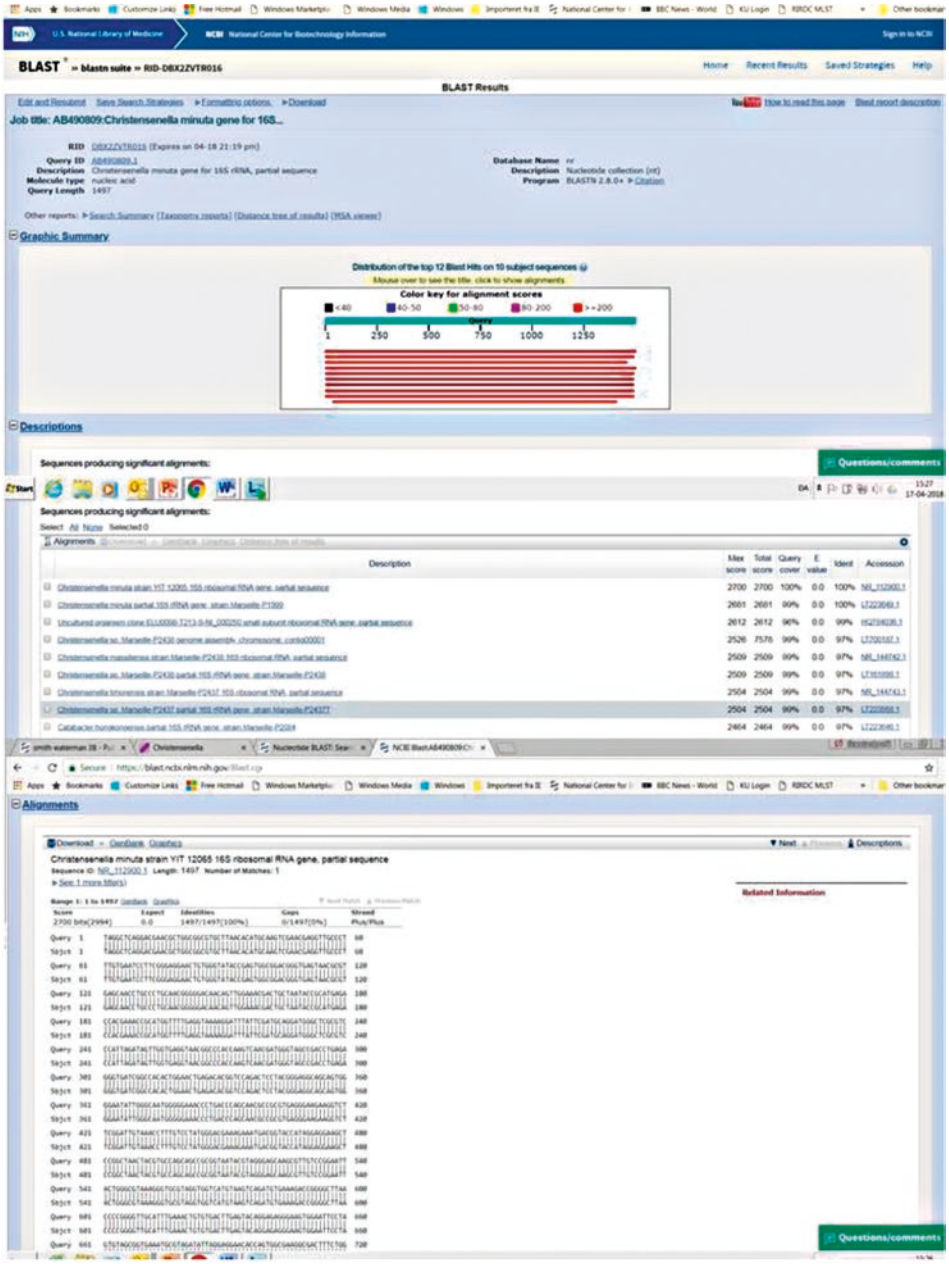


Fig. 4.24 BLAST (Altschul et al. 1990, 1997) output showing three screenshots of the output from BLAST search: **Graphic Summary**, **Description**, and **Alignments**, respectively. The number of Max target sequences was reduced to ten in the alignment parameters section to show it all on one page. (BLAST [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2018 04 17]. Available from: ► https://www.ncbi.nlm.nih.gov/Blast.cgi)

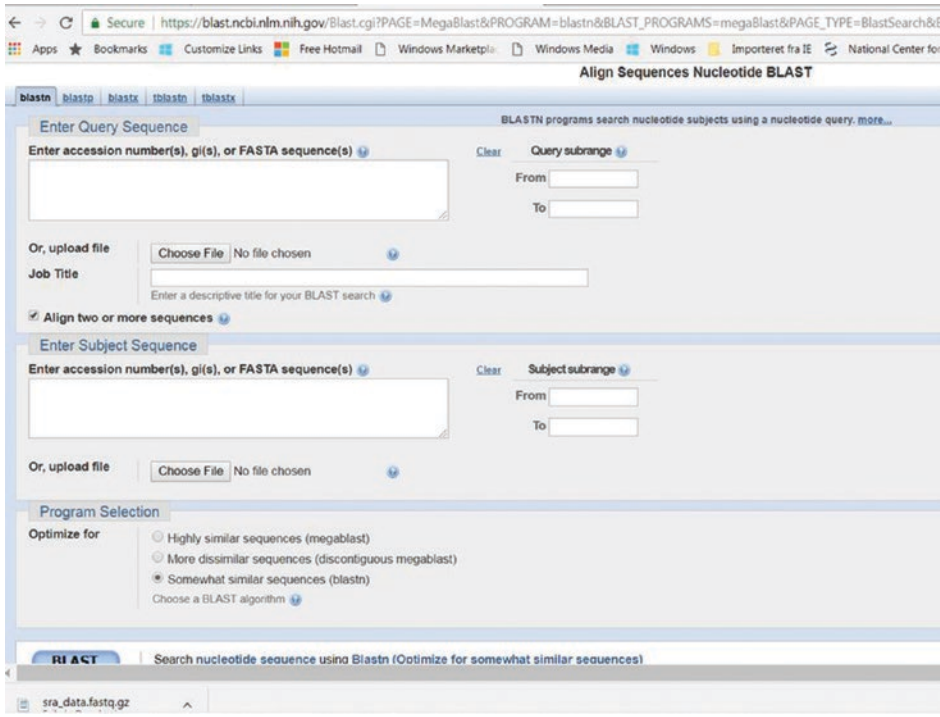


Fig. 4.25 BLAST2 sequences (Altschul et al. 1990, 1997). (BLAST [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2018 04 17]. Available from: ► https://www.ncbi.nlm.nih.gov/Blast.cgi)

the whole genomic sequences including genomes not yet deposited with the databases. The same approach can be used to build a small MLST database (► Chap. 11). BLAST2 sequences is started from the ordinary BLAST page NCBI by marking "Align two or more sequences." It can be used for fast draft comparisons between two sequences for example if you want to look up the location of a primer or gene on a genome. This is your "Swiss knife" always available if you just can access the Internet. Like for an ordinary knife, be careful with BLAST2 sequences! If you need real precise similarities between sequences, they should be obtained by the pairwise alignment programs described in ► Sect. 4.1.

4.3.4 Statistics

Evaluation of the results in the BLAST output is based on the Karlin and Altschul formula $E = k \times m \times N \times e^{-\lambda s}$ (Karlin and Altschul 1990) where m is letter in query meaning the number of nucleotides or amino acids, N is the total letters in database, and S is the actual score.

For amino acids, S is defined by the BLOSUM matrix. S is scaled to bit score to better fit the computer. The bit score = $((\lambda \times S) - \ln \kappa) / \ln 2$. In the pairwise section (Fig. 4.24 above), you can see the raw score in parenthesis along with the bit score,

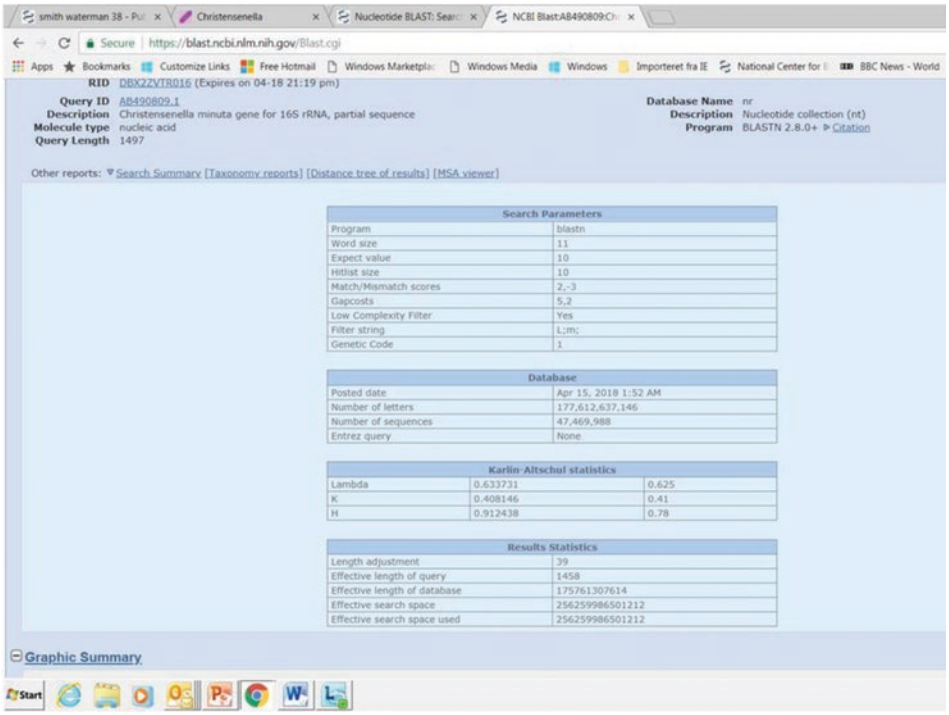


Fig. 4.26 The Search Summary section of the output showing all parameters used for the search (Altschul et al. 1990, 1997). (BLAST [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2018 04 17]. Available from: ► https://www.ncbi.nlm.nih.gov/Blast.cgi)

the constants κ and λ depending on the database. E reflects that we find the sequence in the database by chance (chance for false positives). E is this way related to the size of the database. The smaller E the less likely is it found by chance and the more unique is the sequence – "The smaller E the better." For really "unique" sequences, E is so small that it cannot be represented on the computer; this is the reason for "0.0" with the output.

All parameters related to the search can be found in **Search Summary** (Fig. 4.26); just press the bottom **Graphic summary** on the output page (Fig. 4.24). Here you can find parameters used for calculating the statistics in the Karlin and Altschul formula, and you can understand everything about the BLAST output if you compare these parameters to the statistics just explained.

Note that parameters for ungapped blast (Altschul et al. 1990) differ from gapped blast (Altschul et al. 1997) (Fig. 4.26) and λ and κ are different for gapped BLAST compared to ungapped.

4.3.5 Variants of BLAST

A BLAST search can be based on a query DNA sequence that the program translates to protein and compares to all protein sequence in the database (BLASTx). In this case the search takes more time, and it should only be done when it is absolutely needed (rarely).

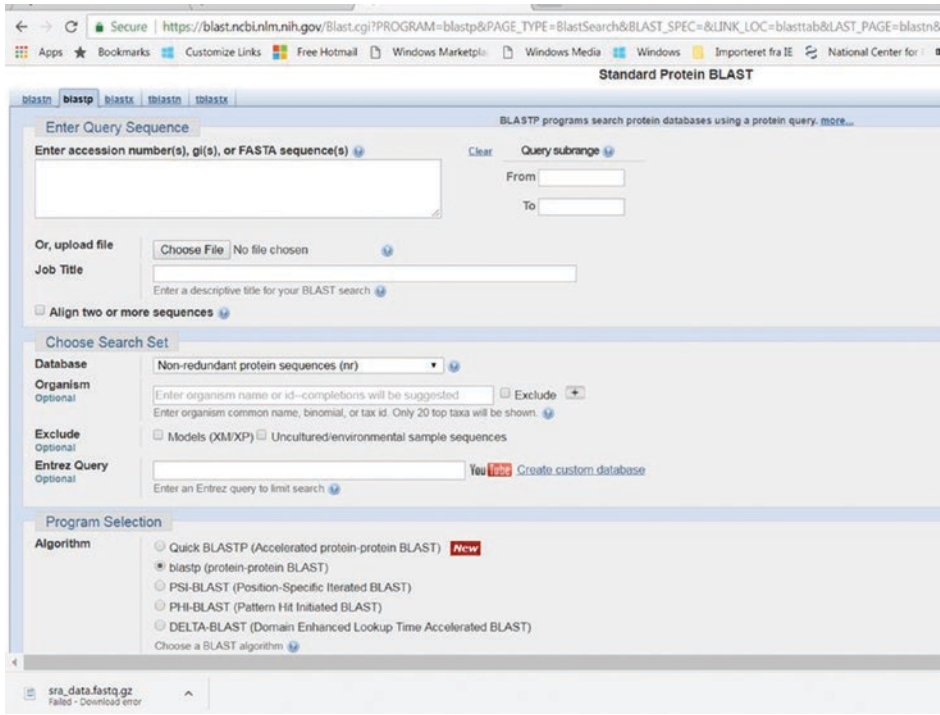


Fig. 4.27 Protein BLAST and BLASTp; note the additional options PSI- PHI- and delta BLAST (Altschul et al. 1990, 1997). (BLAST [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2018 04 17]. Available from: ► https://www.ncbi.nlm.nih.gov/Blast.cgi)

For such comparison, the codon table should be defined according to the type of organisms. Prokaryotes are using codon table 11 (Appendix). tBLASTn searches a translated nucleotide databases using a protein query. tBLASTx searches a translated nucleotide query against the translated nucleotide databases.

For BLASTp it is sometimes convenient to select Swiss-Prot or ref_seq_protein if you want a really precise information about the hits. For protein searches note the additional options PSI-, PHI-, and delta BLAST (Fig. 4.27); they can be used to build up a profile used to search for query proteins which are rather distantly related to the subject proteins in the database.

4.4 Activities

4.4.1 Pairwise Alignment

We will construct local and global pairwise alignments by use of the program "water" and "needle," respectively, available as EMBOSS programs (Rice et al. 2000) on the EBI server. The protein sequences AAA85484 and AAA85485 can be downloaded from NCBI as described in ► Chap. 3 (see ► Sect. 3.3.1 including Activity 3.8.1).

Decide if you want to use "needle" for global – or "water" for local alignment.

Open the FASTA format files, and paste them in the windows of the relevant EMBOSS program on the server:

- ► http://www.ebi.ac.uk/Tools/psa/emboss_needle/
- ► http://www.ebi.ac.uk/Tools/psa/emboss_water/

If you need to do comparisons of nucleotide sequences, the similar programs can be found here:

- ► http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html
- ► http://www.ebi.ac.uk/Tools/psa/emboss_water/nucleotide.html

You can also install the **mEMBOSS** package on your PC and select the programs water and needle from the menu. Download **mEMBOSS** from ► ftp://emboss.open-bio.org/pub/EMBOSS/windows/ Click on ► mEMBOSS-6.3.1.2-setup.exe

And download to local computer. Click on file and install.

Use **Explorer** from **Windows** to navigate the folder with **mEMBOSS**.

Go into the **Jemboss** folder and click **Jar** and **Jemboss MS**; this will allow you to run **EMBOSS** in **Windows** from a graphic interface. All commands to activate the programs are found on the left menu bar. Most useful are **water**, **transeq,** and **revseq** which will do pairwise Smiths and Water alignments, translate from DNA to protein, and reverse and complement sequences, respectively.

4.4.2 Learn How Dynamic Programming Works with Pairwise Alignments

Use the tool at: ► http://www.itu.dk/~sestoft/bsa/graphalign.html.

Use the short protein sequences shown already. Try the different substitution matrices, let the gap costs stay at "linear, gap score −8" to reduce the complexity, or make your own choice. Try with or without traceback. The tool has been designed and is maintained by Peter Sestoft.

4.4.3 Multiple Alignment with ClustalX

From ► ftp://ftp.ebi.ac.uk/pub/software/clustalw2/2.0.10/, download ► clustalx-2.0.10-win.msi, install, and locate the icon to the desktop. Open the program by double-clicking on the icon. **File | Load sequences,** and select the sequences in FASTA format from the location on your computer. You need to edit the input file yourself. The sequences that you want to use as input should be in one file only with all sequences in FASTA format:

```
>sequence1
ATGACGATAC...
>sequence2
GATAGATAGACS...
etc.
```

Select **Alignment | Do complete Alignment | ok**

In the lower left corner, you can follow how the program works.

Scroll through the alignment when it is completed.

At the bottom, the bars show the degree of conservation of columns. Above the alignment the signature * shows columns with conserved nucleotides (for amino acids: with the same properties).

4.4.4 BLAST

We will use this activity to learn how to perform an ordinary BLAST search in GenBank with a DNA sequence. However, we also learn some more about BLAST in this activity. We will perform a search with the sequence with acc. no. AF445297 from GenBank. The expectation is that this will be the best hit from the BLAST output. To a surprise it is not so.

First perform a BLAST search at NCBI: ► https://blast.ncbi.nlm.nih.gov/Blast.cgi. On the graphics select **nucleotide BLAST,** insert the acc. no. AF445297 in the window, mark "**somewhat similar sequences blastn,**" mark "**show results in a new window,**" and press the blue **BLAST** bottom. You will get a nice output and 100% identity to a sequence from an uncultured organism. Where is your query sequence? Usually you would expect the query as the best hit if it comes from the database. The reason can be found if you look up AF445297 at NCBI, since it is labeled as unverified. The unverified sequences are not included as databases in any BLAST search.

Take-Home Messages

- Pairwise alignments and multiple alignments are the basic tools for sequence comparison.
- A quantitative pairwise comparison of two sequences can be performed by aligning two sequences based on considerations of gaps representing insertions or deletions and matches between nucleotides or amino acids.
- Pairwise comparisons can be performed as global alignments if it is known that the sequences are homologous in their full length or by local alignments if it is known that one sequence is shorter than the other.
- BLAST is the most frequently used bioinformatics program to compare your own sequence (query sequence) to all sequences in a database (subject sequences).
- BLAST provides qualitative information about homologous sequences and can quantify the identity of the query sequence to the subject sequences in the database.
- Substitution matrices provide the probability of nucleotide or amino acid substitutions when two or more sequences are compared.
- Dynamic programming is the most frequently used procedure to perform pairwise comparisons of nucleotide or amino acid sequences and can be done based on the Needleman and Wunsch algorithm for global alignments and by the Smiths and Waterman algorithm for local alignments.
- Multiple alignments are most frequently constructed by the progressive profile procedure as implemented in the clustal family of programs.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. 1999. Biological sequence analysis. Cambridge Univ. Press.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797

Feng, D. F. & Dolittle, R. F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25, 351–60.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 1999;41:95–98.

Higgins, D. G. & Sharp, P. M. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237–244.

Karlin, S., & Altschul, S. F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87, 2264–8.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., Higgins, D. G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A. W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüssmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K. H. 2004. ARB: a software environment for sequence data. Nucleic Acids Res. 32, 1363–1371.

Moreno-Hagelsieb, G. & Latimer, K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24:319–24.

Needleman, S. B. & Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

Nei, M. & Kumar, S. 2000. Molecular evolution and phylogenetics. Oxford University Press.

Notredame, C., Higgins, D.G., Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302: 205–217.

Rice, P., Longden, I., & Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genetics 16:276–277.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D. & Higgins, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using ClustalOmega. Mol. Syst. Biol. 7:539.

Smith, T. F., Waterman, M. S. & Fitch, W. M. 1981 Comparative biosequence metrics. J. Mol. Evol. 18, 38–46.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 2011; 28: 2731–2739.

Thompson, J. D., Plewniak, F. & Poch, O. 1999. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15:87–88.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., Barton, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191.

Yamada, K.D., Tomii, K., Katoh, K. 2016. Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. Bioinformatics32, 3246–3251.

Zuker, M. & Jacobson, A. B. 1998. Using Reliability Information to Annotate RNA Secondary Structures. *RNA* 4, 669–679, 1998.

Primer Design

Design of Oligonucleotide PCR Primers and Hybridization Probes

Henrik Christensen and John Elmerdahl Olsen

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_5

What You Will Learn in This Chapter

You will learn to identify different situations where primer design is necessary and to separate exploratory applications from diagnostic applications. In some situations de novo design of primers is needed; in others it will be relevant to test primers already published. You will learn about the basic rules for primer design and how to evaluate sequence differences and T_m, and you will be presented with a list of primer design programs. In the activity you will train primer design based on the single DNA sequence (exploratory) as well as design for PCR primers for diagnostic purpose.

5.1 Background for Oligonucleotide Design

Computer programs are now available allowing users with a limited background knowledge to design oligonucleotides. Most programs are available for free and easy to use from convenient web interfaces. A minimum theoretical background is needed to obtain meaningful results from such analysis. With a limited investment in time including an overview of the different programs available, users can obtain much better performance even without the need to go into details of computer science and algorithms of programs. The aim of the current chapter is to provide an updated overview of the design of oligonucleotides within a bioinformatical framework including presentation of relevant software (Fig. 5.1). The in silico design of oligonucleotides can lead to tremendous savings of time and money in the development of oligonucleotide tests; however, it can never replace experimental verification since it is not possible to fully predict oligonucleotide probe specificity in silico.

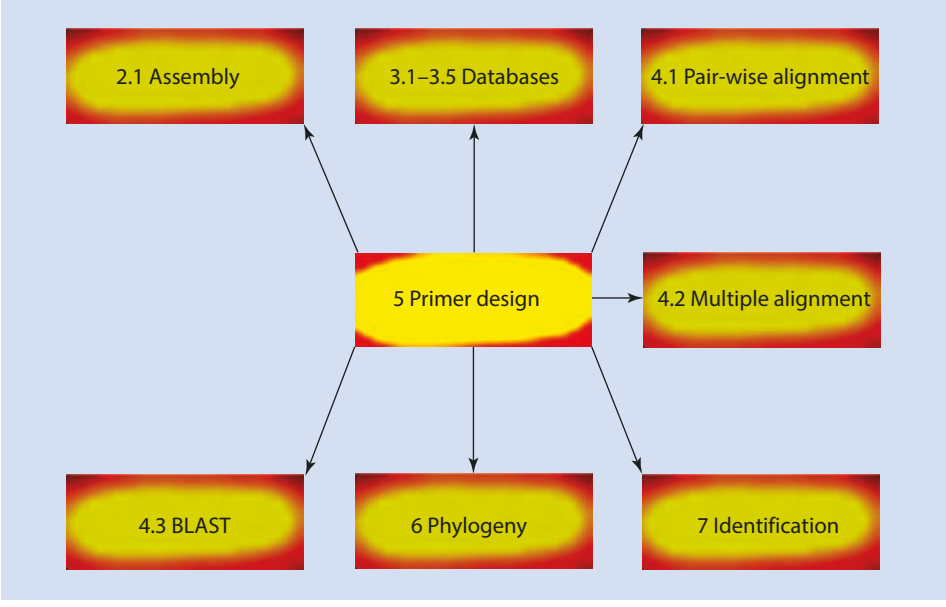


Fig. 5.1 Relation of this chapter to the other chapters in the book. Primer design relates to the generation of sequence, to the multiple alignment of sequences, as well as to databases, identification, and BLAST

Oligonucleotides are in short referred to as "probes" when used for hybridization and "primers" when used for PCR. The design of oligonucleotides will be described with reference to both primers and probes since most researchers use both PCR and hybridization technology. Demands to both probes and primers are formation of stable duplexes with their target sequences and low self-complementarity. In other respects the design of oligonucleotides is different depending on function either in hybridization or PCR, respectively. For many applications, probes and primers have already been designed and are available from publications and databases, for example, accumulated with a database of virus-specific oligonucleotides (▶ http://viroligo.okstate.edu) (Onodera and Melcher 2002). NCBI probe can be searched together with other NCBI databases (▶ http://www.ncbi.nlm.nih.gov/probe) (NCBI 2016).

5.1.1 Practical Approach to Oligonucleotide Design Whether of Exploratory Nature or for Diagnostic Purpose

The use of oligonucleotides can broadly be identified in regard to exploratory projects such as the design of PCR primers to amplify DNA based on a single DNA sequence or diagnostic purpose where the specific amplification of only one or a group of sequences is intended, whereas others are not wanted. For most research projects involving the design of oligonucleotide probes, the work is performed in five steps (◘ Table 5.1). First a database search is performed with respect to target gene(s) and organism(s). The DNA sequences for design of oligonucleotides are mostly downloaded from the International Nucleotide Sequence Database Collaboration (GenBank/EBI/DDBJ; Benson et al. 2018; Karsch-Mizrachi et al. 2018; Kodama et al. 2018; NCBI 2018). For instance, GenBank is searched with BLAST (Altschul et al. 1997). Sequences of importance can be downloaded as described in ▶ Chap. 3. The second step is to compare the selected sequences by multiple alignments and phylogenetic analysis. As a third step, it needs to be considered if more DNA sequences need to be generated to cover the target sequences being investigated. First at step 4, the real sequence comparison allows search for oligonucleotides with the

◘ **Table 5.1** Strategies for the design of oligonucleotide PCR primers and hybridization probes

Task	Action
1. Identification of target sequences and literature	Tables/files of information are generated based on database search (▶ Chap. 3)
2. Identification of conserved and homologous regions	Multiple alignments or/and phylogenetic analysis (▶ Chaps. 4 and 6)
3. Generation of sequence	Experimental sequencing (▶ Chap. 2)
4. Identification of oligonucleotide candidates	Primer search programs (◘ Table 5.3)
5. Verification	Experimental tests including positive and negative controls

software to be described below (▶ Sect. 5.7). The oligonucleotides can further be tested against the database(s) by BLAST to confirm target sequences. The final step is to set up and try the oligonucleotides for real.

It is important to be able to design new probes and primers for three reasons:

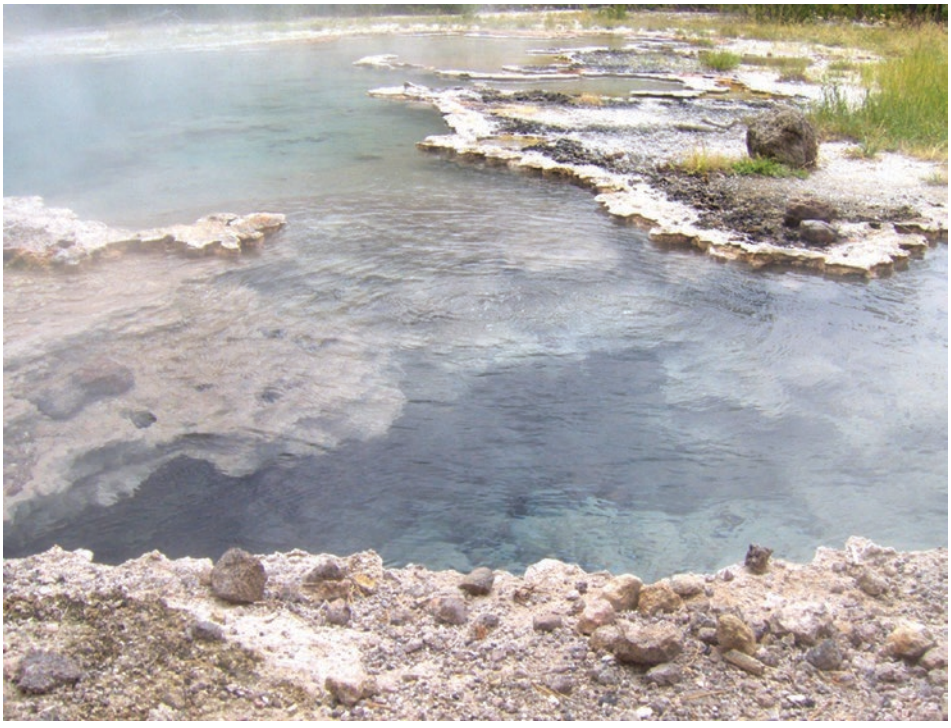
1. New diagnostic problems are faced.
2. Improved knowledge of the organisms including new varieties and genes need to be accounted for.
3. Natural genetic variation occurs in the organisms under investigation especially virus.

The main function of oligonucleotides is to "prime" PCR reactions in the way that they hybridize to single-stranded DNA and prime the replication of DNA by the action of Taq polymerase (◘ Fig. 5.2).

The other main function is with in situ hybridization. Hybridization is also used for microarrays; however, in this case, oligonucleotide probes are usually fixed to a solid support and used unlabeled, and the test DNA is labeled.

5.1.1.1 Exploratory Applications

Exploratory applications include procedures to amplify DNA for sequencing only requiring PCR primers of low specificity and similar tools to manipulate DNA such as PCR primers used for cloning. PCR primers for use with multilocus sequence typing of bacteria (▶ Chap. 11) are also in this category.



◘ **Fig. 5.2** Hot well from Yellowstone National Park, the harsh environment where *Thermus aquaticus* was isolated. From this bacterium the thermostable polymerase was isolated that is essential to perform PCR

Degenerate primers and probes include different nucleotides at some positions of the sequence. Degenerate primers and probes are needed if the target DNA sequences include an unknown diversity, for instance, if the protein sequence is known and then the underlying DNA sequence is known to include codon ambiguities.

The exploratory applications also include PCR primers for cloning with recognition sites for restriction enzymes. Other PCR primers are used to "prime" the reverse transcription of mRNA and viral RNA to DNA.

5.1.1.2 Diagnostics Applications

Oligonucleotides designed for this application are typically used for diagnostics of microbial pathogens of animals and plants where high specificity is needed. Oligonucleotide probes are used for detection of specific DNA target sequences by hybridization when labeled by radioactivity, fluorescence, or other handles facilitating detection.

PCR detection is achieved by overproduction of target DNA sequence, in modern diagnostic assays often combined with fluorescent detection achieved in real-time (RT)-PCR. RT-PCR is not different to ordinary PCR when it comes to oligonucleotide design except for the various types of detection chemistries (Sharkey et al. 2004; Wong and Medrano 2005).

The primary goal of PCR primer design for diagnostic purpose has been formulated to obtain a balance between specificity of amplification and efficiency of amplification (Dieffenbach et al. 1995). Specificity is the tendency for a primer to hybridize to its intended target and not to other targets. In other words, specificity is defined as absence of reaction with negative targets. The same principle exists for hybridization. With PCR, the efficiency is the ease that the expected product is produced. Often a compromise has to be made between specificity and efficiency. For clinical applications, "bedside" specificity might be increased at the cost of efficiency to avoid false-positive results that might result in the wrong treatment of patients (Dieffenbach et al. 1995; Hyndman and Mitsuhashi 2003; Grunenwald 2003).

5.2 General Rules for Design of Oligonucleotides

Both PCR and hybridization techniques are based on the hybridization of an oligonucleotide to a complementary target DNA sequence. Hybridization is the formation in vitro of specific double-stranded nucleic acid molecules from two complementary single-stranded molecules under defined physical and chemical conditions. Figure 5.3 shows the orientation of PCR primers and hybridization probes with respect to targeted double-stranded DNA. As a rule of thumb, the forward primer is always oriented as the sense DNA strand, whereas the reverse primer is both complementary and reverse to the sense DNA strand. Probes are oriented as the antisense DNA strand.

With four possibilities (A, C, G, T) for selection of nucleotides at each position of the sequence, 4^{18} ($\sim 10^{11}$) possibilities exists for the design of an 18 nucleotide oligonucleotide. For this reason, very stringent criteria are needed to select the best oligonucleotide for a specific purpose. However, hybridization, PCR amplification, and probe binding are also possible with a few mismatches under some conditions, and it is not only the sequence but also the conditions that define the outcome.

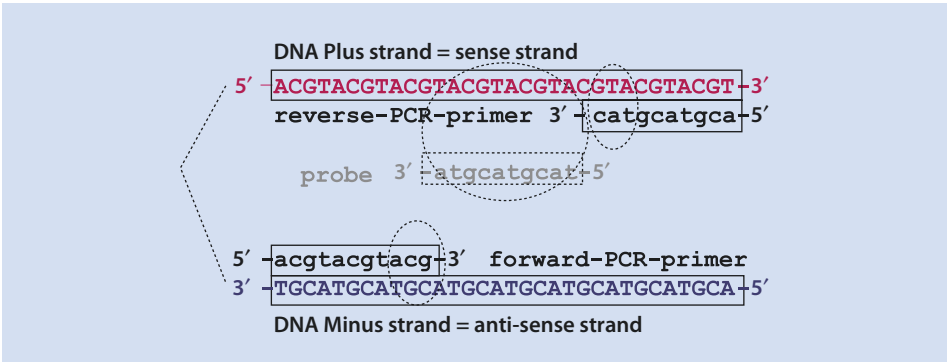


Fig. 5.3 Sketch of orientation of PCR oligonucleotide primer and hybridization probes in relation to DNA target. For PCR, the primer for forward amplification has the same sequence as the sense strand of template DNA and binds to the antisense DNA strand. The primer for reverse PCR amplification has a sequence that is both reverse and complementary to the sense DNA strains. Probes for hybridization are oriented the same way as the antisense DNA strand

The ability of an oligonucleotide to serve as a PCR primer is dependent on the kinetics of:

1. Association and dissociation of probe- or primer-template duplexes at the annealing and extension temperatures
2. The effects on duplex stability of mismatched bases and their location
3. The efficiency that the polymerase can recognize and extend a mismatched duplex

The ability of an oligonucleotide to serve as a hybridization probe depends on the kinetics of:

1. Association and dissociation of probe-template duplexes at the hybridization temperature
2. The effects on duplex stability of mismatched bases and their location

In conclusion, for the design of oligonucleotides, the sequence comparison between oligonucleotide and template is most important. It is also important to calculate T_m (▶ Sect. 5.4) to estimate the annealing temperature in PCR and hybridizations since the actual hybridization temperature in the end will determine the outcome of the experiment. With multiplex PCR many probes and primers, respectively, have to function together at the same temperature, and for this reason, they need to be designed within a common threshold level of T_m.

5.2.1 Lengths of PCR Primers and Products

PCR primers should normally be between 18 and 24 nt (Table 5.2). The length depends on the GC content. For primers of the same length, the T_m will increase with higher GC content related to the three hydrogen bounds in the GC pair compared to only two in the AT pair. To match the T_m of both forward and reverse primers, the length of a primer with high GC content needs to be shorter compared to one with less GCs. PCR primers of 15 nucleotides or shorter are only used for arbitrary or random short priming in the mapping of simple genomes since their sequence is so short that the change that they will be complementary to a DNA sequence in a genome is very high. For ordinary PCR, product

Table 5.2 Rules of thumb for design of oligonucleotide PCR primers and hybridization probes

Application	Rules
All	Purines and pyrimidines should be equally distributed and long stretches of identical nucleotides avoided. The GC content should be similar to or higher than the target DNA sequence. The medium length is 18 nt. The hybridization temperature including annealing temperature should be 5 °C lower than T_m
PCR primers	Primer lengths of 18 and up to 24 nt are preferred with T_m of 54 °C or higher. Perfect base pairing between the 3′ end of the primer and the template is necessary for maximal specificity, and the last 5–6 nucleotides at the 3′ end of the primer must contain minimal mismatches
Hybridization probes	At least one or two nucleotide differences are needed for separation of target from nontarget sequences. Mismatches near the end of probes are less destabilizing than internal mismatches For design of oligonucleotide probes targeting rRNA, probes should bind along one or the other side of hairpins

ranges from approx. 200 to 1500 bp are preferred to allow resolution on ordinary agarose gels. For real-time PCR (RT-PCR), products are usually shorter (<150 bp). Guidelines for RT-PCR are found in Huggett et al. (2013).

5.2.2 Lengths of Oligonucleotide Hybridization Probes

Hybridization probes have been applied with lengths from around 20 nt up to several hundreds of nt. For microarray experiments, the usual size is 25 nt, but probes of 50–70 nt are also used on certain applications (Gibson and Muse 2004). Generally, the longer the probe, the more mismatches are needed to separate target from nontarget sequences.

5.3 Sequence Comparison

For comparison of sequence matches between primer and template, sequence comparison can be performed in two fundamentally different ways. Either sequences can be compared as so-called strings or evaluated in relation to scores between nucleotide pairs (▶ Sect. 5.3.1). This procedure was shown for local pair-wise sequence comparison (▶ Chap. 4). Alternatively, sequences can be compared thermodynamically in relation to their changes in free energy upon forming duplexes; however, this method is only useful for shorter oligonucleotides (▶ Sect. 5.3.2).

5.3.1 String Comparison by Score

Sequences are matched pair-wise by alignments and alignments compared by scores. Scores are defined in favor of canonical base pairing (G-C, A-T) and the alignment with the highest score chosen. The method is by intuition simple in that complementary nucleotide pairs are favored in the comparison. For pair-wise comparison between oligonucleotide and template, precise algorithms are used, whereas for database comparisons, less precise but greedier algorithms are implemented in programs like BLAST.

5.3.2 Nearest Neighbor Comparisons of Duplex Stability

The nearest neighbor calculation is based on the fact that the change in Gibbs free energy (ΔG) of the probe or probe-template complex can be calculated from the enthalpy (ΔH) and entropy (ΔS) at a given temperature (T) as $\Delta G = \Delta H - T\Delta S$. If ΔG for the oligonucleotide binding to itself is negative, the primer should be redesigned by extension/removal of 5′ or 3′ positions in order to reduce self-complementarity. A nearest neighbor pair is defined by two neighboring nucleotides in a sequence. For example, in the probe sequence 5'-ACGT-3′, pairs to be considered for binding to a complementary sequence as a template will be A-C, C-G, and G-T. The calculation can either be used for selection of oligonucleotides by comparison of all nearest neighbor pairs formed between probe and template or for evaluation of probe self-complementarity including loops and for calculation of T_m (▶ Sect. 5.4.3). Considerations are taken to free energy change in relation to helix initiation associated with forming the first base pair in the duplex, the sum of free energies for the subsequent pairs, and correction for self-complementary (Owczarzy et al. 2008). The parameters ΔG, ΔH, and ΔS were empirically determined for all nearest neighbors of DNA by Breslauer et al. (1986) and further refinements performed as reviewed by SantaLucia (1998). These parameters are only used for implementation in computer programs.

5.3.3 Design of Primers for PCR and "Kwok's Rules"

In 1990, Kwok et al. published a set of rules for the design of PCR primers based on their experience with PCR amplification of HIV sequences. The principles determined have found general use known as "Kwok's" rules. The discrimination of different PCR targets is related to the lack of 3′ to 5′ exonuclease activity of Taq DNA polymerase resulting in extension of mismatched 3′ termini at a lower rate compared to complementary termini (Lawyer et al. 1989).

The conclusion of the work of Kwok et al. (1990) was that extension is most dependent on the outermost 3′ base pairing, less on second and third last pair and even less on the other pairs. ◘ Figure 5.4 summarizes some conclusions from the Kwok paper. The experiments with PCR were performed at 55 °C annealing temperature and are therefore comparable to most PCR conditions. The conclusion is that if the aim of the design of primers is to obtain specific detection of certain target sequences without amplification of other nontargets, G-C-pairs in 3′ end will in most cases allow discrimination but A-T pairs will not. It might come as a surprise that amplification also is possible from noncanonical pairs, for example, G-T. G-T pairs should be avoided in the 3′ end since they are more stable than other types of mismatches (Kwok et al. 1995) (see also Newton et al. 1989).

5.3.4 Design of Probes for Hybridization

In parallel to the rules just outlined for PCR, it is also important to account for different sorts of noncanonical base pairing in hybridization experiments. For DNA oligonucleotide hybridization, the pairs G-T and G-A are involved in weak binding, and the pairs G-A, G-G, C-A, A-A, G-T, T-T, and C-T are destabilizing in decreasing order (Ikuta et al. 1987; Pozhitkov et al. 2006) (◘ Fig. 5.5). In correspondence, the stability of RNA/DNA duplexes on microarrays are TG, TU, TC > GU, AC, CC, CU, AA, AG > CA > GG, and GA (Pozhitkov et al. 2006). These observed interactions are mostly a consequence of pyrimidine-pyrimidine mismatches being more stable than purine-purine mismatches

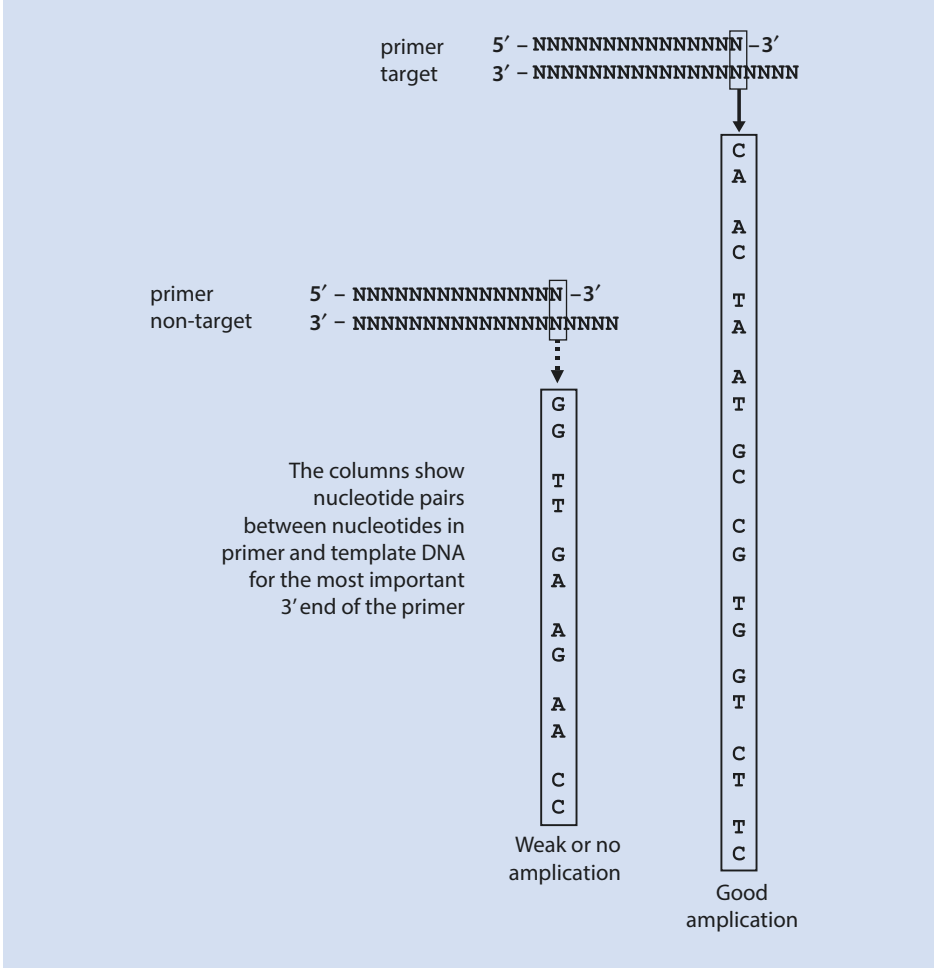


Fig. 5.4 Prediction of amplification from interactions between PCR oligonucleotide primer at the 3′ end and template (based on Kwok et al. 1990). Low degree of mismatch is assumed between oligonucleotide and template for the remaining part of sequence

related to higher steric hindrance for binding of the latter (Pozhitkov et al. 2006). In addition, the G-U pair is stable in RNA compared to the more unstable G-T pair in DNA. For microarrays, oligonucleotides bound at the 3′ end, mismatches nearest the 5′ end are less destabilizing than those in the middle of the oligonucleotide (Szostak et al. 1979; Urakawa et al. 2002; Pozhitkov et al. 2006).

5.4 T_m Calculations

The nucleic acid melting temperature, T_m, is defined as the midpoint in the transition from helix to random coil measured as change in optical density. The original definition implied almost equal proportions of complementary strands. For oligonucleotides, concentrations are orders of magnitude higher than template concentration, and the actual dissociation

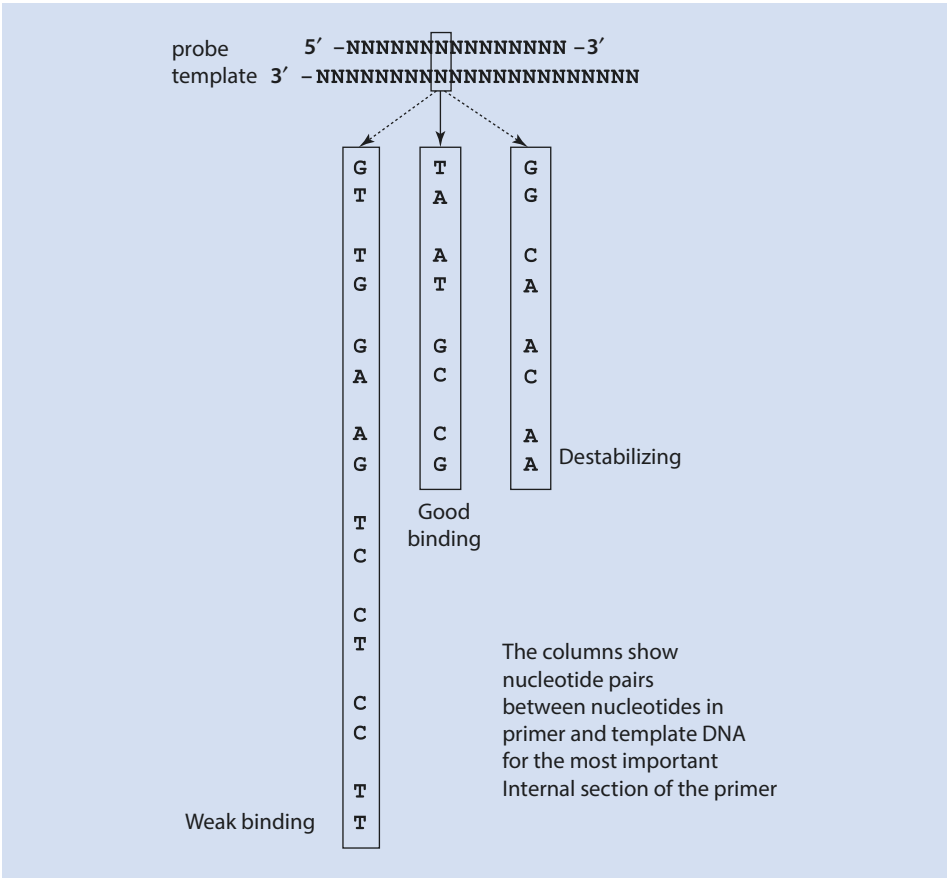


Fig. 5.5 Prediction of binding in hybridization between probe and template. A low degree of mismatch formation is assumed between oligonucleotide and template for the remaining part of the sequence

temperature, T_d, where 50% of oligonucleotides are bound to template out of the maximum possible, might be different from T_m and depend on the concentration of the DNA strands. The thermodynamic approach can also be used for calculation of T_m (▶ Sect. 5.4.3). Alternatively, T_m is determined by a formula (▶ Sect. 5.4.1).

5.4.1 Estimation of T_m by Formula

The crudest way to estimate T_m is used with primers shorter than 20 nt. The formula $T_m = 4$ (G + C + 2(A-T) can be used (Suggs et al. 1981). However, the nearest neighbor method (see below) was found four times more accurate (Rychlik and Rhoads 1989). An empiric derived formula for calculation of T_m was given by Schildkraut (1965) for unsheared DNA fragments. This formula has been modified for short DNA fragments. For oligonucleotides of 10–50 nt, $T_m = 81.5 + 16.6\log M + 0.41(G + C\%) - (820/(\text{length of probe})$ (Lathe 1985). For probes longer than 50 nt, $T_m = 81.5\ °C + 16.6\log M + 0.41(G + C\%) - (500/$

length of probe) (Meinkoth and Wahl 1984). In these formulas, M is the concentration of monovalent cations (N^+ and K^+). Typical values for PCR are 50 mM KCl where K^+ is by far the dominant cation. For 1 × SSC, the concentration of monovalent catios is 195 mM.

5.4.2 Formamide Considerations

For practical purpose formamide can be included during hybridization to decrease the actual temperature required. It is possible to adjust T_m by application of a gradient of formamide concentrations rather than to use a set of different temperatures during incubation. Another reason to include formamide is that temperatures higher than 50–60 °C might damage tissue samples with in situ hybridization as well as reagents, materials, and equipment. The common rule has been that 1% formamide reduces T_m by 0.72 °C (McConaughy et al. 1969). However, recently it was found that 1% formamide reduces T_m by only 0.6 °C in oligonucleotide microarrays (Urakawa et al. 2002). Unfortunately health hazards associated with formamide has limited its use.

5.4.3 Estimation of T_m by Nearest Neighbor Prediction

The composition of the nucleotide sequence has been found to influence T_m more than either length of oligonucleotide or base composition. This is the basis for the use of the nearest neighbor models for calculations of T_m. As mentioned, nearest neighbor calculations are based on the former mentioned stability of base pairs only depending on the immediate up- and downstream neighbors. As an example of such calculation used in the computer programs, $T_d = \Delta H/(\Delta S + R\ln C) + 16.6\log 10\ [M]$, where $T_m = T_d$ for PCR and T_m is less than 7.6 °C of T_d for filter hybridization (Rychlik and Rhoads 1989). R is the universal gas constant (1.99 cal K^{-1} mol^{-1}). C is the molar concentration of all strands when oligonucleotides are self-complementary and is replaced by C/4 for noncomplementary oligo-template interactions (Borer et al. 1974).

Formula and nearest neighbor calculations have been combined. The combination was used by Rychlik et al. (1990) for calculation of annealing temperatures (T_a). T_a was calculated as $T_a = 0.3\ T_m$ primer $+0.7\ T_m$ product $- 14.9$. T_m primer was calculated by the nearest neighbor model and T_m of product by formula. The calculation of T_a was compared to empirically determined T_a in PCR reactions, and they were found to be in good agreement.

5.5 Special Applications

5.5.1 Exploratory Applications

5.5.1.1 Degenerate Primers and Probes

If the purpose is to detect homologous sequences of a particular type, a high degree of variation often must be tolerated, and for this purpose degenerate oligonucleotides are designed. A primer sequence is degenerate if some of its positions have several possible bases. The degeneracy of the primer is the number of unique sequence combination it contains. For example, the degeneracy of the primer GGSABA is six where S means "strong" and either G or C and B can be either C, G, or T resulting in 1 × 1 × 2 × 1 × 3 × 1 = 6 possibilities (GGGACA, GGGAGA, GGGATA, GGCACA, GGCAGA, GGCATA). **The IUB**

codes are shown in the Appendix. Two strategies exist for designing degenerate primers and probes. Either two or more different nucleotides can be incorporated at specific positions of the oligonucleotide at the time of synthesis, or different batches of oligonucleotide primers can be mixed after synthesis. A third possibility is to combine the two strategies. Inosine can be incorporated as an "inert" nucleotide forming base pairs with all four nucleotides. Degeneracies are usually incorporated to account for codon variation at second and third positions. Knowledge of codon bias can be used to limit the degeneracy needed (Kwok et al. 1995). A procedure for this strategy is described with the CODEHOP program (■ Table 5.3). A consequence of the introduction of ambiguous nucleotides is that the actual concentration of the important nucleotides at the 3′ end will decrease. Efficient PCR amplification (detectable on conventional agarose gel) will probably not work with less than 0.1 μM final concentration of primer, and, for example, four ambiguous positions will decrease the actual concentration with 2 × 2 × 2 × 2 = 32 which then requires 3.2 μM for each of the primers to work.

5.5.1.2 Nested PCRs

In nested PCR, the sample is first amplified with an "outer" primer set and a subsample of this product then amplified with the "inner" primer set. The inner primer set is designed to be the specific one. In designing primers for this purpose, primer-dimer formation between outer and inner pairs should be tested for and avoided (primer-dimer formation is the ability of two oligonucleotides to base pair often allowing initiation of short PCR products). The product obtained with the inner pair should be short (<300 bp). The method can be used when the concentration of template is low or the knowledge of the target sequence is limited and amplification cannot be achieved in other ways (Dieffenbach et al. 1995). Programs for the design of nested PCRs include Primer Premier (■ Table 5.3).

5.5.1.3 Primers for Cloning

For molecular cloning, restriction sites are needed in the PCR primers. The program ORFprimer (■ Table 5.3) should be able to handle this application.

5.5.2 Diagnostic Applications

5.5.2.1 Primers for Multiplex PCR

Multiplex PCR is preferred in order to conserve reagents (template) as well as reduce preparation and analysis time compared to conventional PCR. With mixtures of primers with different sequences, the prediction of T_m becomes difficult since one value might only apply to each of the primers included. The programs FastPCR, MuPlex, and Primer Premier (■ Table 5.3) should be able to handle the design of multiplex PCR primers. Experimental testing will be needed to confirm the choice of T_m.

5.5.2.2 SNP Analysis

Single nucleotide polymorphisms (SNP) are differences between individual nucleotides within the same gene which may affect virulence or other properties between isolates of microorganisms. PCRs can be designed to target SNPs. For this the SNPs need to be located in the 3′ of the primer binding to the template. Some programs that can be used to identify SNPs based on PCR tests are included in ■ Table 5.3 (MuPlex, PIRA-PCR, and PrimerZ). In order to increase specificity, mismatches may be introduced according to Newton et al. (1989).

Table 5.3 Programs listed according to application

Program	Description	URL and reference
General purpose		
Amplifx1.7.0	PCR design program for installation on MacOS X and Windows	► http://crn2m.univ-mrs.fr/recherche/equipe-t-brue/jullien-nicolas/programmation/amplifx/?lang=en
AutoDimer	Screen PCR primers for primer-dimer and hairpin	► https://strbase.nist.gov//AutoDimer-Homepage/AutoDimerProgramHomepage.htm
FastPCR	PCR primer design including multiplex PCR	► http://primerdigital.com/fastpcr.html
Netprimer	Numerous parameters of single PCR primers	► http://www.premierbiosoft.com/netprimer/index.html
Oligo	Nearest neighbor calculations of secondary structures and T_m	► http://www.oligo.net/ (Rychlik and Rhoads 1989)
OligoAnalyzer	Nearest neighbor parameters. Hairpin and primer-dimer analysis	► https://eu.idtdna.com/pages/tools (Owczarzy et al. 2008)
OligoCalc	Physical properties of oligonucleotides, self-complementarity, and hairpin loop formation	► http://biotools.nubic.northwestern.edu/OligoCalc.html (Kibbe 2007)
OligoFaktory	DNA microarrays, primers for PCR, siRNAs	► http://www.bioinformatics.org/oligofaktory/
ORFprimer	Large datasets, restriction sites	► http://www.proteinstrukturfabrik.de/ORFprimer/ (Büssow et al. 2002)
Pride and Genome pride	PCR and microarray	► http://pride.molgen.mpg.de/ (Staden package)
PriFi	Search for primers in multiple alignments	► https://services.birc.au.dk/prifi/ (Fredslund et al. 2005)
Primaclade	PCR primers in multiple nucleotide alignment file	► http://primaclade.org/ (Gadberry et al. 2005)
Primegens	Primer design	► http://primegens.org/(Xu et al. 2002)
Primer3	The most frequently used program see also Activity 5.8.1	► http://bioinfo.ut.ee/primer3-0.4.0/primer3/input.htm (Rozen and Skaletsky 2000)
Primer3Plus	Simplified version of Primer3	► http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/
Primer Design Assistant (PDA)	Primer design with all parameters	► http://dbb.nhri.org.tw/primer/ (Chen et al. 2003)
Primer Premier	PCR primer design, multiplex, degeneracies, and nested PCRs	► http://www.premierbiosoft.com/primerdesign/index.html
PrimerQuest	PCR and hybridization probes. Based on Primer3	► https://eu.idtdna.com/PrimerQuest/Home/Index

Table 5.3 (continued)

Program	Description	URL and reference
Primersearch (EMBOSS)	Matches between primers-pairs and DNA template by string comparisons	► http://emboss.sourceforge.net/ (Rice et al. 2000)
Primo Oligo	Calculation of T_m	► http://www.changbioscience.com/primo/oligo.html
Web Primer	PCR primer design	► https://www.yeastgenome.org/cgi-bin/web-primer
SciTools	A series of tools	► http://www.idtdna.com/SciTools/Scitools.aspx (Owczarzy et al. 2008)
T_m calculator	For primer pairs	► http://www6.appliedbiosystems.com/support/techtools/calc/
Special applications		
ARB	Design of rRNA-targeted probes	► http://www.arb-home.de (Ludwig et al. 2004)
Array Designer	Oligo-arrays, cDNA arrays, and SNP arrays	► http://www.premierbiosoft.com/dnamicroarray/index.html
Assembly PCR oligo maker	PCR-based construction of long DNA molecules for RNA molecules by T7 RNA polymerase	► http://www.yorku.ca/pjohnson/AssemblyPCRoligomaker.html
AutoPrime	Design of RT-PCR oligos to specific mRNA amplification of genomic sequences	► http://www.autoprime.de/AutoPrimeWeb
Beacon designer™	Real-time PCR	► http://www.premierbiosoft.com/molecular_beacons/index.html
BLOCKMAKER and CODEHOP	Degenerate primers to genes of proteins	► https://virology.uvic.ca/virology-ca-tools/j-codehop/ (Rose et al. 2003)
ddRNAi	DNA-directed RNA interference (ddRNAi)	► http://www.geocities.ws/zimzwy/ddRNAi.html
Expeditor	QTL design	► https://www.animalgenome.org/cgi-bin/expeditor/expeditor2
Genefisher2	Degenerate PCR primers based on multiple aligned sequences	► https://bibiserv.cebitec.uni-bielefeld.de/genefisher2/
Gene2Oligo	Design of genes in vitro	► http://berry.engin.umich.edu/gene2oligo/ (Rouillard et al. 2003)
MFOLD	Evaluates probes	► http://unafold.rna.albany.edu/?q=mfold
MOP-UP	Targeting one group of sequences and excluding another group	► http://www.hpa-bioinfotools.org.uk/cgi-bin/mopUP.cgi
MuPlex	Multiplex PCR assays for high-throughput genotyping including SNP	► http://cagt.bu.edu/page/MuPlex_about (Rachlin et al. 2005)

(continued)

Table 5.3 (continued)

Program	Description	URL and reference
OligoArray2	Oligonucleotides for microarray experiments	► http://berry.engin.umich.edu/oligoarray2/ (Rouillard et al. 2003)
OligoPicker	Microarray design of oligonucleotides	► https://pga.mgh.harvard.edu/oligopicker/index.html
PIRA-PCR	SNP	► http://primer1.soton.ac.uk/primer2.html (Ke et al. 2001)
PrimerD	Degenerate primer pairs	► http://mblab.wustl.edu/software.html#primerdLink
Primer Explorer	LAMP primers	► http://primerexplorer.jp/e/
PrimerX	Site-directed mutagenesis	► http://www.bioinformatics.org/primerx/cgi-bin/DNA_1.cgi
PrimerZ	Promoters, exons, and human SNP	► https://omictools.com/primerz-tool (Tsai et al. 2007)
ProbeSelect	Design oligonucleotides for array analysis	► https://omictools.com/probeselect-tool (Li and Stormo 2001)
ProbeWiz	Optimal PCR primer pairs for cDNA arrays	► http://www.cbs.dtu.dk/services/DNAarray/probewiz.php (Nielsen and Knudsen 2002)
ProDesign	Oligonucleotide design for microarray	► http://wwwlabs.uhnresearch.ca/tillier/ProDesign/ProDesign.html (Feng and Tillier 2007)
SNPbox	Large-scale design and exons	► http://www.molgen.ua.ac.be/bioinfo/projects/snpbox/snpbox.htm (Weckx et al. 2004, 2005)
ProbeCheck	Especially for 16S rRNA	► http://131.130.66.200/cgi-bin/probecheck/probecheck.pl

5.6 Data Formats

The FASTA format is used most frequently for sequence manipulation (Pearson and Lipman 1988) (see ► Chap. 2). This format is simply a text file starting with a ">" sign and followed by the name of the sequence on the first line. On the second and following lines, the nucleotide string is listed without any additional characters such as numbers or spaces. Some programs will only consider the ten first characters of the name or identifier on the first line. Sequences are usually downloaded in FASTA format from the databases.

For certain programs, the sequences included with multiple alignments need to be trimmed to the same length before they can be further analyzed for oligonucleotide design. First a multiple alignment is prepared, for example, by ClustalX or ClustalW (Larkin et al. 2007) and saved in the *.aln format (► Chap. 4). The program BioEdit (Hall, 1999) can be used to trim sequences to the same length. BioEdit is installed from the URL (► http://www.mbio.ncsu.edu/bioedit/page2.html). "File" and "Open" are activated in BioEdit and the file in *.aln format selected. The "Edit mode" is chosen and regions with gaps identified by "mousing over"

column-wise and deleted when the region is marked. The consensus line at the bottom is not to be touched. The file is then saved by "File" and "Save as" with the choice of FASTA format. The file is reopened with ClustalX or ClustalW and realigned and saved in appropriate format.

Important points to consider using BLAST for evaluation of primer and probes are to use the lowest word size (7) and to turn off dust filters and low complexity filters. Even with these precautions, BLAST searchers in databases with oligonucleotides cannot always be expected to identify all sequences with perfect match. This is related to the short sequence length.

5.7 Programs

A list of relevant computer programs for primer and probe design is shown in ◘ Table 5.3. For these programs, we have briefly tested that the links have been open and that program installation has worked; however, full functionality has not been tested exhaustively. Center for Human and Clinical Genetics at Leiden University Medical Center (▶ http://www.humgen.nl/primer_design.html) maintains a list of programs. Many programs are online installations on servers assessed by the Internet as described yearly with the server issue of the journal, *Nucleic Acids Research* (▶ https://academic.oup.com/nar/issue/45/W1).

These resources can be consulted to fix broken links as well as to search for new programs.

5.8 Activities

5.8.1 Exploratory Primers with Primer3 for Recognition of Single DNA Sequences

The purpose is to amplify DNA from a strain of avian influenza virus. Primer3 is used at ▶ http://bioinfo.ut.ee/primer3-0.4.0/ (Untergasser et al. 2007, 2012). For this activity we will use the DNA sequence of the hemagglutinin (HA) gene of strain 1734 with serotype H5N1 isolated from duck in Fujian in 2005 with GenBank/EBI/DDBJ acc. no. DQ095629. Download the sequence from NCBI or similar database as described in ▶ Chap. 3. Cut and paste the DNA sequence into the window. You first have to select "Pick left primer" and "Pick right primer," and at Product size ranges, delete all ranges except for 100–300, and leave other settings as defaults. Press "Pick Primers." Two primers are suggested with a predicted PCR product of 203 bp. Inspect the location and orientation of the primers (marked with >>>>> on output). Four more sets of primers are further suggested.

5.8.2 Diagnostic Primers with PrimerBLAST

The program is provided by NCBI to select specific PCR primers for one taxon and avoid amplification of others.

This server application at NCBI is based on Primer3 and BLAST. Open the program from: ▶ http://www.ncbi.nlm.nih.gov/tools/primer-blast/

For this example we will design specific PCR primers for the serotype H7N2 of influenza virus. Download GenBank/EBI/DDBJ acc. no. **U20461** in FASTA format (▶ Chap. 2). Include the target sequence by "pasting" in the window or "choosing a file," or simply write the GenBank acc. no. in the yellow field.

In the section **Primer Pair Specificity Checking Parameters** at
Search mode, change **Automatic** to **User guided**.

At **Organism** select family or the specific organisms you are working with (for prokaryotes the genus name), in this case **Influenzavirus A**.

At **Database** select **nr**

Mark **Show results in new window**

Press **Get primers**

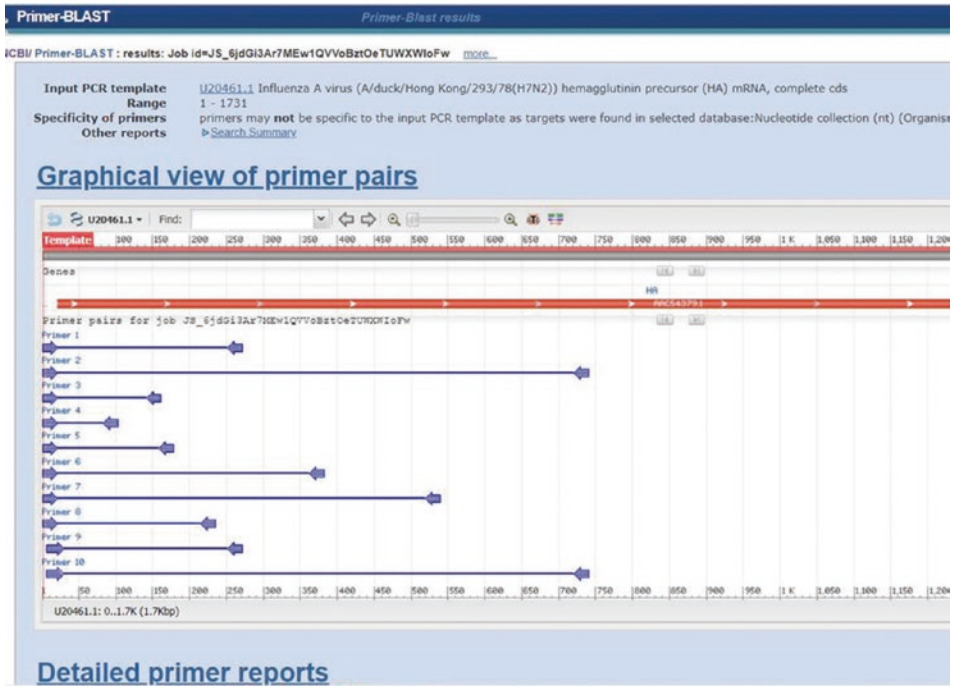
When the search is completed, look at the list to remove unwanted targets and to include more similar targets of the same type (H7N2) (your PCR template is highly similar to the following sequence(s) from the search database. To increase the chance of finding specific primers…). In this case you can mark the three H7N2 sequences (CY006029, CY067681, and AB302789).

Mark **Show results in a new window**

Then proceed by pressing **Submit.**

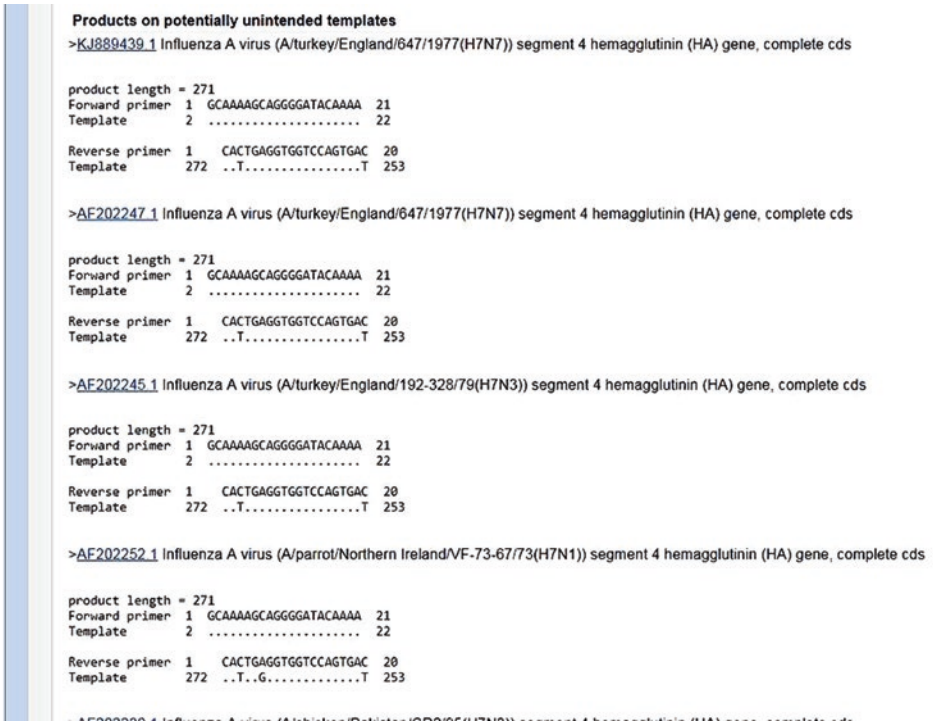
Review the list and consider if a specific primer pair can be designed (it should look like ◘ Fig. 5.6). Are all primers of **Products of intended targets** for **Primer pair 1** matching the type H7N2? Are all **Products of un-intended targets** forming mismatches that are expected not to result in generation of product? (Look at Kwok's rules ◘ Fig. 5.4.)

Note that this is an automatic prediction and you need to judge if the suggestions would be useful in real PCR tests. Among the unintended targets (see also ◘ Fig. 5.7), the forward



◘ **Fig. 5.6** Output from primerBLAST as expected from ► Activity 5.8.2. You see the graphical view of suggested primers for specific amplification of the hemagglutinin gene of the H7N2 type of type A influenza virus. (BLAST [Internet]. Bethesda (MD), National Library of Medicine (USA), National Center for Biotechnology Information; 2004 – [cited 25 18 04]. Available from ► https://www.ncbi.nlm.nih.gov/tools/primer-blast/)

primer is matching, and mismatches are only observed on the reverse primers, and even for them, it is questionable if they will actually not lead to PCR amplification in a real PCR since the C-T mismatch in the 5′ end is expected to result in amplification (◘ Fig. 5.4).



Products on potentially unintended templates

>KJ889439.1 Influenza A virus (A/turkey/England/647/1977(H7N7)) segment 4 hemagglutinin (HA) gene, complete cds

```
product length = 271
Forward primer  1  GCAAAAGCAGGGGATACAAAA  21
Template        2  .....................  22

Reverse primer  1    CACTGAGGTGGTCCAGTGAC  20
Template        272  ..T.................T  253
```

>AF202247.1 Influenza A virus (A/turkey/England/647/1977(H7N7)) segment 4 hemagglutinin (HA) gene, complete cds

```
product length = 271
Forward primer  1  GCAAAAGCAGGGGATACAAAA  21
Template        2  .....................  22

Reverse primer  1    CACTGAGGTGGTCCAGTGAC  20
Template        272  ..T.................T  253
```

>AF202245.1 Influenza A virus (A/turkey/England/192-328/79(H7N3)) segment 4 hemagglutinin (HA) gene, complete cds

```
product length = 271
Forward primer  1  GCAAAAGCAGGGGATACAAAA  21
Template        2  .....................  22

Reverse primer  1    CACTGAGGTGGTCCAGTGAC  20
Template        272  ..T.................T  253
```

>AF202252.1 Influenza A virus (A/parrot/Northern Ireland/VF-73-67/73(H7N1)) segment 4 hemagglutinin (HA) gene, complete cds

```
product length = 271
Forward primer  1  GCAAAAGCAGGGGATACAAAA  21
Template        2  .....................  22

Reverse primer  1    CACTGAGGTGGTCCAGTGAC  20
Template        272  ..T..G.............T  253
```

◘ **Fig. 5.7** Output from primerBLAST as expected from ► Activity 5.8.2. You see the list of predicted primers matching unintended targets for specific amplification of the hemagglutinin gene of the H7N2 type of type A influenza virus. (BLAST [Internet]. Bethesda (MD), National Library of Medicine (USA), National Center for Biotechnology Information; 2004 – [cited 25 18 04]. Available from ► https://www.ncbi.nlm.nih.gov/tools/primer-blast/)

Take-Home Messages

- Oligonucleotides are called primers when used for PCR and probes when used for hybridization.
- Primers and probes can either be for exploratory applications or for diagnostic applications.
- Sequence comparisons of oligonucleotides and their binding to template DNA can either be done by string comparison by score or by nearest neighbor comparisons using the thermodynamic properties of nucleotide hybridization.
- Special applications of primers relate to degenerate PCR, nested PCR, gene cloning, diagnostic applications, and primers for multiplex PCR.
- Computer programs are available for the different applications related to PCR as well as for hybridization such as microarrays and in situ hybridization.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W.and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers EW. 2018. GenBank. Nucleic Acids Res. 46(D1):D41–D47.

Borer, P. N., Dengler, B., Tinoco, I. Jr. and Uhlenbeck, O. C. 1974. Stability of ribonucleic acid double-stranded helices. J. Mol. Biol. 15, 843–853

Breslauer, K. J., Frank, R., Blocker, H. and Marky, L. A. 1986. Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. U. S. A. 83, 3746–3750.

Büssow, K., Hoffmann, S., Sievert, V. 2002. ORFer – retrieval of protein sequences and open reading frames from GenBank and storage into relational databases or text files. BMC Bioinformatics 23:40

Chen, S. H., Lin, C. Y., Cho, C. S., Lo, C. Z. and Hsiung, C. A. 2003. Primer Design Assistant (PDA): A web-based primer design tool. Nucleic Acids Res. 31, 3751–3754.

Gadberry M. D., Malcomber S. T., Doust A. N., and Kellogg E. A. 2005. Primaclade--a flexible tool to find conserved PCR primers across multiple species. Bioinformatics 21:1263–1264.

Dieffenbach, C. W., Lowe, T. M. J. and Dveksler, G. S. 1995. General concepts for PCR primer design. pp. 133–142. In PCR primer a laboratory manual. Cold Spring Harbor Lab. Press.

Feng, S. and Tillier, E. R. 2007. A fast and flexible approach to oligonucleotide probe design for genomes and gene families. Bioinformatics 23, 1195–1202.

Fredslund, J., Schauser, L., Madsen, L. H., Sandal, N. and Stougaard, J. 2005. PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. Nucleic Acids Res. 33, W516 - W520.

Gibson, G. and Muse, S. V. 2004. A primer of genome science. Sinauer, Sunderland.

Grunenwald, H. 2003. Optimization of polymerase chain reactions. In PCR Protocols. 2nd ed. ed. Bartlett, J. M. S. and Stirling, D. pp. 89–99. Methods in Molecular Biology 226. Humana Press, Totowa.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95–98.

Huggett, J. F., Foy, C. A., Benes, V., Emslie, K., Garson, J. A., Haynes, R., Hellemans, J., Kubista, M., Mueller, R. D., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J., Wittwer, C. T., Bustin, S. A. 2013. The digital MIQE guidelines: Minimum Information for Publication of Quantitative Digital PCR Experiments. Clin Chem. 59, 892–902. doi: https://doi.org/10.1373/clinchem.2013.206375.

Hyndman, D. L. and Mitsuhashi, M. 2003. PCR Primer Design. In PCR Protocols. 2nd ed. ed. Bartlett, J. M. S. and Stirling, D. pp. 81-88 Methods in Molecular Biology 226. Humana Press, Totowa.

Ikuta, S., Takagi, K., Wallace, R. B., and Itakura, K. 1987. Dissociation kinetics of 19 base paired oligonucleotide-DNA duplexes containing different single mismatched base pairs. Nucleic Acids Res. 15, 797–811.

Karsch-Mizrachi I, Takagi T, Cochrane G; International Nucleotide Sequence Database Collaboration. 2018. The international nucleotide sequence database collaboration. Nucleic Acids Res. 46(D1):D48-D51.

Ke, X., Collins, A. and Ye, S. 2001. PIRA PCR designer for restriction analysis of single nucleotide polymorphisms. Bioinformatics 17, 838–839.

Kibbe, W. A. 2007. OligoCalc: an online oligonucleotide properties calculator. Nucleic Acids Res. 35, W43–6.

Kodama Y, Mashima J, Kosuge T, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, Takagi T. 2018. DNA Data Bank of Japan: 30th anniversary. Nucleic Acids Res. 46(D1):D30-D35.

Kwok, S., Kellogg, D. E., McKinney, N., Spasic, D., Goda, L., Levenson, C., and Sninsky, J. J. 1990. Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. Nucleic Acids Res. 18, 999–1005.

Kwok, S., Chang, S-Y., Sninsky, J. J. and Wang, A. 1995. Desing and use of mismatched and degenerate primers. In Dieffenbach, C. W. and Dveksler, G. S. PCR primer a laboratory manual. pp. 143-155. Cold Spring Harbor Lab. Press.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007): Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.

Lathe, R. 1985. Synthetic oligonucleotide probes deduced from amino acid sequence data. Theoretical and practical considerations. J. Mol. Biol. 183, 1–12.

Lawyer, F. C., Stoffel, S., Saiki, R. K., Myambo, K., Drummond, R., and Gelfand, D. H. 1989. Isolation, characterization, and expression in Escherichia coli of the DNA polymerase gene from Thermus aquaticus. J. Biol. Chem. 264, 6427–6437.

Li, F. and Stormo, G. D. 2001. Selection of optimal DNA oligos for gene expression arrays. Bioinformatics 17, 1067–76

Ludwig, W., Strunk, O., Westram, R. and 29 other authors. 2004. ARB: a software environment for sequence data. Nucleic Acids Res. 32, 1363–71.

McConaughy BL, Laird CD, McCarthy BJ. 1969. Nucleic acid reassociation in formamide. Biochemistry 8, 3289–3295.

Meinkoth, J. and Wahl, G. 1984. Hybridization of nucleic acids immobilized on solid supports. Anal. Biochem. 138, 267–284.

NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44(D1):D7–19.

NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 46, D8-D13.

Newton, C. R., Graham, A., Heptinstall, L. E., Powell, S. J., Summers, C., Kalsheker, N., Smith, J. C. and Markham, A. F. (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res 17, 2503–2516.

Nielsen, H. B. and Knudsen, S. 2002. Avoiding cross hybridization by choosing nonredundant targets on cDNA arrays. Bioinformatics 18, 321–322.

Onodera, K. and Melcher, U. 2002. VirOligo: a database of virus-specific oligonucleotides. Nucleic Acids Res. 30, 203–204.

Owczarzy, R., Tataurov, A.V., Wu, Y., Manthey, J.A., McQuisten, K.A., Almabrazi, H.G., Pedersen, K.F., Lin, Y., Garretson, J., McEntaggart, N.O., Sailor, C.A., Dawson, R.B., Peek, A.S. 2008. IDT SciTools: a suite for analysis and design of nucleic acid oligomers. Nucleic Acids Res. 36, W163–9.

Pearson, W. R. and Lipman, D. J. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U. S. A. 85, 2444–2448.

Pozhitkov, A., Noble, P. A., Domazet-Loso, T., Nolte, A. W., Sonnenberg, R., Staehler, P., Beier, M. and Tautz, D. 2006. Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. Nucleic Acids Res. 34, e66.

Rachlin, J., Ding, C., Cantor, C. and Kasif, S. 2005. MuPlex: multi-objective multiplex PCR assay design. Nucleic Acids Res. 33, W544 - W547.

Rice P, Longden I & Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genetics 16:276–277.

Rose, T. M., Henikoff, J. G. and Henikoff, S. 2003. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Res. 31, 3763–3766.

Rouillard, J. M., Zuker, M., and Gulari, E. 2003. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. 31, 3057–3062.

Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. pp. 365-386 In Misener, S. and Krawetz, S. A. Bioinformatics Methods and Protocols. Methods in Molecular Biology 132. Humana, Totowa.

Rychlik, W. and Rhoads, R. E. 1989. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res. 17, 8543–8551.

Rychlik, W., Spencer, W. J. and Rhoads, R. E. 1990. Optimization of the annealing temperature for DNA amplification in vitro. Nucleic Acids Res. 18, 6409–6412.

SantaLucia, J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Nath. Acad. Sci. U. S. A. 95, 1460–1465.

Schildkraut, C. 1965. Dependence of the melting temperature of DNA on salt concentration. Biopolymers 3, 195–208.

Sharkey, F. H., Banat, I. M. and Marchant, R. 2004. Detection and quantification of gene expression in environmental bacteriology. Appl. Environ. Microbiol. 70, 3795–3806.

Suggs, S. V., Hirose, T., Miyake, E. H., Kawashima, M. J., Johnson, K. I. and Wallace, R. B. 1981. In ICN-UCLA Symp. Dev. Biol. Using Purified Genes. Brown, D. D. (ed.) Acad. Press., New York. vol. 23, 683-693.

Szostak, J. W., Stiles, J. I., Tye, B.-K., Chiu, P., Sherman, F. & Wu, R. 1979. Hybridization with synthetic oligonucleotides. Methods Enzymology 68, 419–428.

Tsai, M. F., Lin, Y. J., Cheng, Y. C., Lee, K. H., Huang, C. C., Chen, Y. T. and Yao, A. (2007) PrimerZ: streamlined primer design for promoters, exons and human SNPs, Nucleic Acids Res. 35, W63–W65).

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R. and Leunissen, J. A. M. (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 35, W71 - W74.
Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G. 2012. Primer3-new capabilities and interfaces. Nucleic Acids Res. 40:e115.
Urakawa, H., Noble, P. A., El Fantroussi, S., Kelly, J. J., Stahl, D. A. 2002. Single-base-pair discrimination of terminal mismatches by using oligonucleotide microarrays and neural network analyses. Appl. Environ. Microbiol. 68, 235–244.
Weckx, S., De Rijk, P., Van Broeckhoven, C. and Del-Favero, J. 2004. SNPbox: web-based high-throughput primer design from gene to genome. Nucleic Acids Res. 32, W170–2.
Weckx, S., De Rijk, P., Van Broeckhoven, C. and Del-Favero, J. 2005. SNPbox: a modular software package for large-scale primer design. Bioinformatics 21, 385–7.
Wong, M. L. and Medrano, J. F. 2005. Real-time PCR for mRNA quantitation. Biotechniques 39, 75–85.
Xu, D., Li, G., Wu, L., Zhou, J. and Xu, Y. 2002. PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. Bioinformatics 18, 1432–1437.

Further Readings

Introduction to practical work with PCR as well to the historical background is found in Sambrook and Russell (2001).

Sambrook and Russell. 2001. Molecular Cloning. A laboratory manual. CSHL Pres.

Short Introduction to Phylogenetic Analysis of Molecular Sequence Data

Henrik Christensen and John Elmerdahl Olsen

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_6

What You Will Learn in This Chapter

You will learn how to read a phylogenetic tree and evaluate the nature of the data that are suitable for phylogenetic analysis. The models required for phylogenetic analysis such as the substitution matrix, the tree shape, and the weights that can be put on different positions in the multiple alignment are introduced. You are then presented for the different types of phylogenetic methods in order to learn their basic principles and the program packages from where they can be used. You are guided to evaluate the strengths of phylogenetic trees and to optimize model parameters. In the activity you will learn how to construct a neighbor joining phylogenetic tree on your own computer.

6.1 Background

Phylogeny (Phylo- from Greek, family, tribe; -geny, ancestry) has become one of the most fundamental bioinformatical tools. The definition of phylogeny is that it is a model of the relationships between organisms, genes, protein, and other structures based on common ancestry. Lamarck (1809) was first to present an evolutionary tree. The tree was based on bifurcating lines connecting different invertebrate animal groups as well as vertebrates such as fish, birds, mammals, and subdivisions within the mammals. Darwin (1859) used a hierarchical tree-like structure to illustrate common ancestry among organisms, alternating processes of variation and selection, and geological stratification of organisms. The hierarchical tree and Darwin's conceptual "tree of life" was used by Haeckel (1875) graphically to illustrate the relationships between organisms. Hennig (1950, 1966) developed concepts statistically to analyze biological data in a phylogenetic perspective.

Phylogenetic analysis will only make sense if the characters compared are homologous. DNA and protein sequences are homologous if they share common ancestry. The justification for homology based on DNA sequence data comparison has promoted phylogenetic analysis tremendously. Other reasons to the frequent use of phylogenetic analysis are the easy access to molecular sequencing and access to sequence data through the Internet and that hierarchical data representation and dichotomies fit well into construction of computer programs.

Phylogeny sensu stricto is dealing with the ancestral relationships of species. Currently "phylogeny" is also used about numerous biological relationships between populations, genes, and groups of organisms. Data expected to include phylogenetic information are DNA or protein sequences of homologous genes and morphological characters shown to be homologous, while most restriction patterns (PFGE, AFLP, ribotyping) and most DNA hybridization data (micro-array, Southern blot, etc.) are data with dubious phylogenetic information.

Phylogenetic analysis involving molecular sequence comparisons has at least four uses which are central to a range of downstream bioinformatic analysis (Fig. 6.1):

- *Classification (taxonomy)*
- *Grouping of genes, proteins, and other molecular sequences including noncoding sequences*
- *Epidemiological investigations*
- *Analysis of parallel evolution between host and parasite*

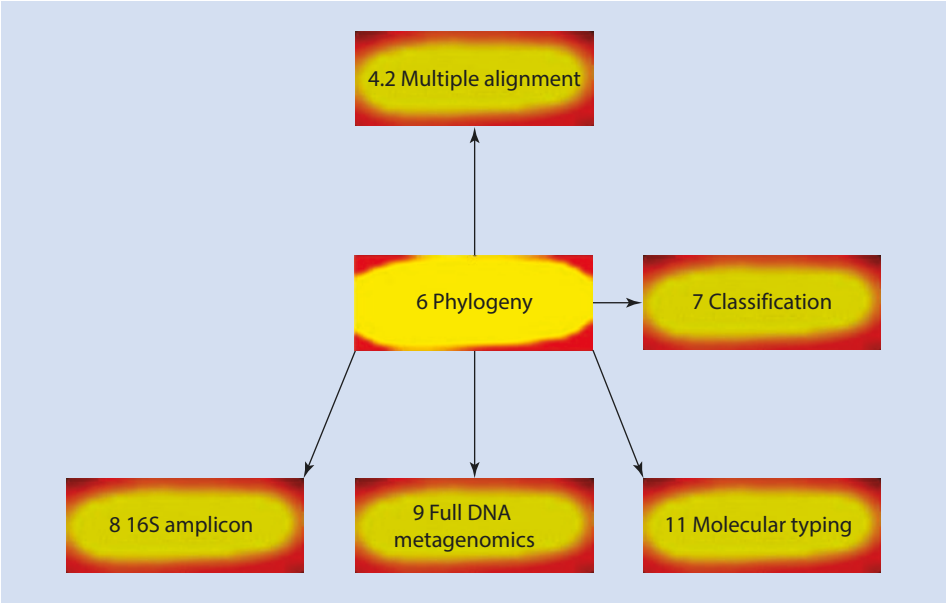


Fig. 6.1 Relation of this chapter to the other chapters in the book. Phylogeny is based on multiple alignment of sequences, and it can be used for classification. Phylogenetic analysis is included as a tool in many other applications such as 16S rRNA amplicon sequencing, full DNA metagenomics, and molecular typing

6.2 Understanding the Phylogenetic Tree

In its most simple form, a phylogenetic tree is represented by straight lines joined as bifurcations. The free ends of lines represent observed sequences (or any other character compared), and the junctions between lines (nodes) represent the hypothetical ancestors for the observed sequences (Fig. 6.2). The lengths of lines to each sequence reflect sequence divergence; however, they can also be of equal length and only show the relationships between species.

The tree can either be on a radial form or on a dendrogram form (Fig. 6.3). On the dendrogram form, all lines from the radial tree are represented vertically, and the horizontal lines included have the purpose only of showing the relationships between the horizontal branches. A radial tree can always be represented as a dendrogram with the same information and vice versa (Fig. 6.3).

To get the most obvious information out of a dendrogram, it needs to be rooted with the out-group. The out-group is a branch selected to be unrelated to all other branches of the tree meaning that it is a branch with a free end which is remarkably longer than the other branches. The root is selected as the vertical line at the most far left of the tree. The root is selected as the starting point of the dendrogram (Fig. 6.4). In the example shown on Fig. 6.4, the internal branch is rotated 180° which will bring the longest branch to the root.

In the strict sense, a root is deliberately selected as a branch with a free end without any label (see Figs. 6.5 and 6.6); however, in practice, all methods demonstrated in the following are producing unrooted trees, and it will not make sense further to consider true rooted trees in this chapter.

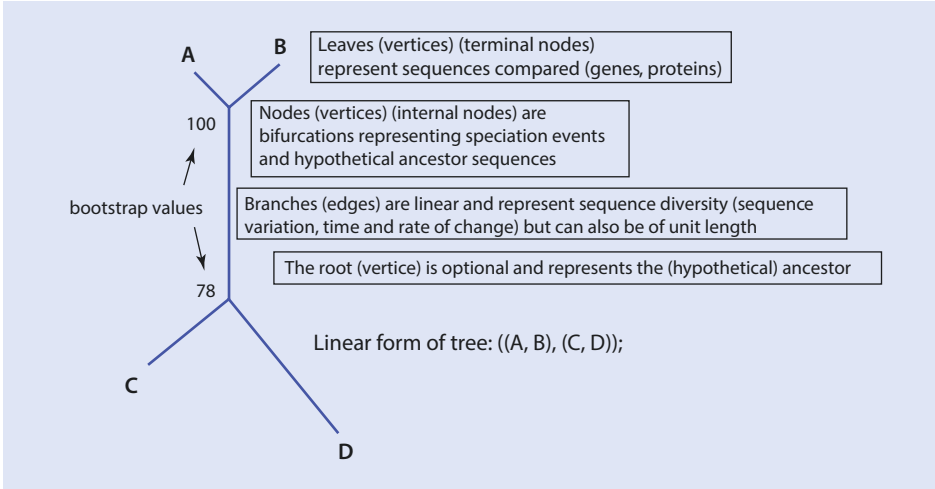


Fig. 6.2 Phylogeny linked to its graphic representation the "tree." A, B, C, and D represents some kind of homologous character that can be compared. In the current context of bioinformatics, they can be DNA or protein sequences. The linear formula for the tree shown below is in Newick format (named after ► http://www.newicks.com/). This is the most simple way to instruct a computer program to handle a phylogenetic tree

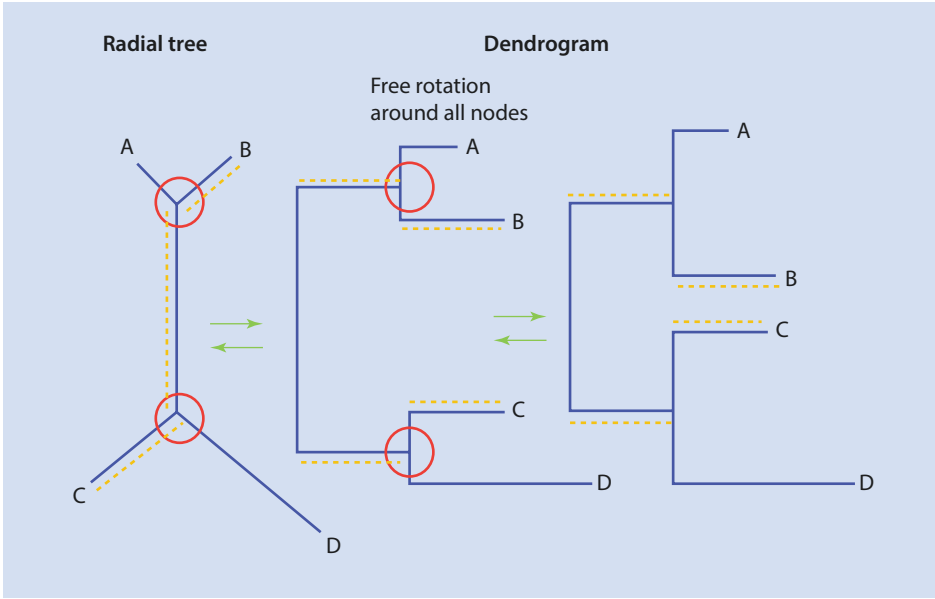


Fig. 6.3 Illustration of free rotation around nodes and visualization as radial tree or dendrogram. All three unrooted trees show exactly the same information. The branch lengths are proportional to sequence divergence or other similar characters. In the radial tree, all branch lengths are related to the sequence diversity. In the dendrogram, it is only the horizontal lines that are related to sequence divergence. The vertical lines are inserted to show how the horizontal branches are related. The calculation of the distance from A to C is illustrated by the stippled orange line, and it is seen that only the lengths of the horizontal lines count in the dendrogram

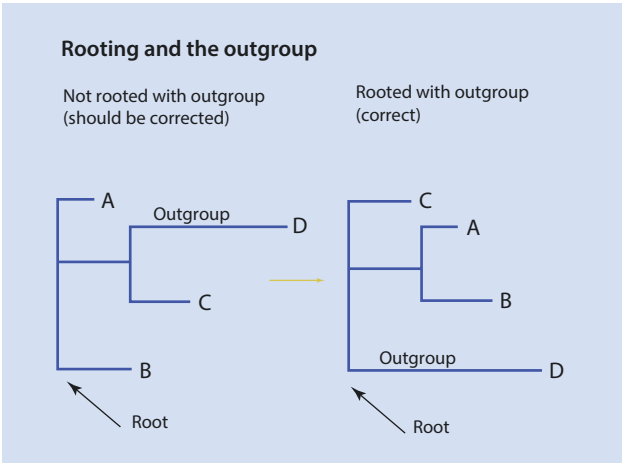


Fig. 6.4 Illustration of rooting with the out-group. The internal branch is mirrored placing the longest branch leading to D at the root. Note that the information obtained from the tree before and after rooting is the same. The rooting with the out-group is only performed to improve the reading of the tree

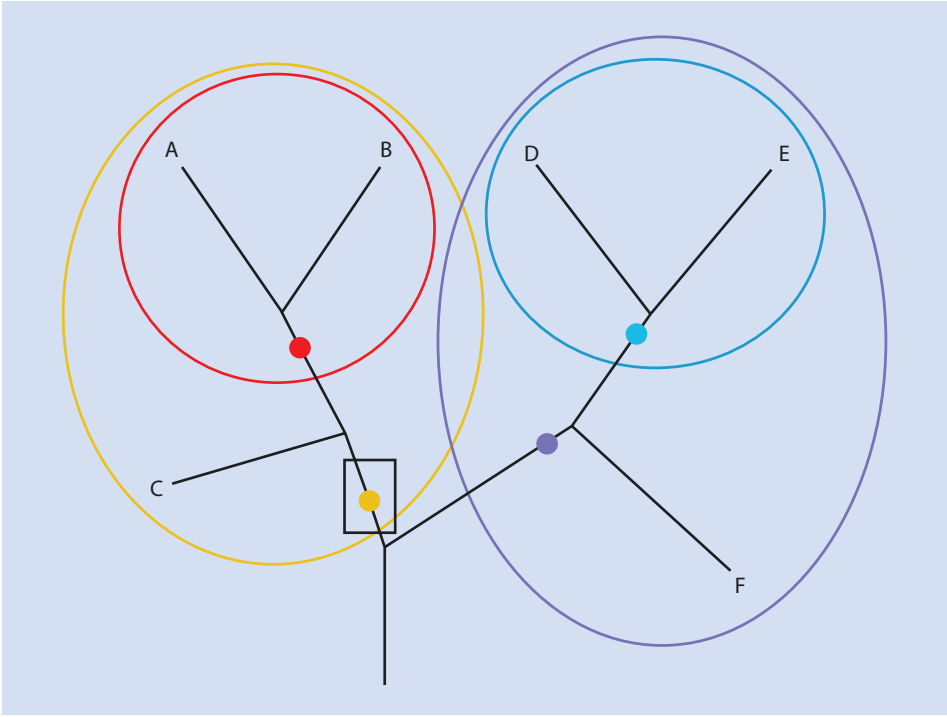


Fig. 6.5 Definition of monophyletic groups (clades). A monophyletic group (clade) is characterized by common descent of all members (at least two) and by all members sharing one and only one common branch (*dot*). This tree includes four monophyletic groups marked in different colors. Note that this tree is rooted. The monophyletic group concept applies both to rooted and unrooted trees

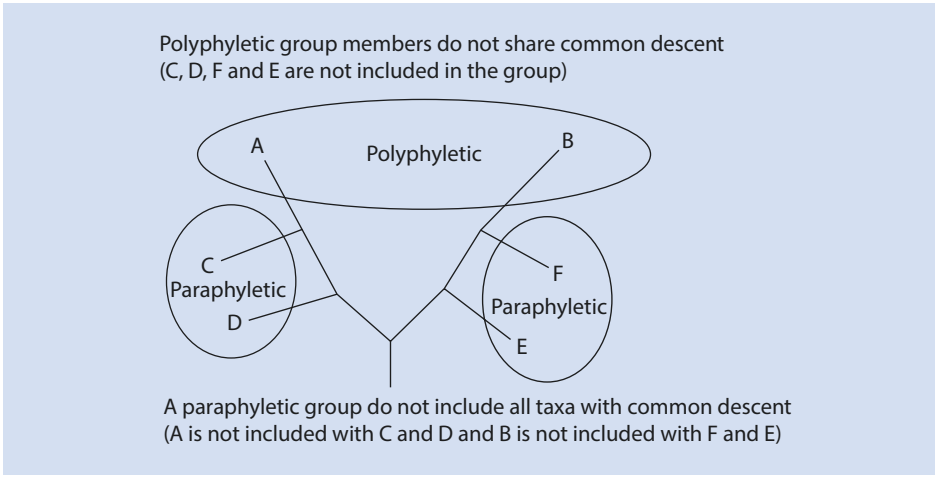


Fig. 6.6 Para- and polyphyletic trees. Note that this tree is rooted. The para- and polyphyletic group concepts apply both to rooted and unrooted trees. Like for the monophyletic groups in Fig. 6.5, many para- and polyphyletic groups can be identified in a phylogenetic tree; however, only a few examples are given above since these groups are less interesting compared to the monophyletic

The most important information that can be read from a phylogenetic tree is the location of monophyletic groups (Fig. 6.5). A monophyletic group (clade) is characterized by common descent of all members (at least two) and by all members sharing a common branch. A phylogenetic tree contains many different monophyletic groups. Only some will have biological meaning. Paraphyletic groups have two ancestors, and groups are polyphyletic if they have more than two ancestors (Fig. 6.6). Ancestral taxa tend to be paraphyletic since they are very difficult to resolve. Polyphyletic groups are without a common ancestor, and polyphyletic groups are normally just considered for their negative nature.

6.3 Assumptions About Data in Order to Perform Phylogenetic Analysis

A list of nine conditions can be made about the demands to data to make a phylogenetic tree (Table 6.1). All nine assumptions will probably never be satisfied, or it cannot be tested if they all are satisfied. The first two about homology and multiple alignment are obligatory. It will only make sense to compare homologous sequences based on a good multiple alignment.

If some conditions cannot be met in order to construct a phylogeny, a network can be made. A network is characterized by trifurcations instead of bifurcations. The program **SplitsTree** (Huson 1998) (▶ http://www.splitstree.org/) can be used to analyze the data. It is used on data that are related by a network. If a tree of the seven species A–G is compared and a "split" between F and B evaluated, ACDF| BEG, "to satisfy a phylogeny, the distances across the split should be pairwise higher than on each side; if this is not possible, it must be a network." A weakness of the method is that it is distance matrix based and cannot account for all different positions of the alignment individually. The main reason to form network structures is weakly defined, but some are related to horizontal gene transfer

Table 6.1 Assumptions about data

No.	Assumption	Tests
1	Sequences compared are homologous	The homology of molecules can normally be verified by high-scoring pairwise alignments, for example, by BLAST or by analysis of shared domains in relation to proteins
2	Positional homology can be established	Sequences can be aligned (multiple alignment)
3	Sequences have been selected in a way that they are representative for the units they are meant to represent (species, genes, proteins)	More sampling with stratified representation of groups at the same level should give the same result
4	Independent sampling of the units in 3 such as species and genes has been performed	Better stratified representation of groups at the same level should give the same result
5	Sequences compared have evolved independently whether they represent species, genes, or proteins	Comparison of other related species, genes, or proteins should give the same result
6	Nucleotide or amino acid positions have evolved independently	Comparison of other related genes should give the same result
7	Events are rare	Comparison of more conserved sequences or conserved regions of sequences compared should give the same result
8	Same rate of evolution in all sequences (some methods)	Further testing of model parameters
9	Long branch attraction can be handled (some methods). Long branch attraction is if the phylogeny includes very short and very long branches, then the risk for the methods to reverse long branches on short branches increases	Reduced mixing of short and long branches should give the same result

(HGT). Horizontal gene transfer is a violation of condition 3 in Table 6.1. HGT will be reflected in a network structure, but a network structure will not always have been caused by HGT.

6.4 Phylogenetic Model Parameters

6.4.1 The Tree Structure

When we see a tree, we often take it for granted. However, in phylogenetic analysis, the tree is only one out of many model parameters that we try to optimize given our data – the multiple alignment. We want to identify the tree which best reflects our data. For some

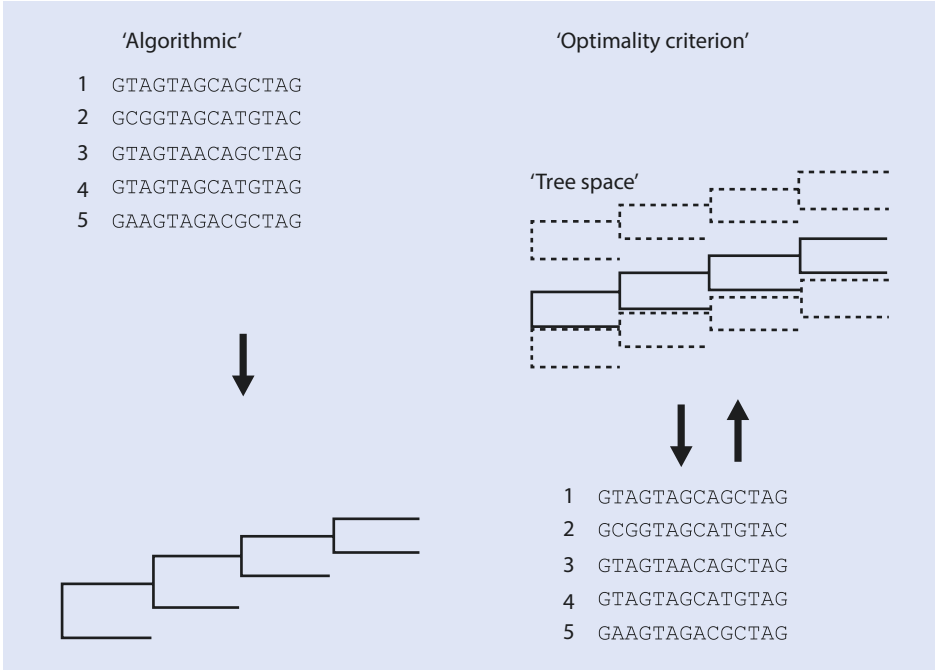


Fig. 6.7 Phylogenetic methods. To make a tree from the multiple alignment or to fit a tree to the multiple alignment

phylogenetic methods, this means we can test different trees against our data. It is shown in the right part of the model on Fig. 6.7. There are two problems with the statistical analysis of tree structures. The first is that the statistical distribution of phylogenetic trees is very complex and cannot be compared to known statistical distributions like the normal – or gamma –distributions. Another problem is that there are unrealistic high numbers of trees to compare (Felsenstein 2004). The solution most often chosen is heuristic tree search where only the trees with the highest probability to represent the data are compared.

6.4.2 Substitution Matrix and Evolutionary Models

The probability of changing one nucleic acid or one amino acid to another is defined by comparing the actual data to a substitution matrix. For this reason, the substitution matrix should be selected to represent the data closely. Substitution matrices were described in ▶ Chap. 4. For DNA, an equal probability for exchange of nucleotides is assumed by the Jukes and Cantor matrix, whereas compensation for a transition/transversion bias is possible by other substitution models. For protein the PAM (point accepted mutations) matrix is most suitable for evolutionary and phylogenetic analysis. With maximum parsimony and maximum likelihood methods described below, the substitution matrix is incorporated into the methods, whereas for the distance matrix/neighbor joining method, calculation of the distance matrix is always the first step later followed by the construction of the phylogenetic tree.

6.4.3 Weighting of Characters

The weighting of characters is sometimes included as an additional model parameter. Weighting of characters is done with the different columns of the multiple alignment, and such weighting is performed to place less weight on characters or nucleotides or amino acids with frequent changes. Different weights w_1, w_2, w_3,..., w_n are placed on n positions of the multiple alignment. For example, with codon positions in DNA, the weights $w_1 = 3$, $w_2 = 4$, and $w_3 = 1$ can be added accounting for higher variation in the third codon position compared to first and second.

6.5 Phylogenetic Methods

Algorithmic (numeric) methods such as neighbor joining analysis work by simply "calculating" a tree based on the data. Only one tree is generated. A specific sequence of steps (algorithm) is defined that leads to the tree (Fig. 6.7). The alternative procedure used in maximum parsimony, maximum likelihood, and Bayesian methods is based on a so-called optimality criterion. An optimality criterion (described by an objective function) is a procedure for evaluating a given tree in relation to the data (Fig. 6.7).

The interpretation of a phylogeny according to the second approach is not always unequivocal. Especially for complex phylogenies, many trees can be considered to reflect the phylogeny given the data. The trees accepted according to certain statistical limits can be interpreted as the tree "space" given the data (Fig. 6.7).

6.5.1 Maximum Parsimony

The maximum parsimony tree represents the minimum number of nucleotide or amino acid changes given the data (multiple alignment) (Table 6.2). This method was the first phylogenetic method to be used with molecular sequences (Eck and Dayhoff 1966). Two fundamentally different ways of comparing data by parsimony exist. One is based on the separation of data into shared and derived characters and use of only the derived common characters (synapomorphs) for parsimony analysis (Hennig 1966). This procedure cannot be used with molecular sequence data because differences between ancestral and derived change are normally unknown. The alternative choice with these data is to relax initial assumptions about the direction of character changes.

The most strict parsimony method is the Camin and Sokal (1965) parsimony. The ancestral state must be known, and only single character changes in one direction are allowed. The method can be used to make a phylogeny of small DNA deletions that are not assumed to revert but cannot be used with sequences per se. Dollo parsimony refers to Dollo's law: "A complex character once attained cannot be attained in that form again" (evolution is irreversible). Only unidirectional changes are allowed although reversals are possible but minimized. This method can be used with restriction sites in DNA but not sequences. The Fitch-Wagner algorithm allows fully reversible changes and for this reason can be used with sequences. It is implemented as DNAPARS and DNAPROT in the PHYLIP package (see Tables 6.2 and 6.3) as well as other programs.

Table 6.2 Phylogenetic methods used with molecular sequence data

Methods	Principle	Benefit	Drawback	Use and limitations	Options for extension
Neighbor joining	Algorithmic	Simple and fast in relation to computational power	Only one tree	Routine draft phylogeny	Bootstrap
Maximum parsimony	Optimality criteria	Includes all informative positions of alignment and more trees are compared based on statistical criterion	More trees can be "equally parsimonious"	Routine phylogeny of closely related sequences	Bootstrap
Maximum likelihood	Optimality criteria	Includes all informative positions of alignment and more trees are compared based on statistical and probabilistic (likelihood criterion)	With complex phylogenies the "maximum likelihood" tree will never be found	Routine phylogeny with fewer sequences	Bootstrap and likelihood ratio tests
Bayesian (MrBayes)	Optimality criteria	Includes all informative positions of alignment and more trees are compared based on statistical and probabilistic (Bayesian) criterion)	Very complex phylogenies are not fully resolved	For complex phylogenies but not too complex	

Most parsimony methods work by "post-order tree traversal" meaning that a tree is evaluated in relation to the data starting with the branch tips and moving toward the base through all nodes.

6.5.2 Distance Matrix/Neighbor Joining

The first distance matrix method was published by Fitch and Margoliash (1967). Neighbor joining (Saito and Nei 1987) belongs to the clustering methods originally developed for numeric taxonomy (Sneath and Sokal 1973) (Table 6.3). It is now one of the most simple

Table 6.3 Most commonly used programs and program packages. For a comprehensive list of all packages and programs, see ► http://evolution.genetics.washington.edu/phylip/software.html

Purpose	Program/ package	URL, references and notes
Neighbor joining, maximum parsimony, and others	PHYLIP	► http://evolution.genetics.washington.edu/phylip.html The use of the package is free with open source code. PC version for use with DOS is available. The programs on the package can also be used from server
Maximum likelihood and maximum likelihood ratio test	fastDNAml	► http://iubio.bio.indiana.edu/soft/molbio/evolve/fastdnaml/fastDNAml.html (Olsen et al. 1994)
Maximum likelihood and maximum likelihood ratio test	PhyML	► http://www.atgc-montpellier.fr/phyml/(Guindon et al. 2010)
Neighbor joining, maximum parsimony, maximum likelihood	MEGA7	► https://www.megasoftware.net/ (Kumar et al. 2016). For simple analysis of phylogeny and population genetics, the use of MEGA is described in the book of Nei and Kumar (2000)
Maximum parsimony advanced and others	PAUP	► http://paup.phylosolutions.com/get-paup/ Source code hidden but see DNAPARS and PROTPARS in PHYLIP. Runs best on MAC. Without preliminary knowledge, the most easy way to start using PAUP is to modify one of the examples included when the program is downloaded
Maximum likelihood advanced applications linked to the specialized population genetics program	PAML	► http://abacus.gene.ucl.ac.uk/software/paml.html
	HyPhy	► http://hyphy.org/w/index.php/Download
Maximum likelihood with large datasets	FastTree	► http://www.microbesonline.org/fasttree/ (Price et al. 2010)
	RAxML	► https://sco.h-its.org/exelixis/software.html (Stamatakis 2014)
Tree manipulation	MacClade	► http://macclade.org/macclade.html

and most frequently used routine methods. First a distance matrix is calculated based on pairwise comparison of all sequences with each other. Then the neighbor joining algorithm constructs the tree. Clustering by the neighbor joining algorithm does not need a molecular clock (ultra-metric data). Neighbor joining analysis is used for draft phylogenies. Neighbor joining analysis can be performed directly from ClustalX (see below in activity ► Sect. 6.9.1).

6.5.3 Maximum Likelihood

The likelihood is the probability of the parameters (tree structure, substitution model) given the data (multiple alignment), and the tree is this way part of the model maximizing the likelihood in relation to data (■ Tables 6.2 and 6.3). Maximum likelihood analysis of phylogeny was first proposed by Cavalli-Sforza and Edwards (1967). It was later used on molecular sequences by J. Felsenstein and implemented in the PHYLIP package as the DNAML program. An extension including many scripts is available at the fastDNAml program (Olsen et al. 1994) (► ftp://ftp.bio.indiana.edu/molbio/evolve/fastdnaml/) (■ Table 6.3).

For up to approx. 25 sequences, it is possible for the tree search algorithm to find the tree that will give the highest lnL. It can be verified by the loop script with fastDNAml. For more sequences it becomes difficult to justify if the tree with the highest lnL is achieved. One solution is to accept a cloud of trees with likelihood above a certain threshold. Another way to come around the problem is to use Bayesian inference (► Sect. 6.5.4).

The easiest way to run maximum likelihood analysis is to use the PhyML program implemented on the server at ► http://www.atgc-montpellier.fr/phyml/ (Guindon et al. 2010). RAxML is another implementation of maximum likelihood found at ► https://sco.h-its.org/exelixis/software.html (Stamatakis 2014). For large datasets up to a million of sequences, FastTree (Price et al. 2010) (■ Table 6.3) can be used. It infers approximately maximum likelihood phylogenetic tress from multiple alignments of DNA or protein sequences.

6.5.4 Bayesian (MrBayes) Inference of Phylogeny

This method is recommended to complex phylogenies not resolved fully by maximum likelihood (■ Table 6.2). The method is based on the Bayesian statistical principle and is nicknamed MrBayes. The method starts with a best guess of a tree to calculate the prior probability. Trees are simulated by MCMC (Markov Chain Monte Carlo), and all the best trees are kept. The posterior probabilities for the branches in all these best trees are then written on to the final tree. These numbers are true probabilities in a statistical sense saying, for instance, that the branch leading to taxon 1 has 93% probability. The program is available from ► http://mrbayes.sourceforge.net/. At this website a very good and comprehensive manual can be found (■ Table 6.3). The installation and execution of the program are not always running so smoothly related to conflicts between versions of the program and versions of Windows. To solve this and similar problems in bioinformatics, great patience is needed.

6.6 Comparison of Phylogenetic Methods

Phylogenies are in most cases hypothesis about evolution, and appropriate approaches to test phylogenetic methods are few. One can either use simulated data to define a phylogeny and then compare methods against this, or one can use real evolution of organisms with fast evolution, for example, virus. Exact phylogenies for organisms with really well-defined paleontologic evidence may also be used.

A phylogenetic method is consistent for an evolutionary model if the results obtained by the method converge on the correct tree as the data become infinite. A phylogenetic method has high efficiency if it quickly converges on the correct solution as more data are applied to the problem. A phylogenetic method is robust if it converges on the correct solution even with violations of the assumptions about the evolutionary model (Hillis 1995). Given the assumptions about data have been optimally handled (Table 6.1), we will of course prefer a phylogenetic method which is consistent, efficient, and robust.

Identical sequences should be avoided since it will only extend the time of computation and not add any new information to the analysis. For complex phylogenies all taxa should be included since it will be very difficult to judge which ones to exclude. A symmetric topology is said to indicate a balance between speciation and extinction, whereas a more comp-like topology should indicate either very high speciation or very high extinction rates.

Probably mainly for prokaryotes, a standard operating procedure for phylogenetic inference (SOPPI) using rRNA marker genes has been described by Peplies et al. (2008).

6.6.1 Bootstrap

Bootstrap analysis is a permutation test. The aim of permutation tests is to show how well supported the branches are by the data. The bootstrap estimates the sampling distribution of a group of sequences by generating new samples by drawing with replacement from the original data (Fig. 6.8). The alternative jackknife analysis computes the statistic of

The original data are simulated by drawing columns randomly with replacement 100 times (or a higher number). The phylogenetic analysis is repeated and the number of branched common in all 100 trees summarized.

original data	1 replicate	2 replicate	3 replicate
Species 1	AGGA	AGGA	GGAA
Species 2	ACGT	AACT	CGTT
Species 3	ACGT	AACT	CGTT
Species 4	ACTT	AACT	CTTT
Species 5	CCGT	CCCT	CGTT
linear form	(2,3)4)5)1;	(2,3)4)5)1;	(2,3)4)5)1;

All 100 trees are compared and the frequency of the same branches in the simulated tree are labelled on to the original tree eg. 100 and 78 on Fig. 6.2.

Fig. 6.8 A simple example with five short sequences (species) of only four nucleotides in length to show how bootstrap analysis works. Three replicates of the original multiple alignment are shown (the original multiple alignment is not shown). The simulated multiple alignments are each constructed by drawing (with replacement) one column at a time from the original multiple alignment. For each simulated multiple alignment, a phylogenetic tree has been constructed. The frequency that each branch occurs in the simulated trees is the bootstrap value. The bootstrap value can then labelled on the same branch in the original tree

interest for all combinations of the data where one (or more) of the original data points is removed. It will not be further described here.

For the bootstrap, the original multiple alignment data are simulated by drawing columns randomly with replacement 100 or 1000 times. The phylogenetic analysis is repeated for each of the 100 or 1000 simulated alignments (Fig. 6.8) and the number of branches common in all 100 or 1000 trees summarized in a consensus tree. These values are then compared to and written on the original tree.

The result of the bootstrap is a support for a particular internal branch in the original tree. As a consequence, the out-group receives no bootstrap since it is not an internal branch but a terminal. If reviewers who are not familiar with phylogenetic analysis request a bootstrap value to the out-group, you can use two out-group taxa, and then their common branch will receive an out-group. Unfortunately, this contradicts the general rule that there should be one and only one out-group of appropriate distance and relatedness to the taxa investigated for monophyly. The bootstrap values cannot be evaluated in a strictly probabilistic sense since it is just a parameter indicating the robustness of the branches in the tree. Bootstrap values at 90% or higher are very "good," those within 50–89% are acceptable, and those below 50% are not considered. Some branches may not receive a bootstrap value at all. This can happen if this branch has not been simulated in the permutation test. With closely related sequence, another problem can be rather low bootstrap values simply because the simulation imposes random noise into the data.

6.7 Data Formats

The most common format is the classical PHYLIP format.

- **Example of PHYLIP Interleaved Format:**

```
4 20
Species001aggcgctagc
Species002agtagctagc
Species003agtccctagc
Species004agtcgttagc

aggcgctagc
aggcgctagc
aggcgctagc
aggcgctagc
```

On the first line are the number of species (four in this example) and sequence length (20 in this example). On the next line the sequences are listed. Each sequence starts with a name no longer than ten characters. The sequence starts on position 11. Longer sequences continue in blocks separated by a blank line shift but without the species name. In the classic PHYLIP format, all blank spaces after position 11 should be removed. The alternative is PHYLIP, non-interleaved, where sequence after sequence is listed with line breaks. Remember to add "I" on the first line.

- **Example of PHYLIP Non-interleaved:**

```
4 20 I
Species001aggcgctagc
aggcgctagc

Species002agtagctagc
aggcgctagc

Species003agtccctagc
aggcgctagc

Species004agtcgttagc
aggcgctagc
```

For more advanced use of the PHYLIP format, different options in the form of a capital letter can be added on the first line and further instructions given on the second. For example, J for jumble on the first line changes the order that the lines in the alignment are read into the phylogenetic analysis.

Another common format is the NEXUS format originally used with PAUP (Table 6.3) but now also required for many other phylogeny programs including MrBayes.

The format looks like this:

```
#NEXUS
   BEGIN DATA;
   dimensions ntax=11 nchar=1495;

   format datatype=dna interleave=yes gap=-;

   matrix
   HinfluNPxx  AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCTTAACACATG
   HinflKW20   AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCTTAACACATG
```

If there are problems getting a "new" phylogeny program to work, one can first try to match an implementation of the program to the correct operating system (Windows, Mac, Linux). The next step will be to look at the data format and try to optimize. It is always a good idea to run a small test data set and even better to run example datasets included with the program. Two main problems often realized with data formats are that the output of the multiple alignments is interleaved but the program only reads non-interleaved and the other that the output of multiple alignments is with line breaks but the program only works without line breaks. Different solutions can be tried. In ClustalW/X it is possible to specify different output formats. For instance, in ClustalX, this can be specified under *Alignment* (top menu bar) and then selecting *Output Format Options*. Here is Clustal shown as default; however, for phylogenetic applications, PHYLIP, NEXUS, and FASTA are also relevant.

The final hard way is "manual" editing. Search and replace (space with nothing) from text editor. Lines can be joined with the UNIX text editor, vi. If these procedures are to be used more frequently, it should be considered to write scripts for editing.

6.8 Phylogenetic Program Packages

Several hundreds of packages are available, but only a few are widely used (■ Table 6.3). The classical package PHYLIP has been used for decades. It includes programs for variants of parsimony analysis, distance matrix analysis, tools to make consensus trees, as well as maximum likelihood analysis. For more advance maximum likelihood analysis, packets like fastDNAml, PhyML, FastTree, and RAxML are available. PAUP is the classical package for parsimony analysis, and MEGA7 can both perform the main phylogenetic analysis and analysis needed in population genetics.

6.9 Activities

6.9.1 Neighbor Joining Phylogeny

Install ClustalX on your PC from ► ftp://ftp.ebi.ac.uk/pub/software/clustalw2/2.0.10/.
Download ► clustalx-2.0.10-win.msi, install, and locate the icon to the desktop.
Mac users should use the big file (9.5 Mb).
Do multiple alignment with **ClustalX** (if not already done in ► Chap. 4).
Open the program by double-clicking on the icon.

File | Load sequences and select the sequences in FASTA format from the location on your computer. The sequences that you want to use as input should be in one file only with all sequences in FASTA format:

```
>sequence1
ATGACGATAC...
>sequence2
GATAGATAGACS...
etc.
```

Alignment | Do complete Alignment | ok

Choose Trees | Draw Tree and then **Trees | Bootstrap N-J tree**. In both cases confirm by "ok" if output files are stored where you expect. The output is automatically saved in the same folder as the input file, and this folder should have write access (if you define 1000 bootstrap replicates (default), the output will be numbers of 1000 and not %).

To draw the tree, you need to install **MEGA7**.

From ► http://www.megasoftware.net/, press **Download** at **Windows** (or what operating system you prefer). Write the required information, download to PC, install the program, and locate the icon to the desktop.

Open **MEGA7** by the icon.

User Tree | Display Newick Trees, browse to the directory where you stored the ClustalX input/output, and choose the file with extension "***.phb**" (or any other extension). At this step the tree should be graphically presented with bootstrap values.

Mark "**Place root on branch**" tool to the left and select an out-group.

Manipulate the tree by "**Swap**" or "**Flip**" to root the tree with the out-group and to bring the rest of tree in shape. Often rotation around a branch horizontally can give better

order. It is also possible to adjust the graphics such as line thickness and fonts in **View | Options**.

Save the tree from **Image | Save as EMF**. The tree can be inserted into PowerPoint or Word by **Insert | Picture,** and select the file in emf-format. In PowerPoint you can manipulate text and lines in the tree further: mark the picture with the tree inserted, right click and select ungroup and then again ungroup, and then you should be able to move text boxes and lines as well as to edit the text.

Note: in the example above, ClustalX is used for multiple alignment and construction of the tree, whereas MEGA7 is only used to construct the tree. MEGA7 can also be used for the whole analysis including multiple alignment and tree construction.

Take-Home Messages

- A phylogeny is a model of the relationships between organisms, genes, proteins, and other structures based on common ancestry.
- An unrooted phylogenetic tree is constructed as a system of lines where the free ends represent actual sequences compared and the ends joined by bifurcations represent hypothetical ancestors.
- The most important information read from a phylogenetic tree is the location of the monophyletic groups.
- A monophyletic group has one and only one common branch, whereas paraphyletic groups have two common branches.
- The basic assumptions about data in order to perform phylogenetic analysis are that sequences compared are homologs meaning that they shared (or are thought to have shared) a common ancestor and that they can be aligned.
- Model parameters required for phylogenetic analysis include the substitution matrix, the tree shape (topology), and the weights put on different positions in the multiple alignment to account for variable and conserved regions.
- The most frequently used statistical principles for phylogenetic methods to construct unrooted trees are parsimony, neighbor joining, maximum likelihood, and Bayesian.
- The robustness of a phylogenetic tree given a dataset can be evaluated by bootstrap analysis which is a comparison of simulated trees based on simulations with replacement of the original multiple alignment.

References

Camin, J. H. & Sokal, R. R. 1965. A method for deducing branching sequences in phylogeny. Evolution 19, 311–326.

Cavalli-Sforza, L. L. & Edwards, A. W. F. 1967. Phylogenetic analysis: Models and estimation procedures. Am. J. Hum. Gen. 19, 233–257.

Darwin, C. 1859. On the origin of species by means of natural selection or, the preservation of favoured races in the struggle for life. 1th ed. (reprinted 1998), Wordsworth, Ware.

Eck, R. V. & Dayhoff, M. O. 1966. Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Spring, Maryland.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. Science 20, 155(3760), 279–84.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59:307–321.

Haeckel, E. 1875. Ziel und Wege der heutigen Entwichklingsgeschichte. Jena, Hermann Duft.
Hennig, W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zenteralverlag. Berlin.
Hennig, W. 1966. Phylogenetic Systematics. Univ. Illinois Press, Urbana.
Hillis, D. M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44, 3–16.
Huson, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14, 68–73.
Kumar, S., Stecher, G. & Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 3. 1870–1874.
Lamarck, J. B. 1809. Zoological Philosophy: An exposition with regard to the natural history of animals. Translated by H. Elliot. Macmillan, London 1914. Reprinted by University of Chicago Press, 1984.
Nei, M. & Kumar, S. 2000. Molecular Evolution and Phylogenetics. Oxford.
Olsen, G.J., Matsuda, H., Hagstrom, R. & Overbeek, R. 1994. fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10, 41–48.
Peplies, J., Kottmann, R., Ludwig, W., Glöckner, F.O. 2008. A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. Syst. Appl. Microbiol. 31, 251–257.
Price, M. N., Dehal, P. S., & Arkin, A. P. 2010. FastTree 2 -- Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5: e9490.
Saito, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–25.
Sneath, P. H. A. and Sokal, R. R. 1973. Numerical taxonomy. Freeman, San Francisco.
Stamatakis, A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies". Bioinformatics 30:1312–1313.

Further Reading

Felsenstein, J. 2004. Inferring Phylogenies 2nd Edition. Sinauer, Sunderland.

Sequence-Based Classification and Identification of Prokaryotes

Henrik Christensen and John Elmerdahl Olsen

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_7

What You Will Learn in This Chapter

You will be introduced to the classification of prokaryotes from species to the higher levels of class and phyla. You will learn that the 16S rRNA gene sequence comparison including phylogenetic analysis provides the main characteristics for classification. For control of the 16S rRNA classification, DNA-DNA hybridization and more recently increasing use of multilocus sequence analysis on the whole genomic level are used. You will learn how DNA-DNA hybridization can be performed in silico by comparison of whole genomic sequences. In the activity you will learn to identify an isolate by 16S rRNA sequence in the EzBioCloud server.

7.1 Introduction

Bioinformatics is important for classification and identification since all prokaryotes have been classified based on the phylogeny of the 16S rRNA gene sequence. This gene sequence has also to great extent been the gold standard for identification. Now whole genomic sequences are increasingly used for classification and identification. The analysis of 16S rRNA gene sequences and also whole genomic sequences involves bioinformatics described in many other chapters of this book (Fig. 7.1). Unfortunately we cannot cover the classification and identification of microfungi and other eukaryotic microorganisms in this chapter; however, the same principles apply to these organisms as well.

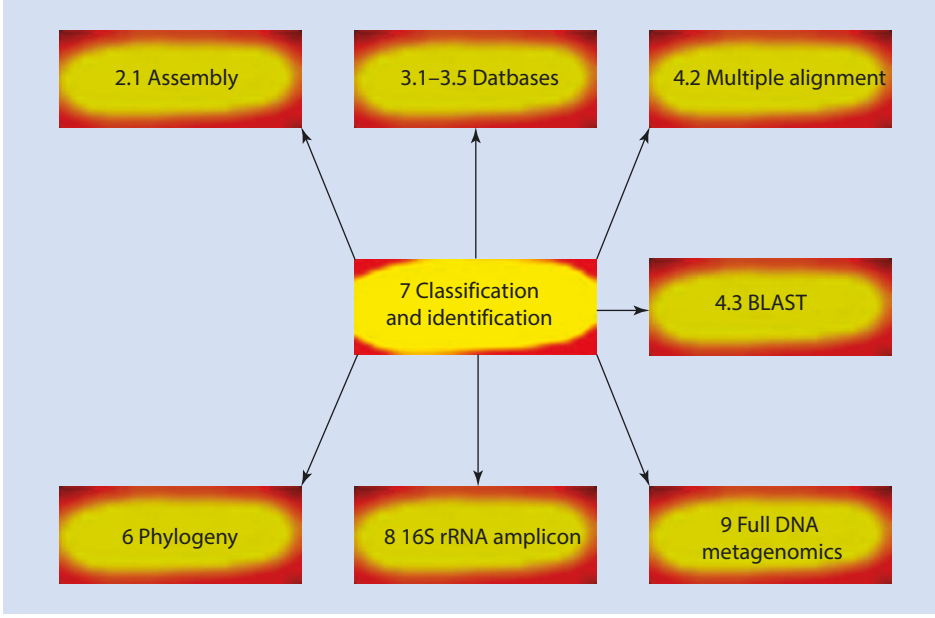


Fig. 7.1 Relationships between sequence-based classification and identification and the other chapters of the book. DNA sequence-based classification involves assembly of sequences, multiple alignments, phylogeny, and search in the databases. 16S rRNA-based identification also involves 16S rRNA amplicon sequencing described in the next chapter. Full DNA metagenomics both include information from the 16S rRNA gene sequence and from all other genes

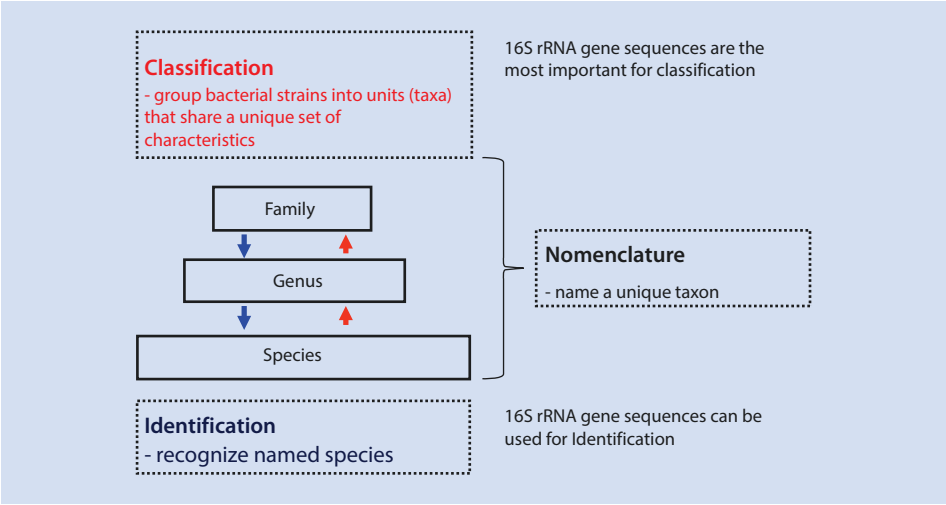


Fig. 7.2 Taxonomy of prokaryotes

The phenotype is still important in classification. New species cannot be classified and named without at least two phenotypic differences to existing species. Phenotypic properties include such diverse characteristics as micromorphology, physiology, biochemistry, and chemotaxonomy. Attempts are being made to predict the phenotype from the whole genomic sequence. Unfortunately only few phenotypic characteristics can be precisely predicted from the gene sequences.

Identification is the part of taxonomy that associates a bacterial strain with a species name (Fig. 7.2). The traditional biochemical and physiological methods used for identification of bacteria have serious limitations with respect to accuracy, and they are very time-consuming. Although fast high-throughput versions of these methods are available, there are still problems with the accurate identification of prokaryotes including some human pathogens and food contaminants. Especially prokaryotes of veterinary importance have been underrepresented in the databases of the fast phenotypic identification systems.

The comparison of 16S rRNA gene sequences has been the first choice to identify isolates that have been problematic to identify by other means. Other gene sequences, for instance, of the *rpoB* gene have often been the choice to identify problematic isolates belonging to species with narrow 16S rRNA gene sequence relationships (Adékambi et al. 2009; Korczak et al. 2004; Case et al. 2007).

7.2 Classification of Prokaryotes

7.2.1 Classification Based on 16S rRNA Gene Sequence Comparison

16S rRNA sequence comparison is the primary character to classify bacteria at species and genus level as well as into the higher taxonomic categories (family, order, class). The reason for the choice of the 16S rRNA gene is that it forms the backbone of the ribosomes and

ribosomes are of only one type in all organisms. The comparison of the 16S rRNA sequence can therefore be used to model the natural relationships and evolution of bacteria. rRNA can be sequenced (DNA) and allow separation of all bacteria to genus (and species) level. For the most there is less than 97% 16S rRNA gene sequence similarity between species and less than 95% 16S rRNA gene sequence similarity between genera. As mentioned some species are closer related than 97% 16S rRNA gene sequence similarity, and for them other genetic methods such as DNA-DNA hybridization need to be performed to confirm species status in regard to classification (see above).

Members of Archaea are so different in their 16S rRNA genes that they need other primer sets. The most simple starting point is a colony of an isolate from an agar plate transferred under sterile conditions to the tube with the PCR mix. For some bacteria, DNA needs to be purified before the PCR. The traditional starting point is PCR amplification of the 16S rRNA gene with conserved primers that in principle can be used for all Bacteria (Table 7.1). Some of the PCR product is used for agarose gel electrophoresis to confirm that a PCR product has been generated and that fragments are of expected size (Table 7.1). The rest of the DNA in the PCR tube is purified to reduce the content of PCR primers and sent to a sequencing company with the sequencing primers (Table 7.1) (Edwards et al. 1989), or you can do it yourself if equipment is available. Traditional Sanger sequencing is the most robust choice in the moment although high-throughput methods such as **PacBio** and **Nanopore** should provide full-length 16S rRNA gene sequences on the genomic level. The DNA sequence can be assembled by a relevant computer program (see ▶ Chap. 2). It is important to assemble the consensus sequence from reads made both by forward and reverse sequencing primers. When the consensus sequence is generated, it is stored in text format and used as a query for comparison to all "subjects" in a database and be used further for multiple alignments and phylogeny (▶ Chaps. 4 and 6).

Full-length 16S rRNA sequence can often not be extracted from the genomic sequence if it has not been fully closed. The 16S rRNA gene is in multiple copies which limit the full assembly of the whole genomic sequence. In this case the 16S rRNA gene sequence needs to be determined by traditional Sanger sequencing (Chun et al. 2018).

7.2.2 Classification Based on DNA-DNA Hybridization Between Total DNA

The way a mixture of denatured DNA preparations from two different bacterial strains hybridize can be used to classify the strains at species level. The degree of DNA-DNA reassociation (DDN) also referred to as DNA-DNA hybridization or DNA-DNA binding is measured by heating the DNA preparations for the two strains to be compared. Heating will denature the double-stranded DNA to single-stranded DNA. Then the DNA preparations are mixed in equal amounts and slowly cooled. The DNA will then start to renature to double-stranded DNA. The DNA of the mixture of the two strains will renature less perfectly compared to DNA from the same strain. The DNA-DNA hybridization between strains of the same species will be more than 70%, whereas the DNA-DNA hybridization between strains of the different species will be less than 70%. DNA-DNA hybridization or the alternative methods are used as a control to classify species when their 16S rRNA gene sequence similarity is higher than 97%. Traditional DNA hybridization is time-consuming and costly, and alternative procedures have been used to estimate.

Table 7.1 Primers for 16S rRNA PCR and sequencing (Sanger)

Primer	Application	Sequence 5′ – 3′	References
The classical set for Bacteria			
8-27f[a]	PCR and sequencing	AGAGTTTGATCCTGGCTNAG	Edwards et al. (1989) and Weisburg et al. (1991)
1390-1408r	PCR and sequencing	TGACGGGCGGTGTGTACAA	Lane et al. (1985)
518-536r	Sequencing	GTATTACCGCGGCTGCTGG	Lane et al. (1985)
785-805r	Sequencing	GACTACCNGGGTATCTAATCC	Dewhirst et al. (1992)
785-805f	Sequencing	GGATTAGATACCCNGGTAGTC	
Alternatives for Bacteria			
5f	PCR and sequencing	TTGGAGAGTTTGATCCTGGCTC	Simmon et al. (2006)
810r	Sequencing	GGCGTGGACTTCCAGGGTATCT	Simmon et al. (2006)
1194f	Sequencing	ACGTCATCCCCACCTTCCTC	Simmon et al. (2006
rD1r	PCR and sequencing	AAGGAGGTGATCCAGCC	Edwards et al. (1989) and Weisburg et al. (1991)
16SUNIr	PCR and sequencing	GTGTGACGGGCGGTGTGTAC	Lane et al. (1985) and Kuhnert et al. (2002)
Archaea			
SDArch0333aS15f	PCR and sequencing	TCCAGGCCCTACGGG	Lepp et al. (2004)
SDArch0958aA19r	PCR and sequencing	YCCGGCGTTGAMTCCAATT	Barns et al. (1994) and DeLong (1992)
SDArch1378aA20r	PCR and sequencing	TGTGTGCAAGGAGCAGGGAC	Lepp et al. 2004

f forward, *r* reverse
[a]Numbering follows *E. coli rrnB* (NCBI/DDBJ/EBI acc. no. J01696)

Whole genome sequence comparison has become a new kind of standard to estimate DDN. The use of whole sequence comparison to separate species was first formulated as average nucleotide identity (ANI) (Konstantinidis and Tiedjem 2005) and later modified by Goris et al. (2007). This index is based on the BLAST (► Chap. 3) program

with certain rules for the fragment length compared and for the sorting of output. The program compares fragments of 1020 nt from one genome against all fragments of similar length in the other genome. It is then calculated how many fragment with at least 30% identity and match in length of at least 70% can be identified out of total fragments compared. Investigations of correlation to DDN have showed that an ANI of 95% corresponds to the conventional 70% DDH cutoff (Richter and Rosello-Mora 2009). For practical applications, an ANI range of 93–96% should "be treated cautiously" (Rosselló-Móra and Amann 2015), and species should only be separated based on this very narrow interval if DDN data are supported by other very stable phenotypic characteristics.

Later the tool GGDC (genome-to-genome distance calculator) has been setup on a server with public access offering the same opportunities as ANI (► http://ggdc.gbdp.org/). The principle behind this program is also based on BLAST; however, this program uses the comparison of high-scoring sequence pairs (HSP) (► Sect. 4.3) for the comparison, and DDN can be estimated. Therefore the output from GGDC is sometimes referred to as digital DDN (dDDN).

JSpecies (► http://www.imedea.uib.es/jspecies/download.html) (Richter and Rosello-Mora 2009) has facilities for ANI using both BLAST and MUMmer approaches (■ Table 7.2). The MUMmer approach is only for closed genomes. With the BLAST approach, non-closed genomes can be used; however, the use of more than 50% of the genome is recommended.

The GC content is of taxonomic value in the way that species of the same genus normally will not differ more than 5% in GC content. However, different genera can have the same GC content. It is standard to determine the GC content for the type strain of the type species of a genus. The GC content can be determined based on the whole genomic sequence which is easier and more precise than previous biochemical methods like HPLC or buoyant density centrifugation methods (Kim et al. 2015) (■ Table 7.3).

Minimal standards have been formulated for the requirements to the quality of whole genomic sequences to be used for taxonomic studies (■ Table 7.3). The coverage used to determine the assembled sequence should be a minimum of 50. Information about genome size, GC content, number of contigs, and N_{50} need to be provided in publications including this type of information. It is a prerequisite that the assembled genomic sequence is deposited with a public database (► Chap. 3).

■ **Table 7.2** Bioinformatic tools to predict DNA-DNA renaturation based on whole genomic sequences

Name	Content	References	URL
JSpecies	ANI, MUMmer	Richter and Rosello-Mora (2009), Goris et al. (2007)	► http://imedea.uib-csic.es/jspecies/
GGDC	GGDC	Auch et al. (2010a, b)	► http://ggdc.dsmz.de/

Table 7.3 Proposed minimal standards for use of genomic taxa for taxonomy of prokaryotes (Chun et al. 2018)

Parameter	Minimum requirements to documentation in scientific publications
GC content	Needs to be stated
Genome size	Needs to be stated
Number of contigs	Needs to be stated, limit depends on taxon
N_{50}	Needs to be stated, limit depends on taxon
Coverage	More than 50
Deposition of sequence	Assembled DNA sequence
Number of conserved genes for phylogeny	More than 30

7.2.3 Classification Based on Multilocus Sequence Analysis (MLSA)

For classification based on phylogenetic relationships, many conserved genes can be used to complement the 16S rRNA analysis. Multilocus sequence analysis (MLSA) is based on DNA sequences of conserved genes which are then concatenated meaning that they are joined end by end. MLSA has been used for control of 16S rRNA classification during the past two decades based on up to ten genes. The increased availability of whole genomic sequences has allowed the comparison of full gene sequences of in principle all conserved genes or protein sequences of a taxonomic unit. For whole genomic MLSA, a lower limit of 31 genes or protein has been proposed (Table 7.3). MLSA for classification is used on species level and higher taxonomic levels. In ► Chap. 11, we will see that MLSA also can be used below the species level for typing.

7.3 Classification of the Taxonomic Hierarchy

7.3.1 Classification of Species

The species represent the main taxonomic unit in classification. Species include prokaryotes with 16S rRNA gene sequence similarities higher than 97.0%. If 16S rRNA sequence similarity higher than 97% between different species are found, then DNA-DNA hybridization or comparable methods have been used to confirm the diversity of such species since members of a species also need to be related with more than 70% DNA reassociation (Tindall et al. 2010). Whole genomic sequence comparison to group species at the genotypic level is now replacing experimental DNA-DNA hybridization (see ► Sect. 7.2.2).

The comparison of housekeeping gene sequences has recently been used for genotypic definition of species. Based on the work of Zeigler (2003), Kuhnert and Korczak (2006) proposed that sequence comparison of the three genes *recN*, *rpoA*, and *thdF* could replace DNA-DNA hybridizations. They further showed that even sequencing of a single gene, *recN*, might serve this purpose.

7.3.2 Classification of Genera

The rules and recommendations are less strict compared to species. As mentioned, usually less than 95% 16S rRNA gene sequence similarity have been determined between genera (Yarza et al. 2008). In metagenomic analysis based on partial 16S rRNA gene sequence, a 96% cutoff has been practiced (Ley et al. 2008). In parallel to ANI for nucleotides, there is also an ANI based on protein comparison (Qin et al. 2014). The percentage of conserved proteins (POCP) formulated by Qin et al. (2014) has been used to verify the classification at genus level. Different genera should have less than half of their proteins in common according to that definition.

7.3.3 Classification of Families, Orders, Classes, and Phyla

The rules and recommendations are again less strict for the higher *hierarchical* units of families, orders, classes, and phyla. Sometimes families have been created to find a home for a new genus which again has created a new order. Analysis of 16S rRNA gene sequence showed a minimum of 87.5% identity between type strains of type species of type genera of different families (Yarza et al. 2008) (Fig. 7.3). There are no 16S rRNA gene sequence limits; however, families, orders, classes, and phyla should

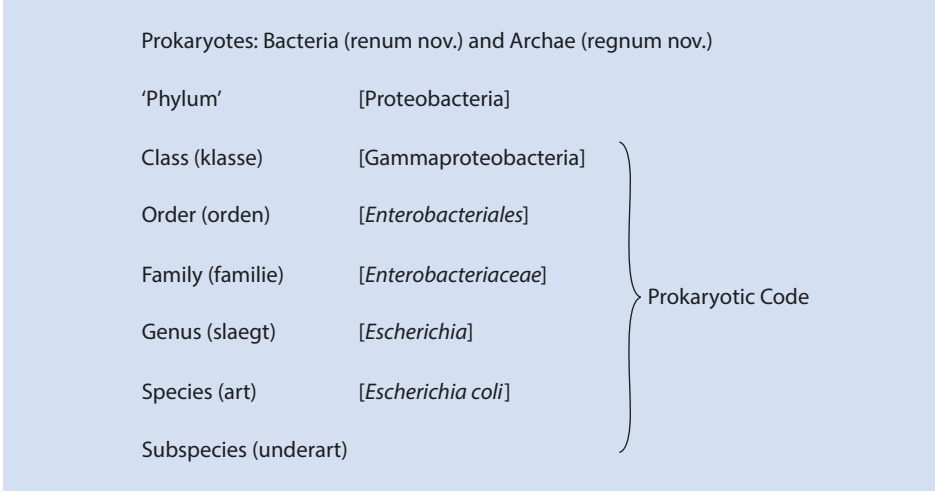


Fig. 7.3 Taxonomic categories of prokaryotes. The ones covered by the International Code of Nomenclature of Prokaryotes are validated names written in italics font and they are universal in all scientific publishing. For new classes the naming should follow the genus (ICNP rule 8). There is a proposal to include phylum as a validated name. Examples are included in brackets

be monophyletic (see ▶ Chap. 6). Work is in progress to define families, orders, classes, and phyla based on whole genomic comparison using the MLSA approach with a weight on the conserved genes. Names of families, orders, and class are covered by the International Code of Nomenclature of Prokaryotes, whereas phyla are not. For new names of families, orders, and classes, the type taxon has to carry the genus names. This has not been made retrospective in order not to cause too much confusion (Escheriaceae, Escherichiales, Pseudomonadia instead of *Enterobacteriaceae*, *Enterobacteriales*, and *Gammaproteobacteria*) (Oren et al. 2015, 2016).

7.4 Rules for the Naming of a New Prokaryote

The way prokaryotes (Bacteria and Archaea) are named is a consequence of International Code of Nomenclature of Prokaryotes (Parker et al. 2015). There are rules for naming of prokaryotes but not for their classification. Names published in *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) are directly validly names, whereas names validly published outside IJSEM can become validly named if accepted for the validation list (VL) in IJSEM. All scientific publishers follow this nomenclature, and they are used as identifiers in databases including those dealing with bioinformatics.

Very few species are homotypic synonyms (two different names can be used for the same species). *Klebsiella mobilis* and *Enterobacter aerogenes* have the same type strain and therefore are homotypic synonyms, and as a consequence either name can be used. Users of the names seem to favor *Enterobacter aerogenes* with 48,200 hits in Google Scholar (▶ https://scholar.google.dk/) compared to only 6860 for the other. Each of the species has its own circumscription, and one can use the name that fits the circumscription one finds most correct. A heterotypic synonym is when the same taxon that has two names but different type strains. In this case the name first published will normally have priority.

Prokaryotic scientific names are put in italic font since they are derived from Latin. They are binary combinations of a genus followed by a single specific epithet. The taxonomy of prokaryotes was originally derived from botanical rules; however, independent rules have existed for bacteria for nearly 100 years. Only names included with the *Approved Lists of Bacterial Names* in 1980 (Skerman et al. 1980) are valid. Names after 1980 are on List of Prokaryotic names with Standing in Nomenclature (LPSN) (▶ http://www.bacterio.net/) (Parte 2014).

7.4.1 Bacterial Species Names Are Linked to the Type Strain

The type strain is one and only one well-characterized strain from the species. To secure loss and access to the strain, it needs to be deposited in at least two culture collections with public access. Such cultures are copies of the same strain but have different strain designations in the form of a letter and number code. The type strain is a "name bearer," and it follows the species even when it is transferred to a new genus. For example, when *Streptococcus faecalis* (type strain ATCC 19433) was transferred to *Enterococcus faecalis* (type strain ATCC 19433), the type strain was unchanged. The type strain may have lost virulence properties due to repeated transfers in culture collections, and it is mostly only a good reference for taxonomic purposes. The type strain should always be included as a reference for classification and identification.

7.4.1.1 Example of an Old Bacterial Name that Never Changed

Staphylococcus aureus was validly published by Rosenbach 1884, and it is the type species of the genus *Staphylococcus*. The name is valid since it was included in the *Approved Lists of Bacterial Names* (Skerman et al. 1980). The type strain is ATCC 12600 = ATCC 12600-U = CCM 885 = CCUG 1800 = CIP 65.8 = DSM 20231 = HAMBI 66 = NCAIM B.01065 = NCCB 72047 = NCTC 8532. The different numbers reflect the deposition of the same strain in different culture collections with public access (List of Prokaryotic names with Standing in Nomenclature).

7.4.1.2 Example of a Taxon with Many Reclassifications and Changes in Genus Name

Brachyspira hyodysenteriae was reclassified from *Treponema* by Ochiai et al. (1997). It was originally named *Treponema hyodysenteriae* (Harris et al. 1972). This name is its basonym, meaning first name (Harris et al. 1972). The name was included with the *Approved Lists* in 1980 (Skerman et al. 1980). In between it was named *Serpula hyodysenteriae* by Stanton et al. (1991) since it was reclassified from *Treponema hyodysenteriae*. This name had to be changed 1 year later to *Serpulina hyodysenteriae* by Stanton (1992) since *Serpula* is illegitimately being used for a fungal genus *Serpula*. In between, *Brachyspira* had been named by Hovind-Hougen et al. (1982) and had to be used instead of *Serpulina* since *Brachyspira* was an earlier synonym of *Serpulina*. Despite of the changes, the type strain of the species was always the same: B78 = ATCC 27164 = CCUG 46668 = NCTC 13041 (List of Prokaryotic names with Standing in Nomenclature) (► http://www.bacterio.net/) (Parte 2014).

7.5 The Benefits of Sequence-Based Identification

The comparison at DNA sequence level is definitive. DNA sequences can be compared on global scale and handled as electronic public information that for many applications substitute the analysis of live strains. Besides the applications of DNA sequences for identification to the species level, DNA sequence-based identification can be useful for molecular epidemiology and population genetics (see ► Chap. 11), and annotations of DNA sequences can be used to predict phenotypic properties.

7.5.1 16S rRNA Sequence-Based Identification, Step-by-Step

The principles for sequencing of the 16S rRNA gene was outlined in ► Sect. 7.2.1. When the 16S rRNA gene sequence has been determined, comparison in EzBioCloud (Yoon et al. 2017 (► https://www.ezbiocloud.net/) allows identification by comparison to all type strains of bacterial species with a validated name (► Activity 7.6). Often such comparison is more precise than comparison in GenBank by BLAST search since the status of strains used to generate the sequence in GenBank is often not known. A phylogenetic 16S rRNA sequence-based analysis can be needed if similarity is obtained to more closely related

species (see ► Chap. 6). The combination of database search and phylogenetic analysis can be used to improve the identification of an isolate to species and genus level. Matrix-assisted laser desorption ionization (MALDI) time-of-flight mass spectrometry (TOF MS) has recently become an alternative method for fast and cheap identification of bacteria. Current limitations are with limited databases and limited resolution between some species. 16S rRNA sequence identification is a gold standard to make the MALDI-TOF database.

7.5.2 16S rRNA-Based Identification Without Culture

In the late 1980s, it became possible to characterize prokaryotes from their natural habitat without culture. The rRNA was extracted, reverse transcribed to DNA, cloned, and sequenced (Ward et al. 1990). Later, PCR was used to amplify the 16S rRNA genes from environmental DNA, and it was then cloned and sequenced. Recently high-throughput sequencing technology has been combined with 16S rRNA target regions V1–V3 to deep sequence environmental samples (see ► Chap. 8). The technology is part of metagenomics and probably the most rapidly evolving field of microbiology. The current drawback with 16S rRNA-targeted metagenomics is the short-read length providing low accuracy for species-level identification.

7.6 Activity

Sometimes you end up with a bacterial isolate that is difficult to identify by phenotypic means, or you do not believe in the result from the identification by a kit. Then you PCR amplify the 16S rRNA gene and sequence the DNA. After assembly of the sequence, you want to perform a database search to obtain the most accurate identification. Traditional nucleotide BLAST in, e.g., GenBank, can be performed. The problem is that so many sequences are deposited and you cannot always rely on the species name associated with a sequence. The optimal reference is the sequence of the type strain. However, it is not always so easy to locate in a BLAST output. To overcome this, a server has been set up by professor Chun and his team where a search is performed against 16S rRNA sequences of type strains only. For the activity we will download a 16S rRNA gene sequence from NCBI and assume it is unknown and recently determined in the lab. Download the sequence with acc. no. KX858032 from NCBI like described in ► Chap. 3. Use EzBioCloud (► http://www.ezbiocloud.net/identify) to identify the sequence. You need to register for a free account. Press *Identify single sequence* and paste the sequence in FASTA format into the window, leave other parameters default, and *Next* and then *Submit* when the sequence has been verified.

What is the top hit taxon and similarity %? Also click at "+" in the *Task column* to see alternative less likely identifications. In this activity the expected result is 96.1% similarity which is lower than the species threshold. Based only on the 16S rRNA-based identification, this strain would represent a new species.

Take-Home Messages

- Bioinformatics is important to classification and identification since all prokaryotes have been classified based on the phylogeny of the 16S rRNA gene sequence and this character also to a great extent has been the gold standard for identification.
- The classification of prokaryotes at the species level based on DNA-DNA hybridization can be estimated in silico by comparison of whole genomics sequences.
- Whole genomic sequences are increasingly used also for the classification of genera, families, orders, classes, and phyla by phylogenetic comparison of common core genes or predicted proteins.
- The *rpoB* gene sequence is frequently used for identification since it allows the separation of some species that cannot be identified by 16S rRNA gene sequence comparison.
- Identification of prokaryotes that cannot be cultured is possible by 16S rRNA gene amplicon sequencing which is further described in ► Chap. 8.

References

Adékambi, T., Drancourt, M. & Raoult, D. 2009. The *rpoB* gene as a tool for clinical microbiologists. Trends Microbiol 17, 37–45.

Auch, A. F., Von Jan, M., Klenk, H-P., Göker, M. et al. 2010a. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2: 117–134.

Auch, A. F., Klenk, H-P., Göker, M. 2010b. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Standards in Genomic Sciences 2: 142–148.

Barns, S. M., Fundyga, R. E., Jeffries, M. W. & Pace, N. R. 1994. Proc. Natl. Acad. Sci. *USA*91**,** 1609–1613.

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., Kjelleberg, S. 2007. Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. Appl. Environ. Microbiol. 73, 278–288.

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., Da Costa, M., Rooney, A. P., Yi, H., Xu, X.-W., De Meyer, S. & Trujillo, M. E. 2018. Minimal standards for the use of genome data for the taxonomy of prokaryotes. IJSEM 68, 461–466.

DeLong, E. F. 1992. Proc. Natl. Acad. Sci. USA 89**,** 5685–5689.

Dewhirst, F. E., Paster, B. J., Olsen, I., & Fraser, G. J. 1992. Phylogeny of 54 representative strains of species in the family Pasteurellaceae as determined by comparison of 16S rRNA sequences. J Bacteriol. 174, 2002–2013.

Edwards, U., Rogall, T., Blöcker, H., Emde, M., Böttger, E. C. 1989. Isolation and direct complete nucleotiwde determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. Nucleic Acids Res. 17:7843–53.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P. & Tiedje, J. M. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 57, 81–91.

Harris, D. L., Glock, R. D., Christensen C. R., Kinyon, J. M. 1972. Inoculation of pigs with Treponema hyodysenteriae (new species) and reproduction of the disease. Vet Med Small Anim Clin.67, 61–64.

Hovind-Hougen, K., Birch-Andersen, A., Henrik-Nielsen, R., Orhold, M., Pedersen, J. O., Teglbjaerg, P. S. & Thaysen, E. H. 1982. Intestinal spirochetosis: morphological characterization and cultivation of the spirochete *Brachyspira aalborgi* gen. nov., sp. nov. J. Clin. Microbiol. 16, 1127–1136.

Kim, M., Park, S. C., Baek, I., Chun, J. 2015. Large-scale evaluation of experimentally determined DNA G+C contents with whole genome sequences of prokaryotes. Syst Appl Microbiol. 38:79–83.

Konstantinidis, K. T. & Tiedjem J. M. 2005. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. 187, 6258–64.

Korczak, B., Christensen, H., Emler, S., Frey, J. & Kuhnert, P. 2004. Phylogeny of the family *Pasteurellaceae* based on *rpoB* sequences. Int J Syst Evol Microbiol. 54, 1393–1399.

Kuhnert, P., Frey, J., Lang, N. P. & Mayfield, L. 2002. Phylogenetic analysis of Prevotella nigrescens, Prevotella intermedia and Porphyromonas gingivalis clinical strains reveals a clear species clustering. Int J Syst Evol Microbiol. 52, 1391–5.

Kuhnert, P. & Korczak, B. M. 2006. Prediction of whole genome DNA-DNA similarity, determination of G+C content and phylogenetic analysis within the family Pasteurellaceae by multilocus sequence analysis (MLSA). Microbiology 152, 2537–2548.

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., Pace, N. R. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A. 82:6955–9.

Lepp, P. W., Brinig, M. M., Ouverney, C. C., Palm, K., Armitage, G. C. & Relman, D. A. 2004. Methanogenic *Archaea* and human periodontal disease. PNAS 101, 6176–6181.

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., Schrenzel, M. D., Knight, R., Gordon, J. I. 2008. Evolution of mammals and their gut microbes. Science 320, 1647–51.

Ochiai, S., Adachi, Y. & Mori, K. 1997. Unification of the genera Serpulina and Brachyspira, and proposal of Brachyspira hyodysenteriae Comb. Nov., Brachyspira innocens Comb. Nov. and Brachyspira pilosicoli Comb. Nov. Microbiol. Immunol. 41, 445–452.

Oren, A., da Costa, M.S., Garrity, G.M., Rainey, F.A., Rosselló-Móra, R., Schink, B., Sutcliffe, I., Trujillo, M. E., Whitman, W. B. 2015. Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes. Int J Syst Evol Microbiol. 65, 4284–7. doi: https://doi.org/10.1099/ijsem.0.000664.

Oren, A., Parte, A., Garrity, G. M. 2016. Implementation of Rule 8 of the International Code of Nomenclature of Prokaryotes for the renaming of classes. Request for an Opinion. Int J Syst Evol Microbiol. 66, 4296–8.

Parker, C. T., Tindall, B. J. & Garrity, G. M. 2015. International Code of Nomenclature of Prokaryotes. Int. J. Syst. Evol. Microbiol. In press.

Parte, A. C. 2014. LPSN--list of prokaryotic names with standing in nomenclature. Nucleic Acids Res. 2014:D613-6. doi: https://doi.org/10.1093/nar/gkt1111. Int J Syst Evol Microbiol. 2016 Oct;66(10):4296–4298. doi: https://doi.org/10.1099/ijsem.0.001319. Epub 2016 Jul 21.

Qin, Q. L., Xie, B. B., Zhang, X. Y., Chen, X. L., Zhou, B. C., Zhou, J., Oren, A., Zhang, Y. Z. 2014. A proposed genus boundary for the prokaryotes based on genomic insights. J. Bacteriol. 196, 2210–2215.

Richter, M. & Rosello-Mora, R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. PNAS 106, 19126–31

Rosselló-Móra, R., & Amann, R. 2015. Past and future species definitions for Bacteria and Archaea. Syst Appl Microbiol. 38:209–16.

Rosenbach, F. J. 1884. Microorganismen bei den Wund-Infections-Krankheiten des Menschen. J.F. Bergmann, Wiesbaden, pp. 1–122.

Skerman, V. B. D., McGowan, V. & Sneath, P. H. A. (editors) 1980. Approved Lists of Bacterial Names. Int. J. Syst. Bacteriol. 30, 225–420.

Simmon, K. E., Croft, A. C. & Petti, C. A. 2006. Application of SmartGene IDNS software to partial 16S rRNA gene sequences for a diverse group of bacteria in a clinical laboratory. J. Clin. Microbiol. 44, 4400–6.

Stanton, T. B., Jensen, N. S., Casey, T. A., Tordoff, L. A., Dewhirst, F. E. & Paster, B. J. 1991. Reclassification of Treponema hyodysenteriae and Treponema innocens in a new genus, Serpula gen. nov., as Serpula hyodysenteriae comb. nov. and Serpula innocens comb. nov. Int. J. Syst. Bacteriol. 41, 50–58.

Stanton, T. B. 1992. Proposal to change the genus designation Serpula to Serpulina gen. nov. containing the species Serpulina hyodysenteriae comb. nov. and Serpulina innocens comb. nov. Int. J. Syst. Bacteriol., 42, 189–190.

Tindall, B. J., Rosselló-Móra, R., Busse, H. J., Ludwig, W., Kämpfer, P. 2010. Notes on the characterization of prokaryote strains for taxonomic purposes. Int. J. Syst. Evol. Microbiol. 60, 249–266.

Ward, D. M., Weller, R. & Baterson, M. M. 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 345, 63–65.

Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 1991. 16S ribosomal DNA amplification for phylogenetic study. J. Bact. 173, 697–703.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K. H., Ludwig, W., Glöckner, F. O., Rosselló-Móra, R. 2008. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol. 31, 241–50.

Yoon, S. H., Ha, S. M., Kwon, S., Lim, J., Kim, Y., Seo, H. & Chun, J. 2017. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol. 67, 1613–1617.

Zeigler, D. R. 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int J Syst Evol Microbiol 53, 1893–1900.

16S rRNA Amplicon Sequencing for Metagenomics

Henrik Christensen, Anna Jasmine Andersson, Steffen Lynge Jørgensen, and Josef Korbinian Vogt

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_8

What You Will Learn in This Chapter

The experimental setup to use of 16S rRNA amplicon sequencing will be introduced. The different steps in the bioinformatical analysis will be explained with background in the major pipelines mothur and QIIME 2. We have extracted examples from mothur in the text, and we have chosen to demonstrate QIIME 2 in more detail in ► Activity 8.10.1. You will be able to get a good start with the analysis of 16S rRNA amplicon data with QIIME 2.

8.1 Use of 16S rRNA Amplicon Sequencing, Generation of Data, and Bioinformatics Pipelines

8.1.1 Use of 16S rRNA Amplicon Sequencing and Generation of Raw Data

The 16S rRNA gene sequence is used to classify all prokaryotes as described in ► Chap. 7, and 16S rRNA amplicon sequencing is an application of the 16S rRNA sequence-based classification concept to characterize biodiversity and to investigate the ecological characteristics of all sample types. The benefit of 16S rRNA amplicon sequencing is that comparison can be made to the extremely detailed and well-curated taxonomic databases. The 16S rRNA amplicon approach to metagenomic analysis started in 2007–2008 by a synergy between the use of high-throughput pyrosequencing analysis by the "454" methodology, bar coding, and the existing framework of 16S rRNA classification of the prokaryotes. The brilliant idea was to amplify a variable region of the 16S rRNA gene from the whole microbiome of a sample and then to barcode samples allowing sequencing of multiple samples in one run and this way to reduce the cost with the "454" method. The theoretical background for this concept was first described by Liu et al. (2007), and later the use was published from the group of professor Rob Knight (Liu et al. 2008; Hamady et al. 2008; McKenna et al. 2008). The combination of high throughput related to "454"- and later Illumina sequencing, relatively low cost related to the use of barcoding, and the high information gained by coupling to the huge 16S rRNA taxonomy databases has revolutionized microbiology, and the technique is now in use in most labs dealing with the metagenomics characterization of diverse environments – from environmental samples to the intestinal microbiome (◘ Fig. 8.1).

The procedures used for DNA extraction are dependent on the type of sample material. For the extraction of DNA from fecal samples and similar complex organic matrices, we can recommend MO BIO PowerMag® Soil DNA Isolation Kit (Qiagen) and FastDNA Spin Kit (MP Biomedicals). We refer the reader to Anderson et al. (2016) for an overview of all lab works related to 16S rRNA amplicon sequencing. Further descriptions of sample preparation and the procedures for DNA sequencing will unfortunately not be described here since the focus will be on the bioinformatical analysis.

Mock samples can be included in the analysis. They are constructed from known prokaryotic isolates pooled at even concentrations. The mock sample is included as a control sample in downstream analysis.

The primers used for PCR amplification depend on the desired degree of variability and length of product. Usually the higher variability of the 16S rRNA gene is near the 5′ end (◘ Table 8.1). Primers are constructed with adaptors of ca. 50 bp which are needed to attach the DNA amplicons to the solid support of the flow cell (◘ Fig. 8.2). In addition,

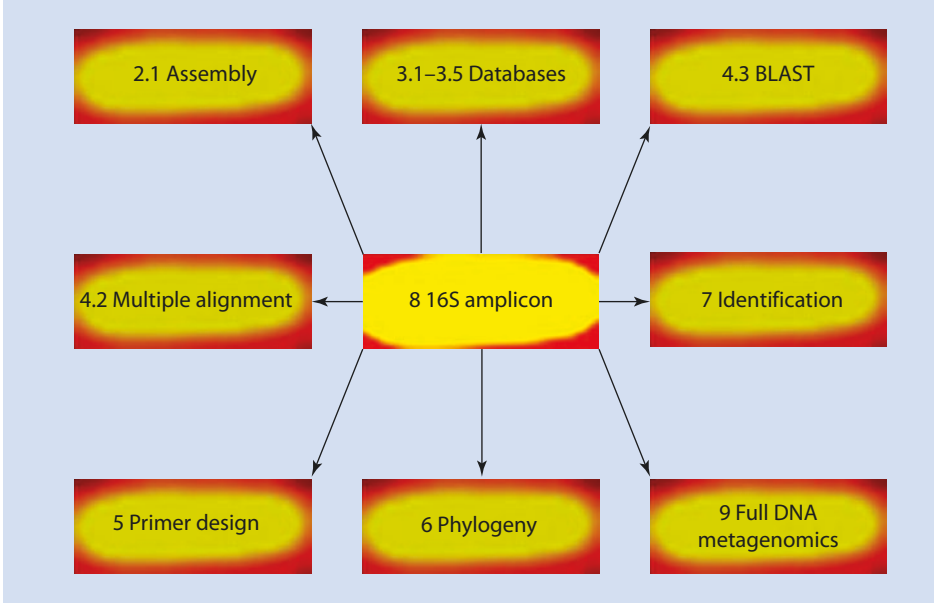


Fig. 8.1 16S rRNA amplicon sequencing relates to nearly all other topics in the book

Table 8.1 Primers for 16S rRNA metagenomics[a]

Target	Forward 5′-3′	Reverse 5′-3′	Size[c]	Reference
V3[b]	CCTACGGGAGGCAGCAG	ATTACCGCGGCTGCTGG	194	Lane (1991); Danzeisen et al. (2011)
V1–V3	AGAGTTTGATCCTGG	ATTACCGCGGCTGCTGG	527	Lane (1991);
				Danzeisen et al. (2011)
V3–V4	GGAGGCAGCAGTRRGGAAT	CTACCTGGGTATCTAATCC	457	Nossa et al. (2010)
V4–V5	GTGYCAGCMGCCGCGGTA	CCCGYCAATTCMTTTRAGT	413	Tang et al. (2014)

[a]The paper of Soergel et al. (2012) presents an exhaustive list of primer combinations and comparisons
[b]Information on V1-V9 (Ashelford et al. 2005)
[c]Referring to acc. no. J01695 of *E. coli rrnB*

barcodes on the 16S rRNA amplicons (multiplexing) are needed in order to include many samples on one flow cell. The final product is called the 16S rRNA amplicon sequencing library. Forward adaptors each with a unique index combined with similar reverse adaptors will give 96 unique indexes meaning that up to 96 independent samples can be sequenced at one time. Bar coding is further outlined in Hamady et al. (2008). Actually, MiSeq allows up to 400 16S rRNA amplicon libraries (50,000 reads per sample) in a single

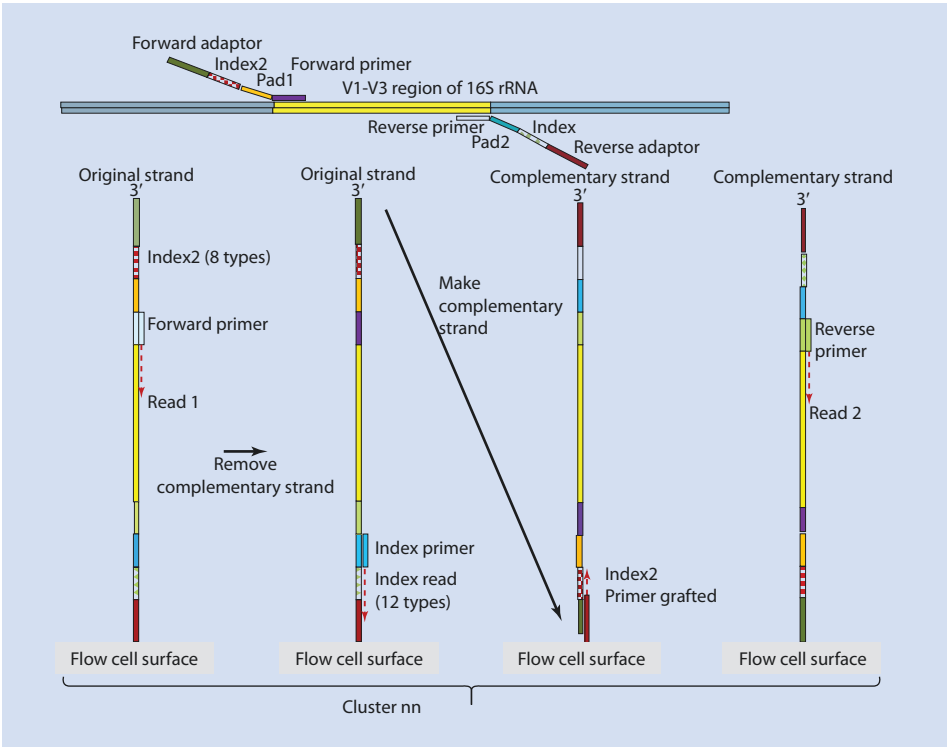


Fig. 8.2 Use of adaptor (fixing the sequence to the flow cell) and index primers (separation of different samples) for Illumina platform sequencing of a genomic DNA fragment which includes the V1-3 region of 16S rRNA gene

sequencing run. As a rule of thumb, there should at least be 50,000 reads per sample and more than 100 reads per bacterium of interest. The main limitation with MiSeq sequencing is the short sequence length of the 16S rRNA gene. Attempts have been made to sequence nearly full-length 16S rRNA genes with PacBio to improve the information content (Schloss et al. 2016; Wagner et al. 2016; Karst et al. 2018).

This chapter focuses on the characterization of the prokaryotes. The same concept can be used with microfungi and other eukaryotic microorganisms; however, other primers need to be used for the initial PCR amplification.

8.1.2 Bioinformatical Pipelines to Analyze Data

The microbiome pipelines allow users to analyze raw DNA sequence data, produce results, apply statistics, and visualize results for publications. A common theme in the analysis of 16S rRNA amplicon sequence data is that all samples can be analyzed at one time as well as the differences between samples can be compared. The two options are possible, thanks to the use of the barcoding of all reads analyzed.

Mothur (Schloss et al. 2009) (► https://www.mothur.org/) was developed to satisfy the need to analyze high-throughput sequence datasets in a modular way and to speed up the existing algorithms. The program is written in C + +. The most simple way to use the

package is from the command (dos) prompt of Windows PC. All material is included in one folder including the raw data and program. The program is started with the mothur icon. The commands can then be copied from the SOP available from the website and full analysis of 16S rRNA amplicon libraries carried out in a series of steps. The program keeps track on all read files and sample types in the "Stability.files." You need to edit this file with a text editor to get started.

QIIME (quantitative insights into microbial ecology) (pronounced chime) (Caporaso et al. 2010) (► http://qiime.org) is a pipeline for analysis of high-throughput sequence data in microbiological ecology studies. From January 1, 2018, QIIME 2 is the supported version of this pipeline. QIIME 2 is based on the computer language Python and is set up with a series of modules which allow the user to handle, manipulate, and analyze the large multiplexed high-throughput datasets.

In this chapter, we will mainly refer to mothur in the text and work with QIIME 2 in the activity. Other equally important pipelines are available such as USEARCH (Edgar 2010) and UPARSE (Edgar 2013). The latter includes quality filtering, trimming, merging of identical reads (dereplication), and clustering. VSEARCH (Rognes et al. 2016) is an open-source version of USEARCH. Sometimes modules from pipelines are combined. For instance, mothur is using the chimera filter from VSEARCH. QIIME 2 makes it possible to easily run VSEARCH via plug-ins.

CLC Genomics Workbench (Qiagen) provides a module for 16S rRNA microbiome analysis. It is recommended for users without past experience. The program will carry out a standard analysis and is a good way to start and save time in the steep learning that is required in this field. Unfortunately the software is licensed and relatively expensive; however, the cost probably pays off in time. The Microbial Genomics Module needs to be added to the CLC Genomics Workbench package. The Help function will provide guidance. The module is accessed from the top menu bar. The following steps should be followed starting with raw data: trim, merge paired reads, trim to fixed length, filter samples based on number of reads, OTU clustering, α- diversity, and β-diversity. Finally, PCoA analysis is offered, however, with limited functionality. The software allows the user to perform a standard 16S rRNA microbiome analysis with a set of samples. For further use, readers are recommended to contact the company's hotline.

8.2 Data Analysis

8.2.1 Quality Trim by Sequence

Denoising is quality filtering of individual reads to remove low-quality reads, and it can also include removal of chimeras (this will be discussed in ► Sect. 8.3), foreign DNA, and individual bases with low-quality scores. In UPARSE, a limit of Phred score (see ► Sect. 2.2.1) was set at 16 (Edgar 2013). Trimming is invoked if two or more adjacent low-quality base calls are identified. If some paired reads are longer than the expected size based on the PCR primers (■ Table 8.1), the sequences that are too long are excluded for analysis (**scree.seqs command, maxlength=**) in mothur.

At this step, some reads will be completely removed for further analysis, and others will be trimmed in length removing the low-quality regions.

8.2.2 Pairing of Reads

At this initial step, forward and reverse paired reads are paired to contigs. With mothur, this is done with the "make.contigs" command. USEARCH also includes this procedure.

8.3 Removal of Chimeras and DNA from Other Domains or Life

Chimeras are biased merged read sequences from two origins. These sequences can artificially increase the diversity of the samples and need to be identified and removed during the bioinformatical analysis. In addition to the tools included in mothur and QIIME 2 pipelines, chimeras can be removed by UCHIME (Edgar et al. 2011).

Filters can be included to remove eukaryotic DNA from mitochondria and chloroplasts and DNA from Archaea if one is only interested in Bacteria.

```
(remove.lineage in Mothur).
```

8.4 Grouping of Reads into OTUs

Dereplication is the grouping of identical reads (**unique.seqs** in Mothur). UPARSE also includes this procedure. Singletons were recommended to be discarded (Edgar 2013).

Reads are grouped into OTUs with a 97% similarity threshold (**cluster.split** in **Mothur**). The 97% limit is based on the 16S rRNA threshold between prokaryotic species (▶ Chap. 7). OTU clustering is also used to improve precision of the analysis since the error of sequencing can be as high as 0.1% per nucleotide in a single read.

8.5 Alignment of OTUs and Association of OTUs with Taxonomic Units

First, a multiple alignment (▶ Chap. 4) is performed to further reduce noise and prepare the dataset for phylogenetic analysis. The multiple alignment will include a mask that allows comparison with the specific formats of the databases. Three main databases, SILVA (Quast et al. 2013), Greengenes (DeSantis et al. 2006; McDonald et al. 2012), and RDP (Cole et al. 2014), are commonly used.

The SOP of mothur will guide you through how to align to the SILVA database. In QIIME 2, MAFFT and FastTree are used for multiple alignment and phylogenetic analysis, respectively. Greengenes is used as the standard database for alignment as described in ▶ Activity 8.10. The QIIME 2 tutorial also explains the eventual need to train a classifier tailored for sample preparation and sequencing parameters, including the primers that were used for amplification as well as the length of the sequenced reads.

Variations between databases have been found to associate OTUs with taxonomic units. The highest assignment success on genus level was obtained with MiDAS-SILVA, whereas Greengenes came second, and RDP was least efficient. However, MiDAS-SILVA required more computing time compared to RDP (Popp et al. 2017).

Unidentified OTUs in the databases are not identified further. The short length of sequence makes further characterization of such unassigned OTUs very difficult.

8.6 α- (Within Group) and β-Diversity (Between Groups) Comparison

Before these indices can be calculated, normalization is needed. The numbers of sequences in all samples are normalized to include the same number. This number is usually set as the number of sequences in the samples with the lowest number.

For α-diversity analysis, different comparisons can be used. In QIIME 2, the choices are Faith's phylogenetic diversity (Faith 1992), Pielou's evenness, observed OTUs, and Shannon's diversity (further explained in ▶ Activity 8.10, ◘ Fig. 8.16). Pielou's evenness is a measure for how much the different species in a sample resembles in terms of numbers, i.e., a low value (scale 0–1) meaning less evenness and more dominant species and vice versa. Qualitative measurements of community richness are Faith's phylogenetic diversity, where the phylogenetic tree is incorporated, and observed OTUs. In Shannon's diversity index for quantitative measure, both richness and evenness are taken into account.

For β-diversity analysis, UniFrac can be used in both mothur and QIIME 2. UniFrac requires a phylogenetic tree to have been constructed, which is why this module is included in both pipelines. UniFrac is based on comparison of the branch length fraction of distances of shared and unshared branch lengths in the phylogenetic tree between the compared samples. A UniFrac distance of zero suggests that the compared samples are identical. There are two types of UniFrac distance metrics – unweighted and weighted. Unweighted UniFrac measures the presence and absence of taxa/OTUs, whereas weighted UniFrac also incorporates the abundances of taxa/OTUs. Bray-Curtis calculates the dissimilarity in the samples composition but is not based on phylogeny.

8.6.1 Rarefaction Analysis

This analysis is performed to see if some samples are represented with low sequencing depths. Based on the analysis, a cutoff read depth is chosen for the following alpha diversity analysis. Samples below a certain threshold are not included in further analysis (◘ Figs. 8.3, 8.4, and 8.16).

The chosen cutoff read depth should be at the state when the slope of the diversity curve approaches 0, which indicates that the maximum diversity is represented at the given read depth.

The phyloseq R package (McMurdie and Holmes 2012) can produce rarefaction curves with different α-diversity metric units, including observed number of OTUs, Shannon diversity index, Simpson diversity index, and Chao1 (◘ Fig. 8.3).

Initial analysis using rarefaction curves can give a visual overview and comparison of the quality of the samples. Samples which demonstrate significantly low sequence depth or α-diversity can be identified and excluded if necessary.

◘ Figure 8.4 demonstrates the rarefaction curves of four different microbiota treatments, using three different α-diversity metric units: observed no. of OTUs, Shannon diversity index, and Simpson diversity index.

The figure demonstrates that the samples have different maximum read depth, with the lowest at app. 20,000 reads and the highest exceeding 50,000 reads. Hence the highest cutoff value can be 20,000 read depth if all samples are included.

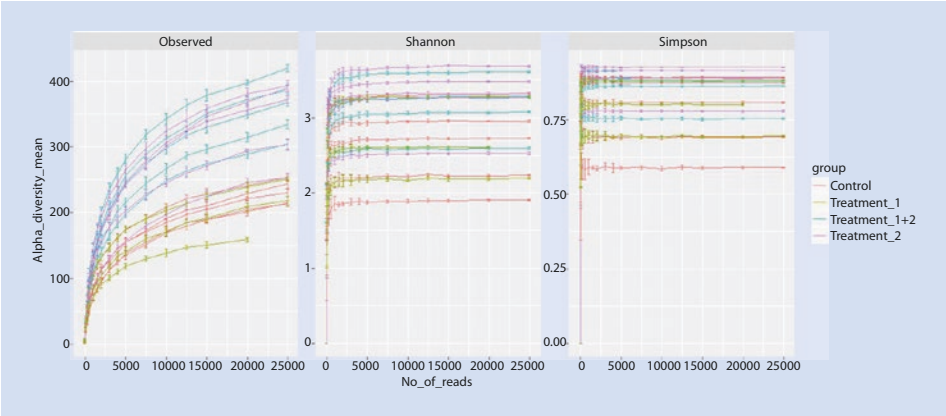


Fig. 8.3 Rarefaction curves for all samples from four different groups, using three alpha diversity metrics: observed richness, Shannon diversity index, and the Simpson diversity index. All plots are visualized at 25,000 read depth

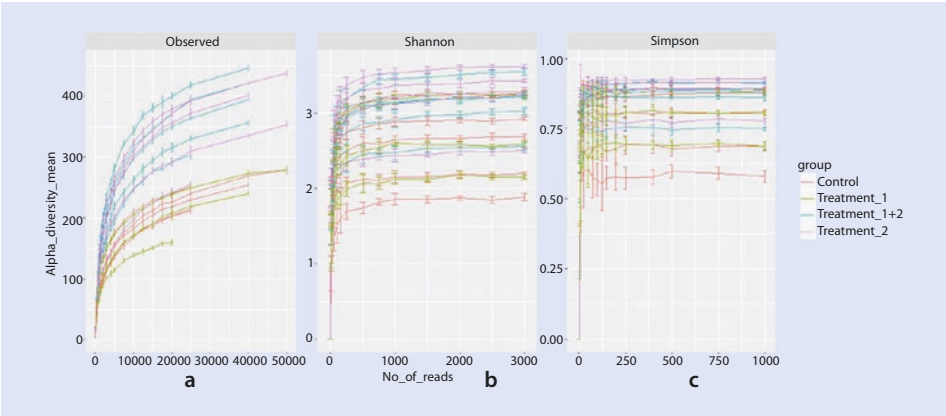


Fig. 8.4 Rarefaction curves with different maximum read depth to demonstrate the slope of the curves

The different α-diversity metrics demonstrate slopes reaching 0 at different read depths. Figure 8.4 visualizes the same dataset, however, with different read depths depending on the α-diversity metric.

While the observed richness (Fig. 8.4a) never reaches a maximum, both the Shannon diversity index curve and Simpson diversity index curve (Fig. 8.4b, c) reaches asymptotes at read depths <1500 reads.

Both the Shannon diversity index and Simpson diversity index demonstrate that overall α-diversity do not significantly change by increasing the read depth after 1500 reads. However, one might suggest that the full α-diversity is not reached even at a read depth of 50,000 reads due to the stump observed richness curves, which demonstrate an increase in identified OTUs when increase read depth.

Similar results can also be visualized when producing boxplots of the alpha diversity at different read depths. The Ampvis2 package can produce boxplots at different read depths and alpha diversity metrics (Fig. 8.5).

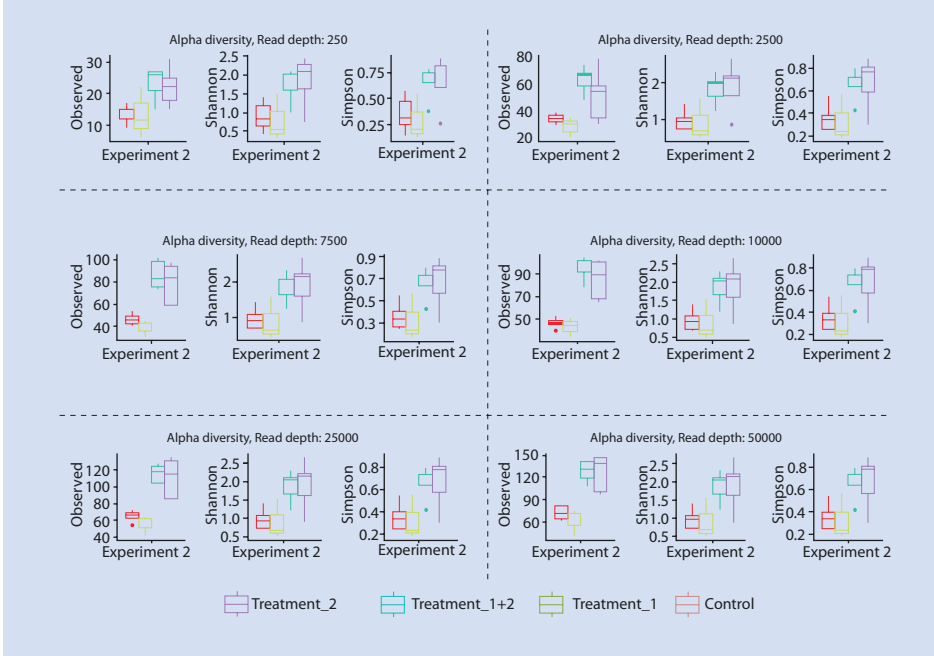


Fig. 8.5 Boxplots of the α-diversity represented by three α-diversity metrics: observed richness, Shannon diversity index, and the Simpson diversity index

Both the rarefaction curves and boxplots demonstrate that increasing the read depth only increases the observed richness but not the overall α-Shannon or Simpson diversity. Hence, these results suggest that all samples are suitable to be included in the further analysis.

These findings suggest that the significant OTUs are already represented at a read depth of 1500–2500 reads, and only OTUs which are only identified in very low and insignificant numbers are added at higher read depths. Especially the *Control* and *Treatment* 1 samples indicate a lower observed richness compared to the *Treatment* 2 and *Treatment* 1 + 2. Additionally, the Simpson diversity index, which is more weighted on evenness, indicates a more uneven diversity in the *Control* and *Treatment* 1. This could be due to an overrepresentation and high abundance of a few OTUs in the *Control* and *Treatment* 1 samples.

8.7 Taxonomic Comparisons

The number of OTUs for each best taxonomic match will provide the raw data for this analysis (**classify.otu** in **Mothur**). The output file from mothur can be opened in Microsoft Excel.

8.7.1 Generation and Interpretation of Heatmaps and Boxplots

Heatmaps and boxplots (Figs. 8.6 and 8.7) can generate an overview of the samples taxonomic makeup. Simple heatmaps of the most abundant OTUs can be useful to

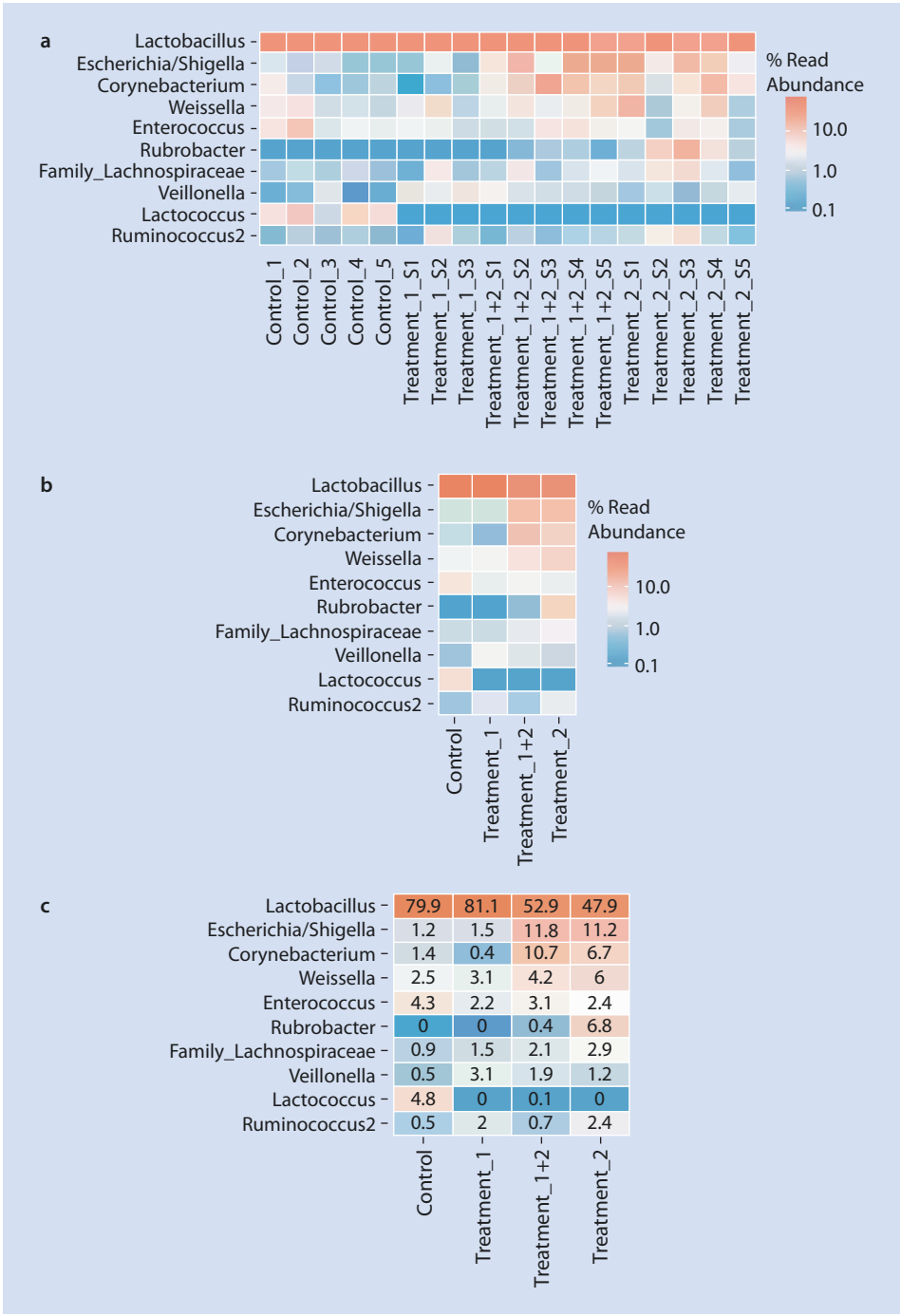


Fig. 8.6 Heatmap of the ten most abundant genera in the analyzed sample set. All three heatmaps are based on the same dataset and presented in three different ways. **a** Includes all samples, with the % read abundance represented by the color temperature of the cell. **b** Each treatment group has been summed up, and the average is represented in the heatmap. **c** Similar to B but with the average % read abundance presented in each cell

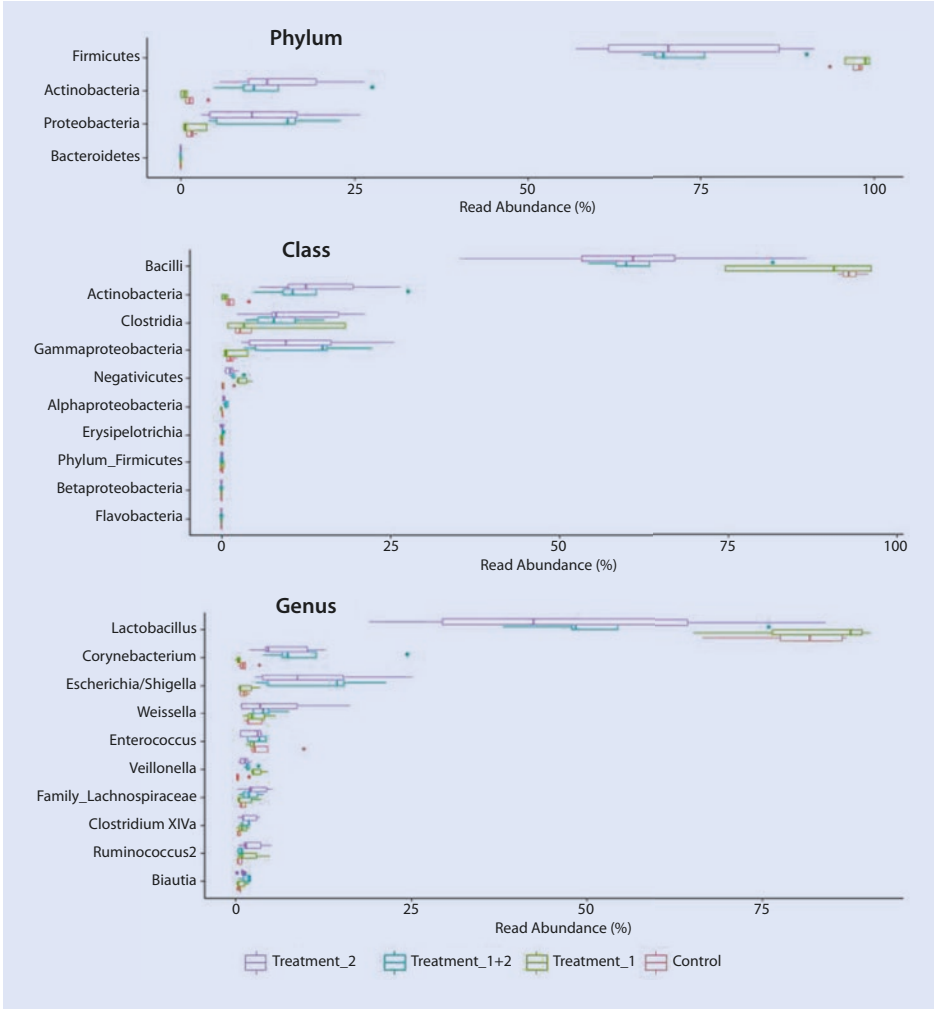


Fig. 8.7 Boxplots of the most predominant OTUs on phylum, class, and genera level in four different groups. All boxplots are based on the average of the samples of the respective group

identify if there are OTUs with very high abundance and for presenting results. The Ampvis2 package can produce heatmaps and boxplots that present a chosen number of OTUs at different taxonomic levels.

Similar to the heatmaps, boxplots are excellent to demonstrate the scale of the predominant OTUs in a sample set.

Both the heatmaps and boxplots can give an overview of the composition of OTUs and the predominant groups, as well as an initial idea of the microbiota similarity of the samples within the group as well as between the groups.

As demonstrated in both the boxplots and heatmaps, all four treatment groups demonstrate high abundance of Firmicutes, especially bacilli and specifically *Lactobacillus*. However, especially the Control and Treatment 1 groups demonstrate very high abundance of *Lactobacillus*, which agrees with the indication of overrepresentation of few

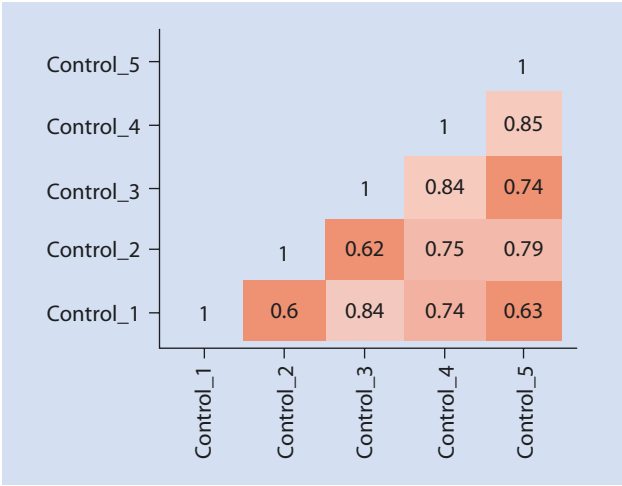


Fig. 8.8 Comparison of the five control samples based on Bray-Curtis similarity distance (1 = 100% similar, 0 = 100% dissimilar)

OTUs in the observed richness alpha diversity from the rarefaction curves and α-diversity boxplots (Figs. 8.3 and 8.4). These results further underlines that increasing the cutoff read depth would not improve the analysis in any way, as the only OTUs with insignificant abundance would be added to the analysis.

Both the boxplots and heatmaps suggest a more similar microbiota composition in Treatment_2 and Treatment_1 + 2 compared to the Control and the Treatment_1 groups. These results confer with the results from the α-diversity boxplots, which demonstrated signs of similar α-diversity in the Treatment_2 and Treatment_1 + 2 and the Control and Treatment_1 groups (Figs. 8.3 and 8.4).

Before β-diversity analysis such as PCoA (▶ Sect. 8.8), initial comparison of the similarity/dissimilarity of the samples within each group, based on non-metric or semi-metric distance, can give an indication of how uniform the samples of the group are. The Ampvis2 extra package offers a simple β-diversity comparison analysis using Bray-Curtis similarity distances (Fig. 8.8).

8.8 Principal Coordinates Analysis (PCoA)

Principal coordinates analysis (PCoA) is a tool to explore 16S rRNA amplicon sequence data visually by reducing a multidimensional sample vector (for 16S data, the feature counts for a sample) to principle components (PCs). PC can be understood as a combination of features capturing a certain degree of variance within all samples. By visualizing with a PCoA plot, sample clusterings can be observed, i.e., similarities or dissimilarities of the data (Fig. 8.9). QIIME 2 offers plotting samples in three dimensions. In the QIIME 2 activity later in this chapter, some of the diversity analysis results are shown as PCoA (Figs. 8.14 and 8.15).

Similar to the initial α-diversity analysis as well as the heatmaps and boxplots, the *Treatment*_2 and *Treatment*_1 + 2 demonstrate a similar microbiota composition, with overlaps in PC1, PC2, and PC3. Meanwhile, the *Control* and *Treatment*_1 groups also demonstrate an overlap in both PC1 and PC3.

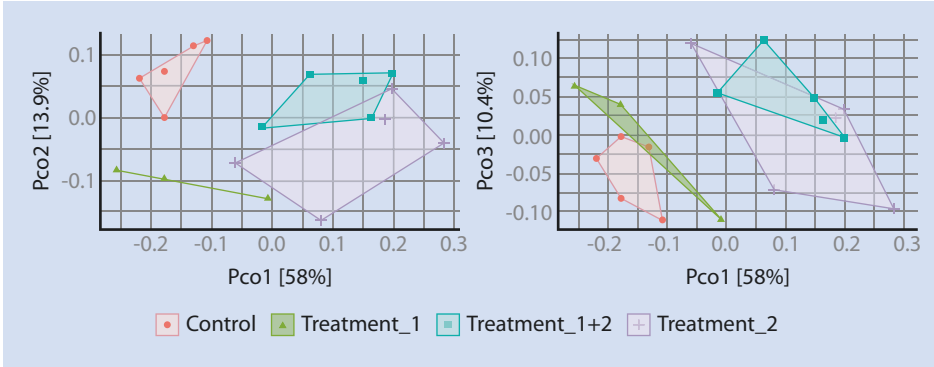


Fig. 8.9 PCoA of the samples from all four treatment groups, visualized by the first three PCs. To the left, PC1 and PC2 visualize 71.9% of the analysis model. To the right, PC1 and PC3 visualize 68.4% of the model

Collectively, these results indicate that Treatment 1 induces a similar microbiota as the Control group. Meanwhile, Treatment 2 induces distinct microbiota, different from the Control and Treatment 1. Both Treatment 1 and Treatment 2 simultaneously result in a microbiota similar to Treatment 2 alone, suggesting that Treatment 1 has no or a very weak effect on the microbiota.

To verify some of the conclusions drawn above, a permutational multivariate analysis of variance (PERMANOVA) is needed. This is a nonparametric multivariate statistical test, and it is used to compare groups of objects and test the null hypothesis that the centroids and dispersion of the groups as defined by measure space are equivalent for all groups.

8.9 Prediction of Function

It is possible to some extent to predict information about function based on comparison to taxonomic profiles. This is a shortcut to full DNA microbiome analysis (► Chap. 9). The drawbacks are that the same functions can often be undertaken by different and even diverse taxonomic groups. For activated sludge, a species taxonomic prediction of function is possible at: ► www.midasfieldguide.org.

PICRUSt (► http://picrust.github.io/picrust/) predicts abundance of gene families in host-associated and environmental communities (Langille et al. 2013).

8.10 Activity QIIME 2

QIIME is one of the major software packages for 16S rRNA microbiome analysis as described in ► Sect. 8.1.2. QIIME 1 was replaced with QIIME 2 in 2018 (current version: ► https://docs.qiime2.org/2018.2/), and it is the version to be installed when working with QIIME. There are continuously updated versions of QIIME 2, with improvements of the pipeline. The user needs to be aware of the version as it has to be specified while working with QIIME 2. The developers recommend the users to work with the latest version available.

There is an official forum for QIIME 2 where users can address questions about different problems encountered. It is a good idea to start searching this forum for questions and problems already discussed and solved before adding a new subject.

There are several tutorials on QIIME's website addressing various aspects of 16S rRNA microbiome analysis. We would recommend starting with the "Moving Pictures" tutorial. "Moving Pictures" was a time series study of the microbial communities of different anatomical sites of humans. This dataset is used as a case for the tutorial. The manual is best suited for macOS, and we recommend that you get access to a new Mac computer with a lot of RAM before starting. The subset of the Moving Pictures dataset includes only five time points and four anatomical sites totaling 20 samples. The tutorial will take you through most relevant analysis for 16S rRNA amplicon sequencing including alternative approaches for some analysis steps. You can use the tutorial to learn and understand 16S rRNA amplicon sequencing.

The commands used in this book can change depending on the version (current version 2018.2). All commands can, however, be found on the version specific documentation site.

8.10.1 Installation

If you try this for the first time, it is recommended to do exactly as the manual says even though the names for directories seem a little odd. You can copy the whole command strings from the QIIME 2 tutorial including the \ that will join lines on the command prompt. You can get the command prompt from **Go | Utilities** select **Terminal. App**. You can copy commands from the tutorial by using copy-paste functions.

Open the manual ► https://docs.qiime2.org/2018.2/, and select **Getting started** and **Install QIIME 2**. Select Natively **install QIIME 2 and install Miniconda and then QIIME 2** within a **Miniconda** environment. Select **macOS**. From the command prompt, run **source activate qiime2-2018.2** (or actual version).

If you are a Windows user, install MobaXTerm (serves as your terminal), and then use the manual Installing QIIME 2 using virtual machines.

8.10.2 Running QIIME 2

The first time you start QIIME 2, you will have to open the terminal window and write or copy-paste **source activate qiime2-2018.2** (or actual version). Then continue to do **exactly** as the tutorial says: **mkdir qiime2-moving-pictures-tutorial** and then the command **cd qiime2-moving-pictures-tutorial**. It will create the directory "qiime2-moving-pictures-tutorial" and place you in that directory.

The next time you start QIIME 2, you will have to open the terminal window and write or copy-paste **source activate qiime2-2018.2** (or actual version). Then continue with the **cd qiime2-moving-pictures-tutorial**.

There are three choices for downloading data in general in the tutorials. "Browser" clicking on the link and download data. "wget" and "curl" are utilities to download data conveniently via a terminal. Copy-paste the "wget" or "curl" command to your command prompt, and the data is saved in your directory.

Download URL: ► https://data.qiime2.org/2018.2/tutorials/moving-pictures/sample_metadata.tsv

Save as: **sample-metadata.tsv**

OR

```
wget-O "sample-metadata.tsv""https://data.qiime2.org/2018.2/tutori-
als/moving-pictures/sample_metadata.tsv
```

OR

```
curl -sL "https://data.qiime2.org/2018.2/tutorials/moving-pictures/
sample_metadata.tsv" > "sample-metadata.tsv"8.3. Download metadata
```

The metadata file contains information about the samples that are necessary for further analysis. For example, the categories included in the tutorial are Sample-ID, BarcodeSequence, and BodySite. If you analyze your own samples, this file has to be created to fit your own analysis. The file has to be in TSV format (tab separated values) and saved in your working directory. For more information about this topic, see "Metadata in QIIME 2" tutorial.

Choose one of the three alternatives for downloading the sample metadata. For the rest of this exercise, the curl command will be used.

8.10.3 Download Data

First, you need to make a subdirectory **emp-single-end-sequences** to the one you are standing in (qiime2-moving-pictures-tutorial). This is done by the same procedure as used in ► Sect. 8.10.2:

```
mkdir emp-single-end-sequences
```

Now, you download the data from the QIIME 2 homepage to this directory by the command:

```
curl -sL "https://data.qiime2.org/2018.2/tutorials/moving-pictures/
emp-single-end-sequences/sequences.fastq.gz" > "emp-single-end-
sequences/sequences.fastq.gz"
```

At this step, it can be a good idea to check if the files are located in your working directory.

```
cd emp-single-end-sequences
ls -l
```

Here, you should see the two files **barcodes.fastq.gz** at a size of around 3783785 bytes and **sequences.fastq.gz** of 25303756 bytes. Remember to navigate back to the qiime2-moving-pictures-tutorial directory by the command **cd**.

8.10.4 Change Data Format to QIIME 2 Artifacts

The imported data is multiplexed and has to be converted into "QIIME 2 artifacts" before continuing the analysis:

```
qiime tools import --type EMPSingleEndSequences --input-path emp-
single-end-sequences --output-path emp-single-end-sequences.qza
```

8.10.5 Demultiplexing Sequences

Demultiplexing is the step where the barcodes of all the single-end-reads are associated with their samples as described above.

```
qiime demux emp-single --m-barcodes-file sample-metadata.tsv --m-bar-
codes-column BarcodeSequence --o-per-sample-sequences demux.qza
```

Create a summary of the distribution of results from demultiplexing. The important information is the summary of the sequence quality at each base in the reads (interactive quality plot in demux.qzv artifact).

```
qiime demux summarize --i-data demux.qza --o-visualization demux.qzv
```

See the results with command:

```
qiime tools view demux.qzv
```

This command will invoke the browser and show the result in a browser window (**Firefox**). The graphics can be downloaded in pdf format. As the manual elaborates, you can visualize all similar graphics output by the command:

```
qiime tools view new_file.qzv
```

where new_file means that the file names will change for different outputs for the different analysis.

If your own project data is already demultiplexed, refer to ▶ Sect. 8.10.14.1 to proceed.

When working with paired-end reads, it is a good idea to merge the forward and reverse read; refer to ▶ Sect. 8.10.14.2.

8.10.6 Denoising

Denoising is quality filtering of individual reads to remove low-quality reads. In QIIME 2, one can choose to use DADA2 or Deblur. Here, DADA2 is used.

```
qiime dada2 denoise-single --i-demultiplexed-seqs demux.qza
--p-trim-left 0 --p-trunc-len 120 --o-representative-sequences
rep-seqs-dada2.qza --o-table table-dada2.qza
```

Then, rename the two files created with the command:

```
mv rep-seqs-dada2.qza rep-seqs.qza
mv table-dada2.qza table.qza
```

In this example, all reads are truncated to 120 nucleotides. The dataset is from the early 2010s when reads were shorter than the current 250 nucleotides. The truncation value of 120 is chosen based on the quality scores that dropped at this base number. To choose this value, inspect the demux.qzv artifact with the command **qiime tools view demux.qzv,** and in this visualization, view the interactive quality plot (Fig. 8.10).

8.10.7 Visualization Summaries of the Data

In the previous exercise you created FeatureTable (table.qza) and FeatureData (rep-seqs.qza) artifacts. Now, you want to view the content in these files. To do this, use the command below.


```
qiime feature-table summarize --i-table table.qza--o-visualization
table.qzv--m-sample-metadata-file sample-metadata.tsv
qiime feature-table tabulate-seqs --i-data rep-seqs.qza --o-visual-
ization rep-seqs.qzv
```

The output files are **table.qzv and rep-seqs.qzv** that can be viewed with **qiime tools view** as described earlier.

In table.qzv, there are three main summaries: overview, interactive sample detail, and feature detail (Fig. 8.11).

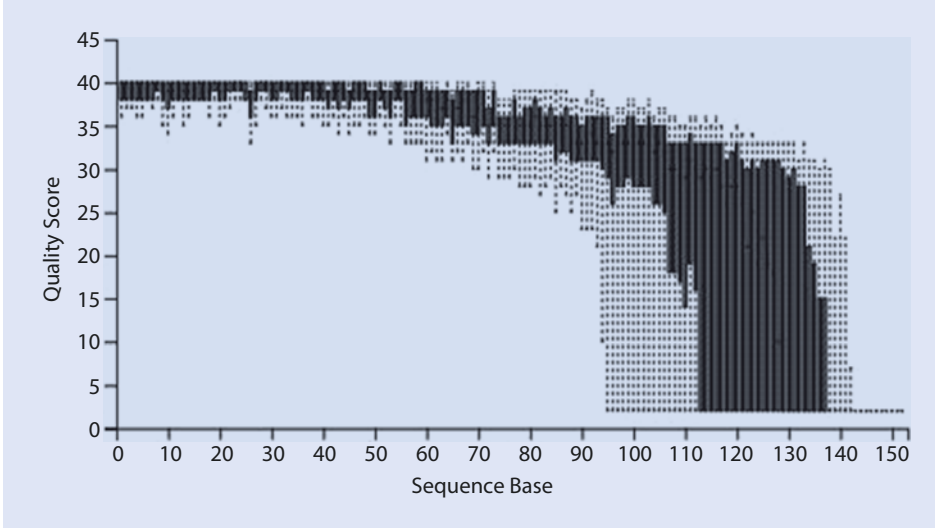


Fig. 8.10 Sequence quality control of the DNA sequences in DADA2: ▶ https://qiime2.org/

Overview | Interactive Sample Detail | Feature Detail

	Frequency	# of Samples Observed In
4b5eeb300368260019c1fbc7a3c718fc	11,497	13
fe30ff0f71a38a39cf1717ec2be3a2fc	9,092	16
d29fe3c70564fc0f69f2c03e0d1e5561	8,935	25
868528ca947bc57b69ffdf83e6b73bae	7,834	10
154709e160e8cada6bfb21115acc80f5	7,448	13
1d2e5f3444ca750c85302ceee2473331	7,270	23
0305a4993ecf2d8ef4149fdfc7592603	5,402	11
cb2fe0146e2fbcb101050edb996a0ee2	4,909	15
997056ba80681bbbdd5d09aa591eadc0	3,978	16
9079bfebcce01d4b5c758067b1208c31	3,364	11
3c9c437f27aca05f8db167cd080ff1ec	3,048	14
bfbed36e63b69fec4627424163d20118	2,508	14

Fig. 8.11 Feature detail from the artifact table.qzv shown below with feature IDs (OTUs) and their corresponding frequency and samples observed in ► https://qiime2.org/

8.10.8 Phylogenetic Tree

The α- and β-diversity metrics such as Faith's phylogenetic diversity and UniFrac use the phylogenetic tree, and therefore this tree has to be created before these diversity metrics can be performed.

```
qiime alignment mafft --i-sequences rep-seqs.qza --o-alignment
aligned-rep-seqs.qza
qiime alignment mask --i-alignment aligned-rep-seqs.qza --o-masked-
alignment masked-aligned-rep-seqs, qza
qiime phylogeny fasttree --i-alignment masked-aligned-rep-seqs.qza
--o-tree unrooted-tree.qza
qiime phylogeny midpoint-root --i-tree unrooted-tree.qza --o-rooted-
tree rooted-tree.qza
```

For the next exercise, the **rooted-tree.qza** artifact is used.

8.10.9 α- and β-Diversity Analyses

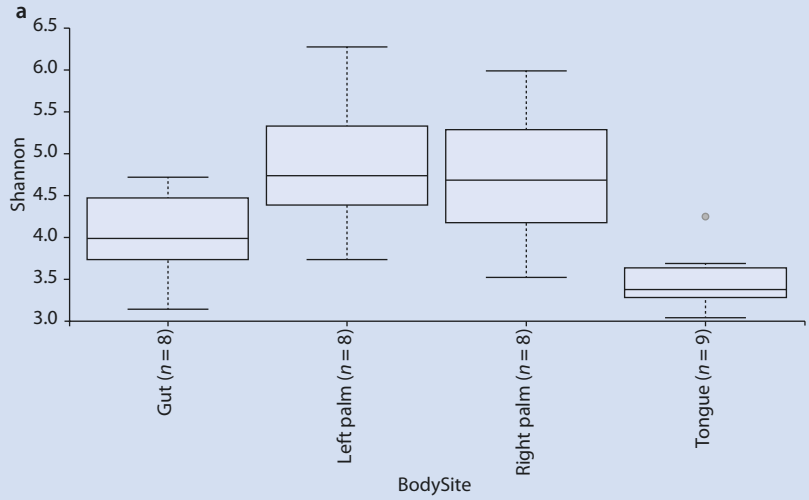
For the different diversity metric calculations, you have to choose the value for the **p-sampling depth** by viewing the **table.qzv** artifact (interactive sample detail). The number of reads in each sample has to be normalized, i.e., rarified. This is done by the command **--p-sampling-depth**. It can be difficult to choose this value. The chosen sampling depth should reflect the balance between capturing the alpha diversity in your samples

with high number of reads but low enough to not exclude too many samples with low depths. In this example, 1109 was used as threshold.

```
qiime diversity core-metrics-phylogenetic --i-phylogeny rooted-tree.
qza--i-table table.qza --p-sampling-depth 1109--m-metadata-file
sample-metadata.tsv --output-dir core-metrics-results
```

A new directory (core-metrics-results) is created, and it contains all the necessary files you need to explore for the α- and β-diversity (Figs. 8.12 and 8.13).

Shannon's diversity for category BodySites from the tutorial is shown as example (Fig. 8.12). The statistical test used in the boxplots for the different α-diversity metrics is



Kruskal-Wallis (all groups)

	Result
H	16.028817587641115
p-value	0.0011186613628180366

b

Kruskal-Wallis (pairwise)

Download CSV

Group 1	Group 2	H	p-value	q-value
Gut (*n* = 8)	**Left palm (*n* = 8)**	3.981618	0.045999	0.068999
	Right palm (*n* = 8)	2.481618	0.115184	0.138220
	Tongue (*n* = 9)	4.898148	0.026886	0.053771
Left palm (*n* = 8)	**Right palm (*n* = 8)**	0.099265	0.752714	0.752714
	Tongue (*n* = 9)	10.703704	0.001069	0.006415
Right palm (*n* = 8)	**Tongue (*n* = 9)**	8.898148	0.002855	0.008564

Fig. 8.12 Boxplot for Shannon's diversity index for category BodySite from the tutorial "Moving Pictures" in QIIME 2: ► https://qiime2.org/

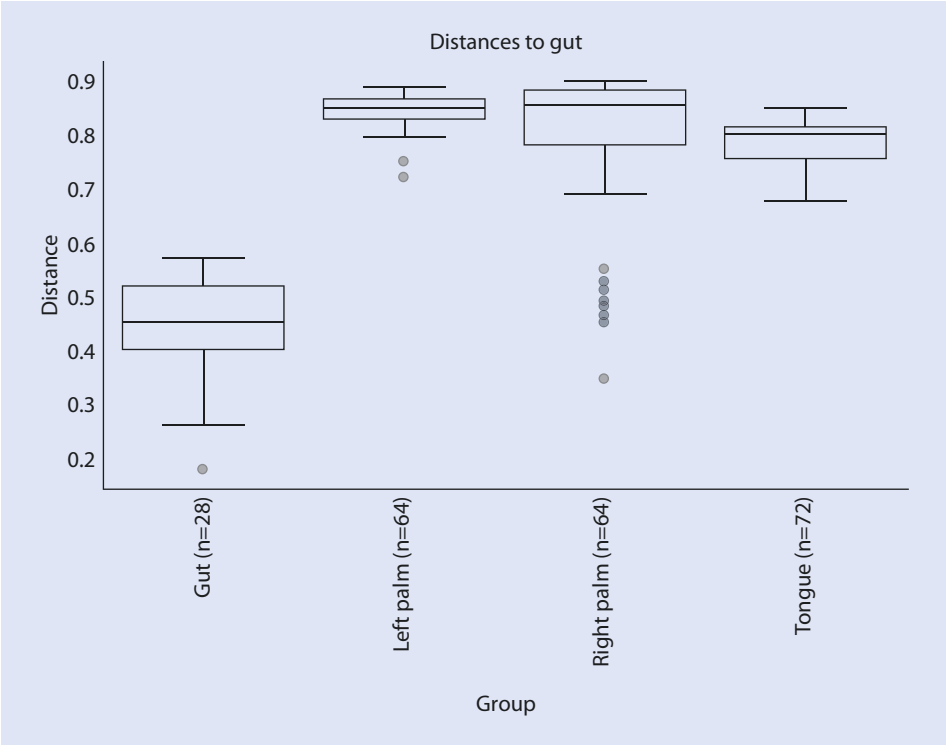


Fig. 8.13 Boxplot for β-diversity with weighted UniFrac as output example for category BodySites from the "Moving Pictures" tutorial in QIIME 2. PERMANOVA was used as statistical significance test (► https://qiime2.org/)

nonparametric Kruskal-Wallis test, with p-value at 0.05 considered as statistical significant. Shannon's diversity indexes were significant for BodySites ($p = 0.001$). The highest Shannon value, i.e., greater diversity, was observed for palms when compared to the tongue and gut. However, the gut/right palm comparison was not significant.

For β-diversity, weighted UniFrac and Bray-Curtis dissimilarity were used as example (Fig. 8.13). Bray-Curtis is a non-phylogenetic method using the abundance information of OTUs, where the distance metric is calculated by the species difference of abundance divided by total abundance in both samples. The PCoA plot shows relatively stable microbial communities during the time period (Figs. 8.14 and 8.15). PERMANOVA is used as significance test with p-value at 0.05 as statistical significant.

8.10.10 Alpha Rarefaction Plotting

This step in the analysis process is to explore if you have sequenced deep enough, i.e., captured the diversity in the samples. The rarefaction curve demonstrates that the sequencing depth (1109) was sufficient for almost all samples (Fig. 8.16). The sample with the highest sequence count was not sequenced deep enough as the curve is not plateauing as the rest of the samples. So for this sample, the richness would be underestimated for the top curve in red (Fig. 8.16).

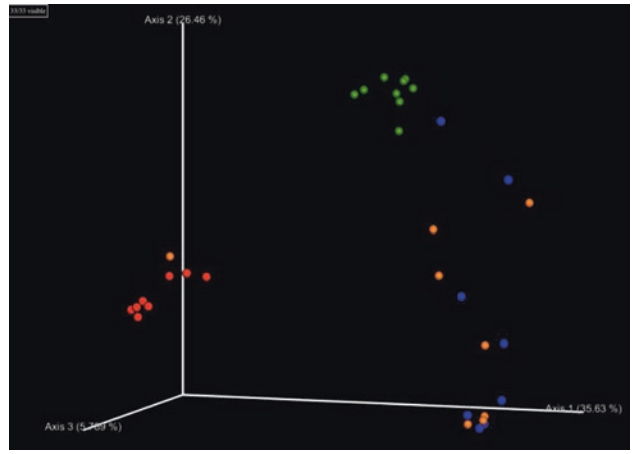


Fig. 8.14 Plots of PCoA based on UniFrac distance matrices of microbial communities in the gut, left palm, right palm, and tongue from the "Moving Pictures" tutorial in QIIME 2: ► https://qiime2.org/ (**Colors:** red = gut, blue = left palm, orange = right palm, green = tongue)

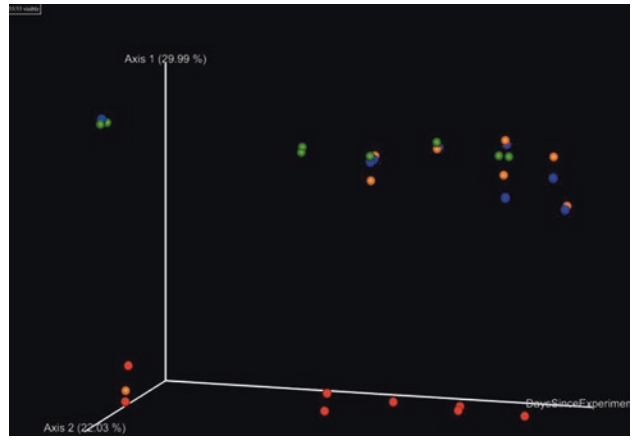


Fig. 8.15 Plots of PCoA based on Bray-Curtis dissimilarity of microbial communities in the gut, left palm, right palm, and tongue during DaysSinceExperimentStart. "Moving Pictures" tutorial in QIIME 2 (► https://qiime2.org/) (**Colors:** red = gut, blue = left palm, orange = right palm, green = tongue)

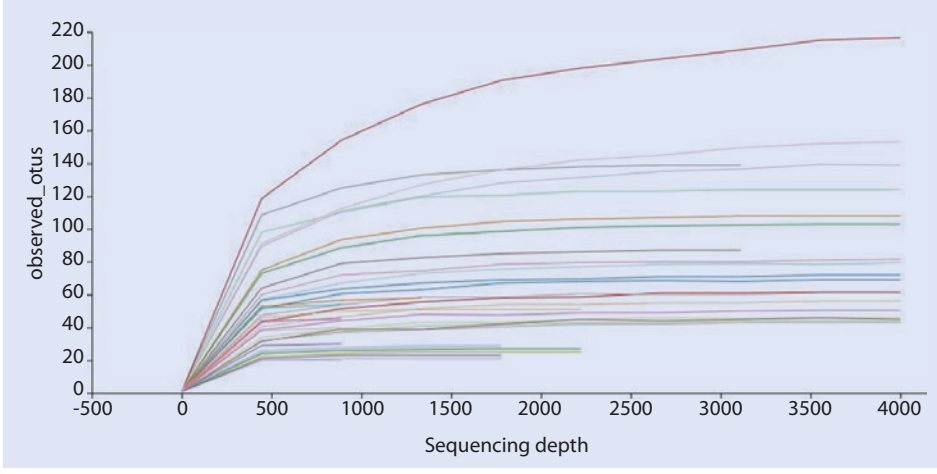


Fig. 8.16 Alpha rarefaction curve from the "Moving Pictures" tutorial in QIIME 2: ► https://qiime2.org/

8.10.11 Taxonomic Analysis

In this step, the bacterial DNA is annotated to the Greengenes database using the classifier (gg-13-8-99-515-806-nb-classifier.qza). The rep-seqs.qza artifact is used here; it contains the feature IDs and direct link to the NCBI database.

To improve taxonomic classification, creating your own classifier might be beneficial. To do that please follow the instructions in the tutorial "**Training feature classifiers with q2-feature-classifier**" in ▶ Sect. 8.10.14.3.

Results from this analysis are shown in ◘ Fig. 8.17 (genus level).

8.10.12 Exporting Data

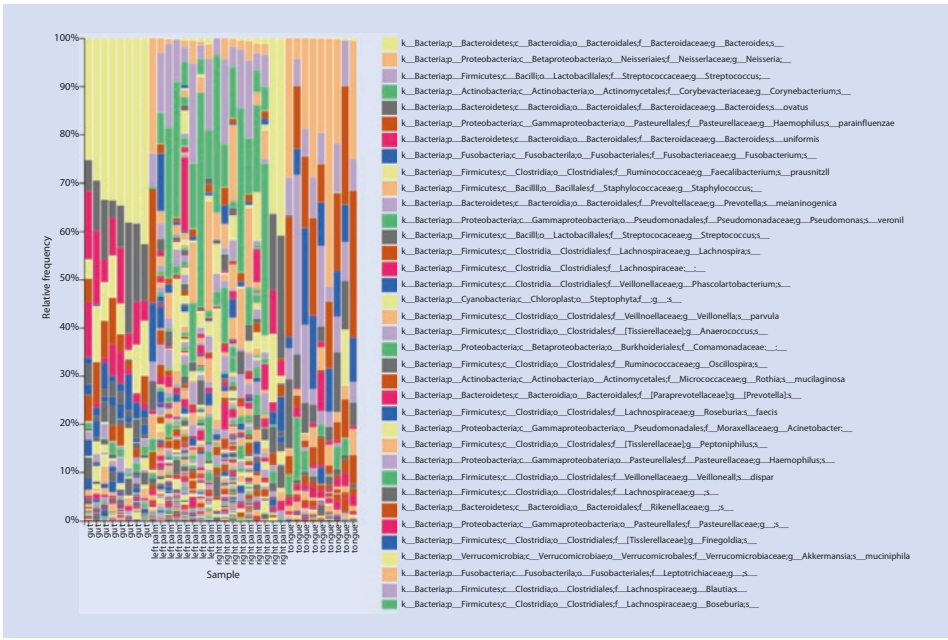
QIIME 2 is a good tool to start analyzing your data. However, to further explore and visualize your data, we recommend you to export relevant files from QIIME 2.

To do this, you have to convert the QIIME 2 artifacts into formats that are suited for export and visualization in other programs. Use the "Exporting data tutorial" in QIIME 2 to export your files, e.g., exporting the feature table (table.qza):

```
qiime tools export table.qza –output-dir exported-table
```

Note: you might also have to convert your BIOM table to TSV format for convenience.

```
biom convert exported/feature-table.biom --o-feature-table.tsv --to-tsv
```



◘ **Fig. 8.17** Taxa bar plot with classified bacteria and their relative abundance in the samples for category BodySites. (Example from the "Moving pictures" tutorial in QIIME 2 (▶ https://qiime2.org/))

8.10.13 Filtering Data

The tutorial "filtering data" in QIIME 2 explains how to filter feature tables, sequences, and distance matrices, e.g., if you want to modify a taxa bar plot (Fig. 8.10). The q2-taxa plug-in sorts the taxa as you want to display the results.

An alternative for this step is to convert the **taxonomy.qza** file to TSV format (see above) and import into your favorite statistics program (e.g., R).

8.10.14 Good to Know When Working with QIIME 2

8.10.14.1 Already Demultiplexed Data

Your project data might already be demultiplexed and still paired. Use the "Importing data" tutorial and create a "Fastq manifest" file in CSV format, and then follow the instructions for single-end reads or paired-end reads. The final output files are paired-end-demux.qza (or single-end-demux.qza for single-end reads). Paired-end reads with PHRED offset 33 were chosen as the shown example from the tutorial.

The manifest file must be in CSV format with the header line *sample-id*, *absolute-filepath*, and *direction*. There can only be one line per sample-id as shown below:

```
sample-id, absolute-filepath, direction
sample-1, $PWD/some/filepath/sample1_R1.fastq, forward
sample-1, $PWD/some/filepath/sample1_R2.fastq, reverse
```

Import the manifest file (e.g., pe-33-manifest) you have created with command:

```
qiime tools import --type 'SampleData [SequencesWithQuality]'
--input-path se-33-manifest --output-path single-end-demux.qza
--source-format SingleEndFastqManifestPhred33
```

When you have converted the manifest file to a QIIME artifact, visualize it with:

```
qiime demux summarize --i-data paired-end-demux.qza  --o-visualiza-
tion paired-end-demux.qzv
```

If you use Deblur instead of DADA2, continue to the step merging reads (refer to the following Chap. 8.10.14.2); otherwise follow the instructions in "Sequence and quality control and feature table construction" for "Moving Pictures."

8.10.14.2 Merging of Paired-End Reads

The sequence quality control can be performed either with DADA2 or Deblur. If you use Deblur, the forward and reverse reads have to be merged first as it only works for single-end reads. The plug-in vsearch join-pairs is used for this step of the analysis (► https://docs.qiime2.org/2018.2/plugins/available/vsearch/). The commands for join pairs are shown below.

```
qiime vsearch join-pairs --i-demultiplexed-seqs paired-end-demux.
qza --o-joined-sequences demux-joined.qza
```

```
qiime demux summarize --i-data demux-joined.qza --o-visualization
demux-joined.qzv
```

Now, you can run the **demux-joined.qza** through the sequence quality control with Deblur (refer to "Moving Pictures" tutorial in QIIME 2).

8.10.14.3 Train Your Classifier

The tutorial "Training feature identifiers with q2-feature-classifier" guides you through the steps that create the specific classifier for your dataset (► https://docs.qiime2.org/2018.2/tutorials/feature-classifier/). The elements required are Greengenes reference sequences, your own rep-seqs.qza, primer sequences used for your dataset, and the length of the reads (value chosen in Deblur's quality check).

8.10.14.4 Plug-Ins

If you cannot find the desired functions in the tutorials provided, view the available plug-ins (► https://docs.qiime2.org/2018.2/plugins/). For example, heatmaps are often used for visualization of the count data. In QIIME 2, there is a plug-in called feature-table (with heatmap) for this purpose. Bar plots generate an overview of the samples taxonomic makeup. These plots can, however, be difficult to read on lower taxonomic levels. The plug-in "feature-table" in QIIME 2 can be used to plot heatmaps from the features tables.

8.10.14.5 Decontamination

Contamination disrupts the accuracy of microbiome studies and is a known problem in this field. Upon until recently, there was no standard statistical method to handle these contaminants in the bioinformatical analysis. Davis et al. (2017) published a study about a new available open-source R package, decontam, that they created for the purpose to identify and remove the contaminants in marker gene (especially low-biomass samples) and metagenomics (Davis et al. 2017). Download it from (► https://github.com/benjjneb/decontam).

Take-Home Messages

- The 16S rRNA amplicon sequencing technique is a microbiota analysis of prokaryotes where different samples are analyzed at the same time using multiplexing.
- 16S rRNA amplicon sequencing results can be used to evaluate microbial diversity at species, genus, family, order, class, and phylum levels.
- The main limitation of 16S rRNA amplicon sequencing is related to the short read length of 250 nucleotides inhered from the Illumina sequencing techniques which results in insufficient taxonomic prediction of the species level.
- Raw data for 16S rRNA amplicon sequencing are mainly generated on the Illumina sequencing platform and bioinformatical pipelines used to analyze data include quality trimming of reads, removal of chimeras and sequences from other domains of life, grouping of reads into OTUs with a threshold of 97%, **alignment of OTUs with a database, and a**ssociation of OTUs with taxonomic units.
- The data analysis allows the diversity within samples, called α-diversity to be calculated, and the diversity between samples which is the β-diversity.

References

Anderson EL, Li W, Klitgord N, Highlander SK, Dayrit M, Seguritan V, Yooseph S, Biggs W, Venter JC, Nelson KE, Jones MB. 2016. A robust ambient temperature collection and stabilization strategy: Enabling worldwide functional studies of the human microbiome. Sci Rep. 25;6:31731.

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is esti-mated to contain substantial anomalies. Appl Environ Microbiol. 71, 7724–36.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 7, 335–336.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42, D633–42.

Danzeisen JL, Kim HB, Isaacson RE, Tu ZJ, Johnson TJ. 2011. Modulations of the chicken cecal microbiome and metagenome in response to anticoccidial and growth promoter treatment. PLoS One. 6:e27949.

Davis NM, Proctor D, Holmes SP, Relman DA, & Callahan BJ. 2017. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. bioRxiv preprint. https://doi.org/10.1101/221499

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 72, 5069–5072.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.

Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 10:996–998.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194–2200.

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation 64, 1–10.

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methds. 5:235–7.

Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, & Albertsen M. 2018. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat Biotechnol. 36:190–195.

Lane DJ. 1991. 16S/23S rRNA Sequencing. In: Stackebrandt, E. and Goodfellow, M., Eds., Nucleic Acid Techniques in Bacterial Systematic, John Wiley and Sons, New York, 115–175.

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower, C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31, 814–821.

Liu Z, Lozupone C, Hamady M, Bushman FD & Knight R. 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 35, e120.

Liu Z, DeSantis TZ, Andersen GL, Knight R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 36:e120.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6, 610–618.

McKenna P, Hoffmann C, Minkah N, Aye PP, Lackner A, Liu Z, Lozupone CA, Hamady M, Knight R, Bushman FD. 2008. The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. PLoS Pathog. 4:e20.

McMurdie PJ, & Holmes S. 2012. Phyloseq: a bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. Pac Symp Biocomput. 12:235–46.

Nossa et al. 2010. World J. Gastroenterol. 16, 4135–44.

Popp et al. 2017. Biospektrum Abstractbook, p. 199, 2017.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41:D590–6.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 18:e2584.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.

Schloss PD, Jenior ML, Koumpouras CC, Westcott SL & Highlander SK. 2016. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. Peer J 4:e1869.

Soergel DA, Dey N, Knight R & Brenner SE. 2012. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J. 6, 1440–4.

Tang Y, Underwood A, Gielbert A, Woodward MJ, Petrovska L. 2014. Metaproteomics analysis reveals the adaptation process for the chicken gut microbiota. Appl Environ Microbiol. 80, 478–485.

Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. 2016. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC Microbiol. 16:274.

Full Shotgun DNA Metagenomics

Henrik Christensen and John Elmerdahl Olsen

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_9

What You Will Learn in This Chapter

Full DNA metagenomics includes the ultimative way to sequence all DNA of a sample and to obtain all information of the microbiome. The microbiome can be analyzed with respect to taxonomic comparison and to functional characteristics. We will mainly focus on the pipeline MG-RAST to demonstrate different steps in the analysis. MG-RAST can also be used to archive data, to download data and to share data with the scientific community.

9.1 Background

Full DNA metagenomics is the sequencing of all DNA from a sample followed by assembly of DNA sequence reads and annotation and assignment of sequence information to known organisms and functions. The assembly attempts to reconstruct genome fragments to draft genomes. The technique involves most of the other subjects described in the book (Fig. 9.1).

The first real shotgun sequencing project was performed by Tyson et al. (2004). A very simple microbial community (with respect to diversity) of acid mine drainage was investigated (Fig. 9.2). The scientists managed to assemble two major bacterial genomes, *Leptospirillum* groups II and III (Bacteria) and a third composed of populations of *Ferroplasma* type II (Archaea). The study was based on 103,462 reads obtained by Sanger sequencing of plasmid inserts.

Another outstanding study was the in silico assembly of the genome of the organism *Candidatus Chloracidobacterium thermophilum*, an aerobic phototrophic acidobacterium (Bryant et al. 2007). The metagenomic data from the phototrophic microbial mats of an alkaline siliceous hot spring in Yellowstone National Park allowed the assembly of this

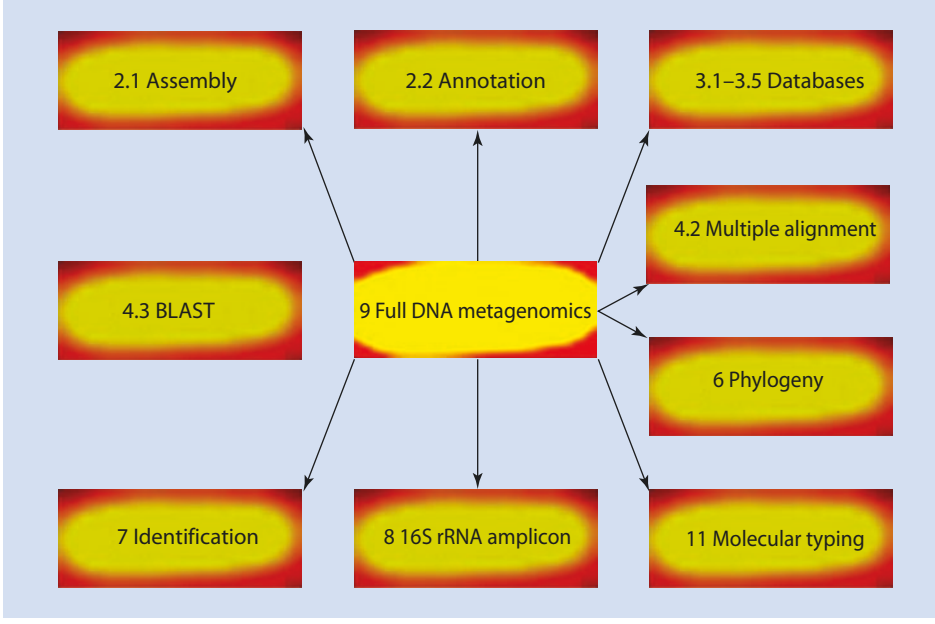


Fig. 9.1 Full DNA metagenomics is related to nearly all other chapters in the book

Fig. 9.2 Picture of acid mine discharge. This type of sample was used for the first full DNA metagenome study of Tyson et al. (2004). US Geological Survey: ► http://wwwbrr.cr.usgs.gov/projects/GWC_chemtherm/ironmtn.htm, Public Domain, ► https://commons.wikimedia.org/w/index.php?curid=2085336

distinctive bacteriochlorophyll (BChl) – synthesizing phototrophic bacterium without any previous isolation and cultivation.

The full DNA metagenomics concept has early on been used to diagnose infections caused by microorganisms in relation to human and animal medicine (Nakamura et al. 2008). In that investigation, *Campylobacter jejuni* was identified in a patient with diarrhea. A similar investigation of *Shigella* was reported more recently (Liu et al. 2018). A full DNA metagenomic approach to clinical diagnostics will overcome the problems with non-culturable bacteria that are problematic to identify by other methods. The level of detection is depending on the sequencing "depth" meaning the sequenced coverage of an expected target. This relationship is not yet standardized but has to be evaluated from case to case.

Also in food production, full DNA metagenome has been tested for detection of *Escherichia coli*, *Salmonella enterica*, and *Clostridium botulinum* (Yang et al. 2016). However, it was recognized that the technique was impractical for routine use.

The benefit of full DNA metagenomics is that the general diversity of all organisms can be investigated. The most serious drawback with the full DNA metagenomic technique is the limitations of the databases in particular with respect to taxonomy. Some pipelines are not able to extract 16S rRNA sequence information and provide search in the 16S rRNA databases which is further limiting the analysis. Another drawback is the higher cost of sequencing the full DNA metagenome compared to 16S rRNA amplicon sequencing.

9.2 Sequencing Strategies and Data Types

For experimental design and DNA extraction in general, we refer to other publications. Here as well as for 16S rRNA amplicon sequencing (► Chap. 8), we will refer to the MO BIO PowerMag® Soil DNA Isolation Kit (Qiagen) that has been used for DNA extraction. For the investigation of microbial communities, the output of DNA after extraction for sequencing should include at least 50% microbial DNA.

Table 9.1 Sequencing strategies in full DNA metagenomics

Strategy	Outline	Benefit	Drawback	Application
Full DNA shotgun	All DNA of a sample is high-throughput sequenced	High information content	Costly and analysis demanding	Main method
Full DNA with cloning	All DNA of a sample is shotgun cloned and sequenced	High information content, libraries can be kept	Extremely costly and demanding to analysis	Used before high-throughput sequencing. Now only used for special applications where libraries are needed
Full DNA with cloning and end sequencing	All DNA of a sample is shotgun cloned, but clones are only end sequenced	Less work demanding compared to sequencing of full libraries	Lower information content (only end reads of clones)	Used for initial screening of full libraries

For sequencing strategies, see Table 9.1. In the past, before high-throughput sequencing was invented, full DNA shotgun sequencing of cloned libraries was used. Figure 9.3 shows an example of such analysis where a dataset of the microbiome of mother and child was generated by sequencing of fosmid libraries. The taxonomic groups were predicted by searching GenBank.

End sequencing was a shortcut to the analysis of the cloned libraries of BAC (bacterial artificial chromosome) or fosmid libraries. Conserved primers complementary to the cloning vector are used to sequence the 5′ and 3′ ends of the inserts in the vectors. Such a sequence can be used to determine the type of organisms inserted and maybe to locate the insert on a fully sequenced organism. The information generated from the end reads have been used to select clones for full sequencing.

The benefit of full DNA shotgun sequencing with cloned libraries is that the cloned libraries can be kept and reanalyzed compared to the approach of using extracted DNA without cloning where the possibility for reanalysis ends when the DNA tube is empty. However, the sequencing capacity with cloned libraries is very limited compared to sequencing of DNA directly by high throughput without cloning.

Full DNA shotgun high-throughput sequencing without cloning is now the preferred method for sequencing. Sequencing platforms from Illumina are the most frequently used. DNA extracts from samples can be processed with the Nextera XT DNA sample Prep Kit (Illumina). The same principles of barcoding and indexing samples as used for 16S rRNA amplicon sequencing (▶ Chap. 8) are used. Paired-end sequencing can, for instance, be performed on the MiSeq Platform (Illumina).

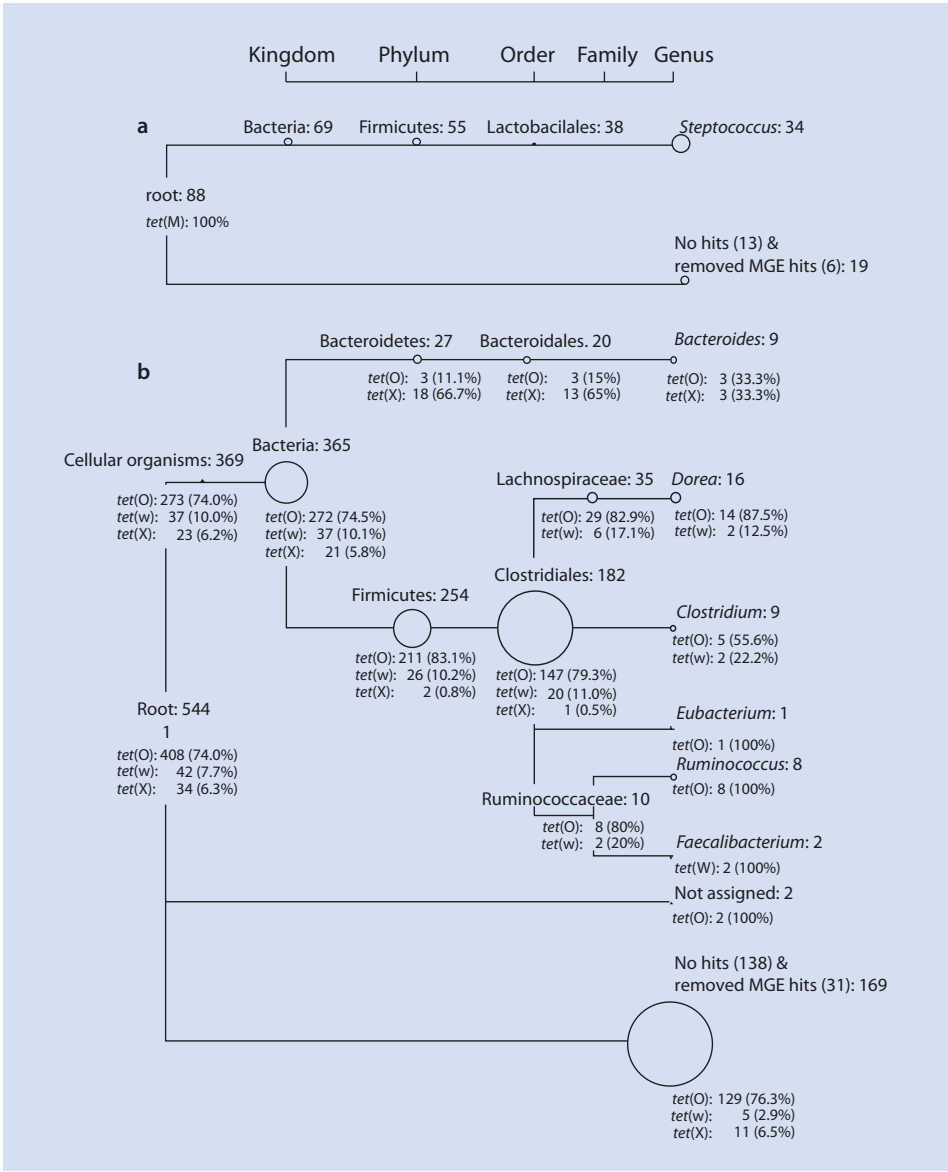


Fig. 9.3 Microbial diversity of tetracycline resistance fosmid clones in infant **a** and mother **b**. The MEGAN (Huson et al. 2016) (Table 9.2) tree is collapsed at genus level and summarizes the numbers of reads assigned at different taxonomical levels with the size of the node proportional to the number of reads assigned to the specific node. For each taxon level, the number of reads assigned to different tetracycline resistance genes is shown (From de Vries et al. 2011) (► http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021644)

9.3 Analysis of Full DNA Shotgun Sequence Data

After sequencing, the raw reads need to be de-multiplexed (to associate barcodes with samples) and converted to FASTQ format files. This step can be done on the MiSeq sequencing platform. Further handling of data requires a bioinformatic pipeline. It is recommended to perform a control for completeness and contamination by CheckM (Parks et al. 2015) (▶ http://ecogenomics.github.io/CheckM/) or similar program.

Quality guidelines have been published for metagenome-assembled genomes (Bowers et al. 2017). According to the paper, there are a range of criteria but no strict definition or guidelines to assemble genomes from metagenomes. Like for whole genomes from cultured bacterial isolates, key parameters are total assembly size, contig N_{50}, and maximum contig length. Sequences should be assembled to the longest possible contigs (▶ Chap. 2). The assembly may involve assembly of scaffolds (genome frame).

For high-throughput sequencing data, the initial steps in the bioinformatical analysis with quality control and filtering are the same as described for 16S rRNA amplicon sequencing (▶ Chap. 8). Full DNA metagenomics differs from 16S rRNA amplicon sequencing in that open reading frames (ORF) are predicted to encode for proteins that are identical. ORFs are compared to the existing databases (▶ Chap. 3). The database search can link information about organism (taxonomy) with function. This way taxonomy is predicted from the ORF in the databases.

The pipelines QIIME and Mothur described in relation to 16S rRNA amplicon sequencing (▶ Chap. 8) can be used; however, for full DNA shotgun sequencing, other pipelines are also available (◘ Table 9.2). MEGAN was introduced linking taxonomy with function (◘ Fig. 9.3). The platform has been updated (Huson et al. 2016). MEGAN assigns taxonomic relationships by comparison to NCBI taxonomy (▶ https://www.ncbi.nlm.nih.gov/taxonomy) using the (lowest common ancestor) LCA algorithm. Comparison can be made to the most common databases for functional identification. However, the focus in this chapter will be on another pipeline, MG-RAST (metagenomics RAST).

9.4 MG-RAST

The focus in this chapter will be on MG-RAST (metagenomics RAST) (Meyer et al. 2008) since it is probably the most user-friendly pipeline. MG-RAST can automatically perform assessment of sequence quality and annotation with respect to multiple databases based on uploaded raw metagenomic sequence data (Keegan et al. 2016). In MG-RAST, preprocessing including removal of low-quality reads, dereplication (merging of identical reads), DRISEE, screening, and removal of host-specific sequences including human and gene calling, AA clustering, protein identification, annotation mapping, and abundance profiling are done in the pipeline. DRISEE (duplicate read inferred sequencing error estimation) accounts for artificial overrepresentation of certain genes (PCR artifact). AA clustering is clusters of proteins build on 90% similarity. This is done to simplify the computational analysis.

MG-RAST has several databases available, for instance, can functional annotation be done against a multi-source protein database M5nr (MD5 based on nonredundant protein databases) (◘ Fig. 9.4). Subsystems under SEED (see description of RAST in ▶ Sect. 2.5) can be used for further functional annotation (Keegan et al. 2016) (◘ Fig. 9.5). The search in the databases is performed by BLAST (▶ see Chap. 4) or BLAT (Kent 2002). If 16S

Table 9.2 Bioinformatics pipelines for analyzing full DNA shotgun data

Name	Functions	URL	References
Blast2GO®	Functional annotation and analysis	► https://www.blast2go.com/	Conesa et al. (2005)
EBI metagenomics	Analysis and archiving of metagenomic data. Phylogenetic diversity as functional and metabolic potential of a sample	► https://www.ebi.ac.uk/metagenomics/	Mitchell et al. (2018)
GhostKOALA	Metagenomic annotation in the KEGG database	► http://www.kegg.jp/ghostkoala/	Kanehisa et al. (2016)
IGC	Dedicated human microbiome analysis	► http://meta.genomics.cn/meta/home	Li et al. (2014)
Kraken	Default database is NCBI RefSeq complete genome database. To investigate environments with well-known target organisms	► https://www.kraken.com/	Wood and Salzberg (2014)
MEGAN	Linking taxonomy with function	► http://ab.inf.uni-tuebingen.de/software/megan6/	Huson et al. (2016)
MetaPh1An	Human microbiome project	► http://huttenhower.sph.harvard.edu/metaphlan	Segata et al. (2012) and Huttenhower et al. (2012)
Metavir 2	Viral metagenome analysis	► http://metavir-meb.univ-bpclermont.fr/	Roux et al. (2014)
Mothur	Analysis of data	► https://www.mothur.org/	Schloss et al. (2009)
MG-RAST	Automatic quality control, storage of data, annotation, and comparative analysis of samples on a simple interface	► https://www.mg-rast.org/	Meyer et al. (2008)
QIIME	Analysis of data	► http://qiime.org/	Caporaso et al. (2010)

rRNA sequences can be extracted from the dataset, the databases SILVA, Greengenes, or RDP can be used for taxonomic identification like described in ► Chap. 8. The abundance plot shows the abundances of specific taxonomic categories (Figs. 9.6 and 9.7). MG-RAST is able to calculate α-diversity (Fig. 9.8).

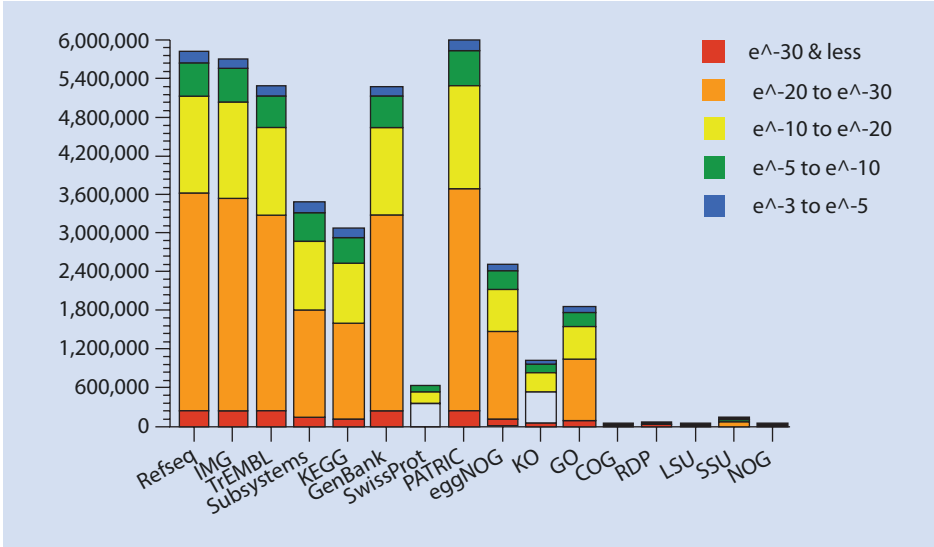


Fig. 9.4 Distribution of database hits based on the MG-RAST (Meyer et al. 2008). An example is shown based on the acc. no. mgm4629274.3 reported in Young et al. (2016) (► https://www.mg-rast.org/). The sample is from cat's feces, and the figure shows how many annotated sequence reads are found in the different databases. The annotated reads are finding most hits in the databases PATRIC (part of MG-RAST), RefSeq, IMG, TrEMBL, and GenBank. The expected value, e, indicates the chance that the sequence is present in the database (see explanation of e in ► Sect. 4.3.4)

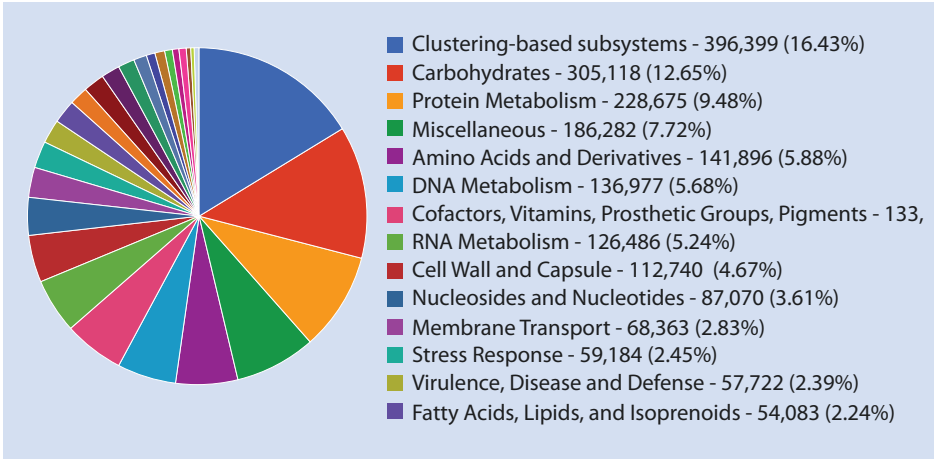


Fig. 9.5 Distribution of subsystem in MG-RAST (Meyer et al. 2008) using acc. no. mgm4629274.3 reported in Young et al. (2016) (► https://www.mg-rast.org/) as an example. The sample is from cat's feces, and the figure shows the distribution of annotated sequence reads on the functional protein subsystem categories in the databases of MG-RAST

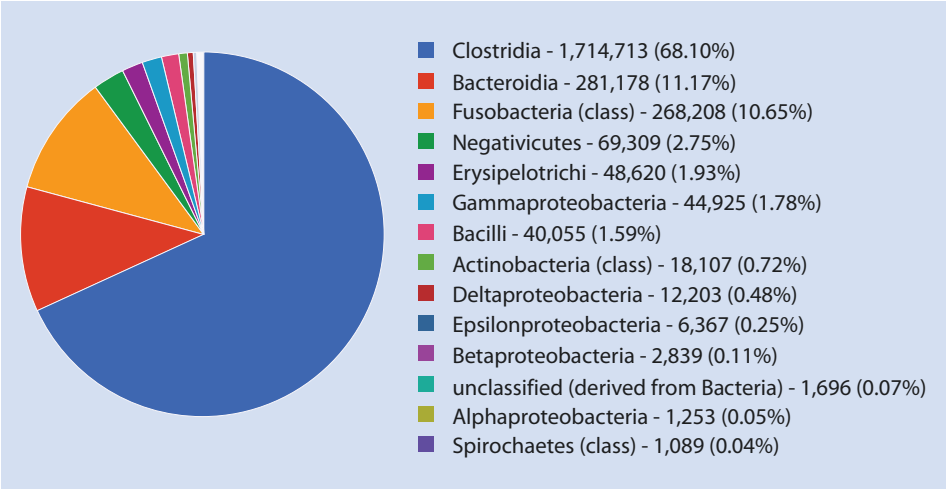


Fig. 9.6 Distribution of taxonomic hits based on the MG-RAST (Meyer et al. 2008) based on an example with acc. no. mgm4629274.3 reported in Young et al. (2016) (► https://www.mg-rast.org/). The sample is from cat's feces, and the figure shows the distribution of sequence reads on class level of prokaryotes

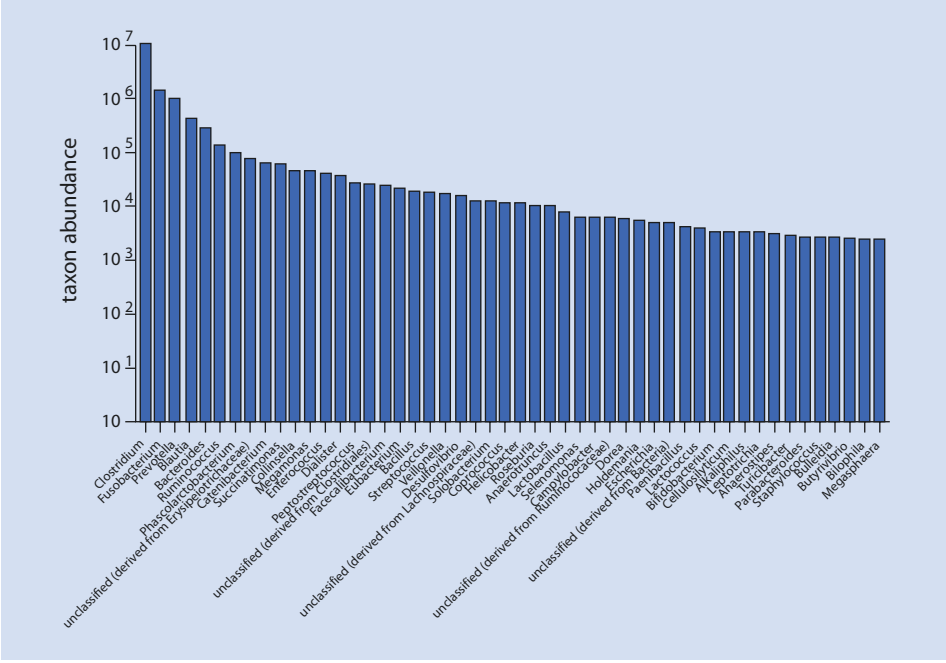


Fig. 9.7 Distribution of taxonomic hits at genus level based on the MG-RAST (Meyer et al. 2008) using acc. no. mgm4629274.3 reported in Young et al. (2016) (► https://www.mg-rast.org/) as an example. The sample is from cat's feces, and the figure shows the abundance of sequence reads with respect to genera of prokaryotes

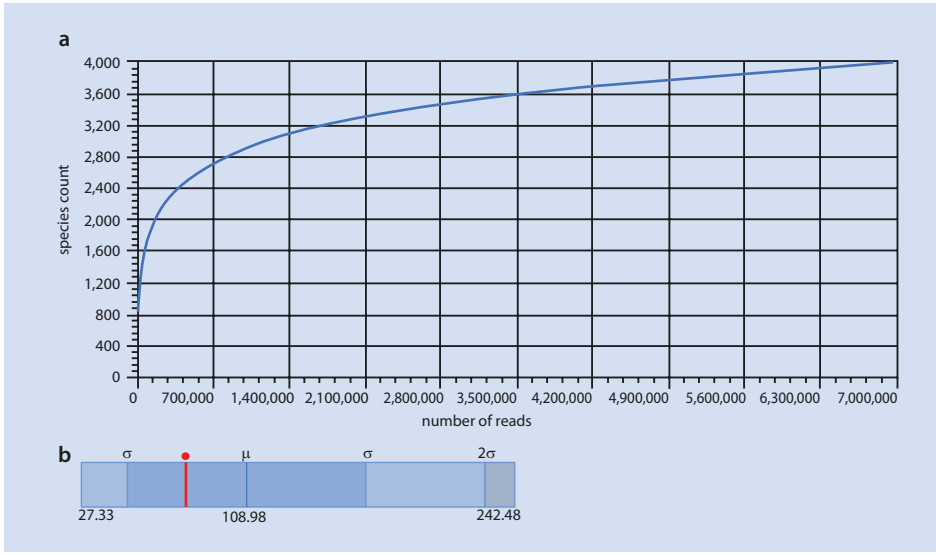


Fig. 9.8 Rarefaction plot **a** and α-diversity **b** based on the MG-RAST analysis (Meyer et al. 2008) of the sample with acc. no. mgm4629274.3 reported in Young et al. (2016). The sample is from cat's feces, and the plot in A shows the total number of distinct species annotations as a function of the number of sequences sampled. The more flat the curve, the less likely it is to find additional species. When this type of curve levels off, a sufficient number of reads have been sequenced for a given sample. The plot in B shows that the α-diversity is 79 (red spot) in this example meaning that the 79 different species have been estimated. The min, max, and mean values are shown, with the standard deviation ranges (σ and 2σ) (► https://www.mg-rast.org/)

MG-RAST allows both taxonomy prediction based on the approach in ► Chap. 8 based on the rRNA databases SILVA, Greengenes, and RDP as well as the prediction achieved from the ORF. The databases used for the ORF are very much limited with respect to organism diversity, and for data where a high degree of non-cultured prokaryotes are expected, the 16S rRNA function in MG-RAST will probably improve the taxonomic prediction.

9.5 Activities

9.5.1 Full DNA Metagenome: Shotgun DNA Metagenomics

We will use data reported in Young et al. (2016) which are available from MG-RAST. This is a metagenomics study of feces from 20 cats. The 20 datasets are listed in the paper, and we will only look at the first dataset mgm4629274.3. This is the MG-RAST acc. no., and you can view a lot of information about the dataset in MG-RAST.

Open MG-RAST from ► https://www.mg-rast.org/ (Fig. 9.9). Type in the number mgm4629274.3 in the search field, and press the search bottom. Note that some of the metadata seems odd (temperate broadleaf and mixed forest biome). It is just because we expect something about cats. These metadata are referring to the geographic location of the cats including the natural vegetation type.

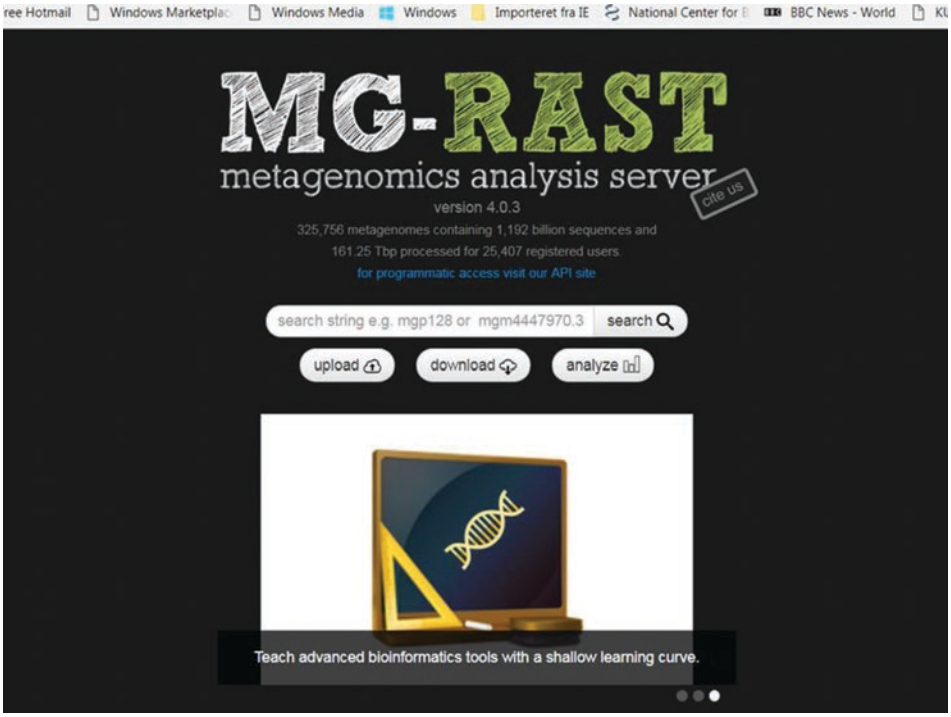


Fig. 9.9 MG-RAST (Meyer et al. 2008) pipeline with the simple interface of download, upload, and analyze bottoms (► https://www.mg-rast.org/)

Use no. t174 in the column "study" as link. Here you will get a list of all 20 datasets including mgm acc. numbers. Note that the initial no. mgm4629274.3 is on the list as the first cat. At each line in this table, an overview of number of reads and number of total nucleotides is provided. For the first sample, use number BY170 as link. This brings you to a paradise of graphics including taxonomic overview and breakdown on functional categories all in nice pie charts. These representations have been used for the figures of this chapter.

Take-Home Messages

- Full DNA metagenomics is the sequencing of all DNA from a sample followed by assembly and annotation and assignment of sequence information to organisms and functions.
- The assembly of DNA sequence reads attempts to reconstruct genome fragments to draft genomes.
- The microbiome is the common gene pool of all microorganisms in a sample analyzed by full DNA metagenomics and the microbiota the organisms predicted based on the microbiome.
- The computer program MG-RAST can be used to analyze, archive, and download data and to share data with the scientific community.

References

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C; Genome Standards Consortium, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol.;35(8):725–731.

Bryant DA, Costas AM, Maresca JA, Chew AG, Klatt CG, Bateson MM, Tallon LJ, Hostetler J, Nelson WC, Heidelberg JF, Ward DM. 2007. Candidatus Chloracidobacterium thermophilum: an aerobic phototrophic Acidobacterium. Science 317(5837):523–526.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 7, 335–336.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–6.

de Vries LE, Vallès Y, Agersø Y, Vaishampayan PA, García-Montaner A, Kuehl JV, Christensen H, Barlow M, Francino MP. 2011. The gut as reservoir of antibiotic resistance: microbial diversity of tetracycline resistance in mother and infant. PLoS One 6:e21644.

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R. 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol. 21;12(6):e1004957.

Huttenhower C and collaborators (248). 2012. Structure, function and diversity of the healthy human microbiome. Human Microbiome Project Consortium. Nature 486(7402):207–214.

Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol. 2016 428(4):726–731.

Keegan KP, Glass EM, Meyer F. 2016. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods Mol Biol. 2016;1399:207–33.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Res. 12:656–64.

Li et al. 2014. An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology 32, 834–841.

Liu J, Almeida M, Kabir F, Shakoor S, Qureshi S, Zaidi A, Li S, Tamboura B, Sow SO, Mandomando I, Alonso PL, Ramamurthy T, Sur D, Kotloff K, Nataro J, Levine MM, Stine OC, Houpt E. 2018. Direct Detection of Shigella in Stool Specimens by Use of a Metagenomic Approach. J. Clin. Microbiol. 24;56(2): e01374–17.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, & Edwards RA. 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes BMC Bioinformatics 2008, 9:386

Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, Salazar GA, Pesseat S, Boland MA, Hunter FMI, Ten Hoopen P, Alako B, Amid C, Wilkinson DJ, Curtis TP, Cochrane G, Finn RD. 2018. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res. 46(D1):D726-D735.

Nakamura S, Maeda N, Miron IM, Yoh M, Izutsu K, Kataoka C, Honda T, Yasunaga T, Nakaya T, Kawai J, Hayashizaki Y, Horii T, Iida T. 2008. Metagenomic diagnosis of bacterial infections. Emerg Infect Dis. 14, 1784–1786.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–55.

Roux S, Tournayre J, Mahul A, Debroas D, Enault F. 2014. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics. 19;15:76.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-

source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 0;9(8):811–814.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004 Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978):37–43.

Wood DE, Salzberg SL. 2014 Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 3;15(3):R46.

Yang X, Noyes NR, Doster E, Martin JN, Linke LM, Magnuson RJ, Yang H, Geornaras I, Woerner DR, Jones KL, Ruiz J, Boucher C, Morley PS, Belk KE. 2016. Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Pathogens within the Microbiome of the Beef Production Chain. Appl Environ Microbiol. 82(8):2433–2443.

Young W, Moon CD, Thomas DG, Cave NJ, Bermingham EN. 2016. Pre- and post-weaning diet alters the faecal metagenome in the cat with differences vitamin and carbohydrate metabolism gene abundances. Sci Rep. 6:34668.

Transcriptomics

RNA-seq

Rikke Heidemann Olsen and Henrik Christensen

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_10

What You Will Learn in This Chapter

A gene is active when it is being transcribed to mRNA. mRNA is acting as a template for protein synthesis. Measuring of how much mRNA is transcribed can be used to estimate how active a gene is under a given circumstance. It is possible to get insight to the regulation status of all genes of a prokaryotic strain at the same time by sequencing the total amount of mRNA (the transcriptome) and annotating the sequences against the fully sequenced genome of the prokaryotic strain. The method is called RNA sequencing (RNA-seq), and this chapter will teach you methods on how to prepare RNA material for sequencing, considerations on experimental design, and how data are analyzed.

10.1 Introduction to Transcriptomics

The ability to measure how genes are regulated under certain developmental stages or physiological conditions has expanded the knowledge of the biology of both human and prokaryotic cells tremendously. A technique called Northern blotting, developed in 1977, was the first method to study gene expression (Alwine et al. 1977). More methods to investigate the regulation of genes, including quantitative PCR (qPCR) and microarrays have become available. The methods, however, have several limitations, for instance, in qPCR analysis, only few genes can be studied at the same time. For all the methods, the major drawback is hybridization probes (needed, e.g., for Northern blotting and microarrays), and specific primers (for qPCR) need to be manually designed, and consequently "you only find what you are looking for." Despite the bias that the researcher has to decide which contents to put on the hybridization plate, microarrays have been the gold standard to study differential gene expression during the 2000s. As the price of novel high-throughput sequencing has decreased promptly in the same period, RNA sequencing (RNA-seq) is now rapidly replacing hybridization techniques in genome-wide expression studies (Wang et al. 2009).

RNA-seq relies on high-throughput sequencing, and it will allow a genome-wide detection of "active" genes by measuring the level of the genes transcribed. The technique relates to many bioinformatic techniques already described including sequencing, annotation, databases, alignments, and BLAST (◘ Fig. 10.1).

For RNA-seq experiments, often different conditions of the same prokaryotic strain are compared, for instance, the highly oxacillin-resistant strain of *Staphylococcus aureus*, USA300, cultured in normal growth media can be compared to growth media with therapeutic concentrations of oxacillin (the concentration of oxacillin normally used to treat infections with oxacillin-sensitive *S. aureus*). RNA-seq will enable the identification of differently expressed (DE) genes between the two conditions. Such an experiment will allow the identification of DE genes in *S. aureus* under exposure to oxacillin compared to control condition and may elucidate the mechanism to why *S. aureus* USA300 is able to withstand the exposure to oxacillin. Apart from this example on how RNA sequencing can be used for functional studies, RNA-seq may also be applied for detection of SNPs, finding of novel genes, or a total transcriptome assembly, just to give some examples of the use of transcriptomic data.

In overview, the workflow for RNA-seq is relatively simple: extracted RNA is converted to cDNA; cDNA is sequenced on a next-generation sequencing platform (NGS) such as either Illumina, Helicos, or SOLiD; and finally, the sequence data are matched to genes by sequence alignment (► Chaps. 2 and 4).

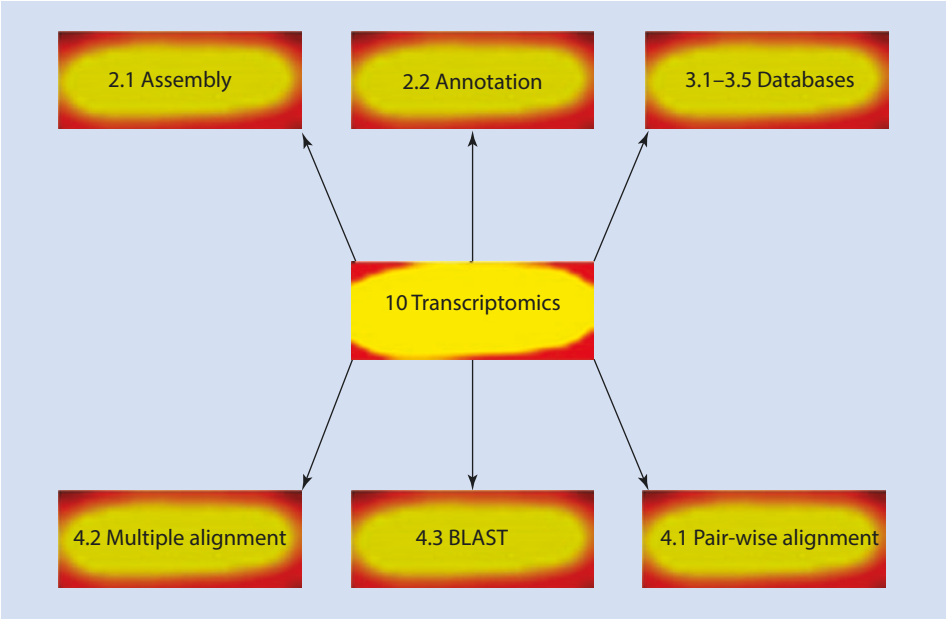


Fig. 10.1 Relationship of the chapter to other chapters of the book. RNA-seq relates to many preceding chapters dealing with sequence assembly, annotation, databases, alignments, and BLAST

The application of NGS has several advantages. In contrast to hybridization-dependent methods, the transcription is studied in an unbiased manner, since no probe sequences are needed to be specified. Secondly, the experimental design does not need to be altered in accordance with differences in genome sequences. Finally, the probe-less sequencing allows the discovery of new genetic features (Buermans and Dunnen 2014). Although the analysis is slightly more user-friendly for microarray data than for RNA sequencing data, RNA-seq is giving a more comprehensive overview of the transcriptome and a better dynamic range and gives the possibility to detect SNPs (Merwe et al. 2013). These features taken together are probably the key reasons of why RNA-seq has exceeded microarrays as the method of choice to analyze gene expression (Malone and Oliver 2011).

10.2 Experimental Design

There are several types of RNA that can be sequenced such as total RNA, small RNA, or mRNA. mRNA only constitutes about 2% of the total RNA in the cell; however, the assumption for transcription profiling is that changes in the transcriptional mRNA level correlate with the phenotype (protein expression). This chapter is only focusing on the sequencing of mRNA in a single culture and sequencing by the Illumina short-read technology, although mixed samples (metagenomics) and other sequencing platforms may also be suitable for RNA sequencing (Chu and Corey 2012).

The first thing to do before beginning an experiment is to decide on the experimental design (Fig. 10.2). The prokaryotic growth phase needs to be considered to assess gene expression since gene expression may vary significantly under exponential – compared to stationary growth phase (Rolfe et al. 2012). The number of replicates of each sample needs

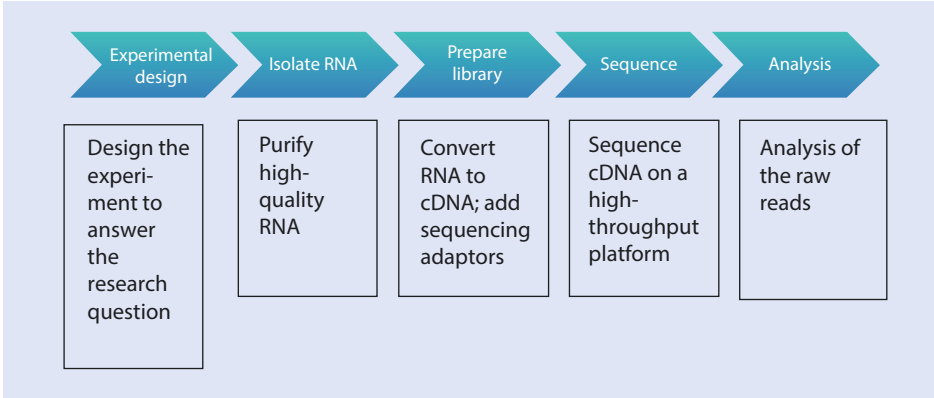


Fig. 10.2 Overview of RNA sequencing workflow

to be considered. It needs to be considered if rRNA needs to be removed. The sequencing depth required to answer a particular research/biological question needs to be defined. The optimal read length needs to be considered, and it needs be considered if samples can be pooled. The type of library needs to be selected and the sequencing platform decided including if single-end or paired-end sequencing is required.

It is important to make a distinction between biological and technical replicates. The biological replicates assess variations between samples, whereas the technical replicates can determine variation within sample preparations. RNA extracted from one sample and divided into three samples to be sequenced would be three technical replicates, whereas three samples of RNA harvested from three independently cultured colonies treated under similar conditions would represent three biological replicates. More biological replicates should always be favored over technical replicates, but be aware to stratify samples over time. Hence, do not extract RNA for all control samples Monday and all treated samples Wednesday, as factors such a humidity in the weather or even slight changes of the temperature in the culture water bath may influence the RNA profile. Differences between control and treated samples in this setup could then be due to external (environmental) factors rather than due to the treatment. More biological replicates (at least two, three is better) will increase the statistical power in the subsequent analysis.

The adequate read depth needed has to be assessed from experiment to experiment, depending on prokaryotic species and the research question under investigation. If you do not have sufficient read depth, the vast majority of reads will be associated with the highly expressed genes, which may not be the biologically most important genes needed to answer your research question, e.g., how does antibiotic x affect genome-wide gene expression (Depardieu et al. 2007). That being said, there is a trade-off between more read depth and replicates, meaning you can add more replicates rather than increase the read depth (Sims et al. 2014).

For highly expressed genes, little effect of an increased sequencing depth is gained on the number of differentially expressed (DE) genes detected. In this case increasing the number of biological replicates will be more beneficial. However, for low-expressed genes, both sequencing depth and biological replicates increase the power to detect DE. According to ENCODE guidelines, 10–20 million reads are sufficient for differential genes expression, but additional unique transcripts are still being found at one billon read (Lui et al. 2014).

The online tool named "Scotty" may guide for in the determination of sufficient reads for your RNA experiment (► http://scotty.genetics.utah.edu/).

The Illumina platform can perform single-end or paired-end sequencing. Paired end is more expensive than single end but improves mapping to reap sequences and improve the accuracy for detection of differential expression for low-expressed genes.

10.3 Preparing a RNA-seq Library

The Illumina TruSeq RNA protocol is the most commonly used protocol. Using the Illumina TruSeq RNA protocol, there are six steps in preparing an RNA-seq library: Step 1, isolation of RNA; Step 2, depletion or removal of rRNA; Step 3, conversion of rRNA into complementary, double-stranded DNA (cDNA); Step 4, addition of sequencing adaptors; Step 5, PCR amplification (enrichment); Step 6, quality control of the library.

■ Step 1: Isolation of RNA

There are several companies offering sequencing of RNA, e.g., ZF Genomics, Netherlands. It is, however, the responsibility of the researcher to provide a high RNA sample quality, which is essential for successful RNA-seq experiments. Many methods are available for purification of RNA from prokaryotic cells, including different manual protocols (e.g., an acidic phenol-chloroform RNA extraction protocol) and commercial kits (e.g., Qiagen RNA extraction kit). It is important to use the same RNA extraction protocol for all samples to be compared, since differences between protocols may slightly influence the RNA material (Sultan et al. 2014; Kumar et al. 2017). It is also important to acknowledge that irrespectively of protocol, RNA extraction is much more fragile than DNA extraction due to the ubiquitous and hard RNases that degrade RNA. Furthermore, RNA extraction is a challenging task due to the short half-life of mRNA (Tan and Yiap 2009). Gloves should always be used when handling/isolating RNA (human skin carries RNases) and all samples be kept on ice. It is acceptable to spin down prokaryotic cells at room temperature, but as soon as the cells have lysed, the temperature should be kept low (by keeping samples on ice) which will inhibit the activity on any luring RNases. Preferably use pipettes that are only handled for work with RNA. If that is not possible, at least make sure always to use RNase-free tips with filter. Be aware that RNases are very stable and will not be eliminated by autoclaving; hence, do only use certified RNase-free water.

The RNA quantity and purity can be evaluated to measuring the UV absorption of the sample using a spectrophotometer, e.g., on a NanoDrop spectrophotometer (Thermo Fisher Scientific). RNA has a maximum absorption at 260 nm, and the RNA concentration is determined by the OD reading at 260 nm. In addition to the OD260, measurements should also be taken at 280 nm and 230 nm. The A260/A280 ratio provides an indication of the level of protein contamination in the sample. Pure RNA has an A260/A280 ratio of 2.1; however, values between 1.8 and 2.0 are considered acceptable for many protocols. Be aware that OD absorbance measurements can change depending upon the pH of the RNA solution. The best results are obtained when RNA is solubilized in TE buffer. In general, RNA concentrations must be above 20 μg/mL to give reliable readings.

It is needed to control the integrity of the RNA preparation since RNA may be degraded and not perform well in downstream applications. The easiest and cheapest way to assess RNA integrity is by the means of a 1% standard agarose gel and examining the ribosomal

RNA (rRNA) bands. The upper ribosomal band (23S in prokaryotic cells) should be about twice the intensity compared to the lower band (16S in prokaryotic cells) and should be crisp and tight. If the rRNA bands are of equal intensity, then it suggests that some degradation has occurred. mRNA runs between the two ribosomal bands and might be seen as a smear. This is acceptable; however, smearing below the rRNA bands suggests that you have poor-quality RNA. Higher molecular weight bands that might indicate that the RNA is contaminated with DNA must not be observed.

A second method to control the integrity of your RNA that has become more popular, especially with in microarray analyses, is to use a bioanalyzer, such as the Agilent Bioanalyzer. Bioanalyzers use small amounts of RNA (1–2 μL) and microfluidics to determine the quantity and quality of RNA samples. The analyzer measures the sizes of the rRNA bands and determines an RNA integrity number (RIN) to standardize between RNA samples. Bioanalyzers are expensive, but they can often be found in core facilities.

The input needed for the TruSeq-stranded mRNA kit (Illumina) is 0.1–4 μg of total RNA. Using the Agilent Bioanalyzer, the RNA integrity number (RIN) value should be greater than or equal to 8.

■ Step 2: Depletion or Removal of rRNA

The majority of the total RNA is rRNA, and rRNA must either be depleted from the sample, or alternatively, mRNA must be captured. For the latter, the selection of mRNA can be done by poly(A) selection, which is done by filtering RNA with 3′ polyadenylated poly(A) tails. The RNA with 3′ poly(A) tails is a mature, processed, coding sequence. Poly(A) selection is performed by mixing RNA with poly(T) oligomers covalently attached to a substrate, typically magnetic beads (Cui et al. 2010). If whole RNA is to be sequenced instead, rRNA must be depleted. There are a number of commercially available kits for rRNA depletion, such a RiboZero (TaKaRa).

■ Step 3: Conversion of RNA into Complementary DNA (cDNA)

The first stage is fragmentation of the RNA to be sequenced. Fragmentation is achieved by using divalent cations, under elevated temperature, which ensures good coverage of the transcriptome. The cleaved RNA fragments are copied into first strand cDNA using reverse transcriptase and random primers.

One of the advantages of the Illumina TruSeq protocol is that it is "stranded" meaning that it will provide information from which of the two DNA strands a given RNA is derived. This provides a large, complete picture of the transcriptome (Wang et al. 2009).

The strand specificity of the samples is achieved by replacing dTTP with dUTP in the second strand synthesis process, which quenches the subsequent amplification of this strand because the polymerase used in the assay will not incorporate past this nucleotide. The final product of this step is blunt-ended, double-stranded cDNA molecules. The addition of an A-base to the ends of the cDNA prevents the blunt-ended fragments from ligating to one another during the adapter ligation reaction.

■ Step 4: Addition of Sequencing Adaptors

A ligation reaction takes place, which ligates multiple indexing adapters to the ends of the DNA fragments, preparing them for hybridization onto a flow cell.

- **Step 5: PCR Amplification (Enrichment)**

The adaptor-added products are purified and enriched with PCR to create the cDNA library. This step is necessary to ensure that the sequencing signal will be strong enough to be detected unambiguously for each base of each fragment.

- **Step 6: Quality Control of the Library**

To assess library quality, 1 µl of the post-enriched library is loaded on one of the following instruments: Advanced Analytical Technologies Standard Sensitivity NGS Fragment Analysis Kit (Advanced Analytical, Heidelberg, Germany) or Agilent High Sensitivity DNA chip (Agilent, Santa Clara, USA). The size of the library is controlled for distribution of the DNA fragments with a size range from approximately 200 bp–1 kb. The manufacturer's instructions should be followed for the respective instruments depending on the kit you are using.

10.4 Sequencing

When preparation of cDNA is done with TruSeq RNA, cDNA sequencing can be done on Illumina MiSeq or HiSeq platform as descripted elsewhere in the book (▶ Chap. 2). Overall, the sequencing of RNA does not differ from sequencing of genomic DNA. The suggested read length is 50–250 bp.

10.5 Data Management (Sequence Reads)

Data from sequencing will be provided in FASTQ format. Data management includes assessing data for the quality, alignment of the reads to a reference genome and normalization of the data, before the differential gene expression analysis can be conducted. Some of the most important bioinformatical programs are listed in Table 10.1, and examples on their use will be given in the text.

10.5.1 Raw Data

The Illumina platform provides raw sequence reads in FASTQ format, which may be stored directly at the Sequence Read Archive (SRA) (▶ Chap. 3). When assessing the raw data, start by checking whether the FASTQ file is consistent. The format can be validated with software such as FastQValidator (▶ https://github.com/statgen/fastQValidator). Then the base calling quality score, which is part of the sequencing data output, needs to be assessed. Quality scores reflect how confidently the right bases have been called. FastQC (part of the FastQValidator) is an excellent tool for assessing the quality of the sequencing run. The base calling quality score is called a Phred score, Q, which is proportional to the probability p that a base call is incorrect, where $Q = -10 \log_{10}(p)$. For example, a Phred score of 10 corresponds to one error in every 10 base calls ($Q = -10 \log_{10}(1/10)$), or 90% accuracy; a Phred score of 20 corresponds to one error in every 100 base calls, or 99% accuracy. A higher Phred score thus reflects higher confidence in the reported base. There is no defined cutoff on how low an acceptable Phred score can be, but one should aim for

Table 10.1 Examples of software programs to manage and analysis RNA sequencing data

Design of RNA-seq experiment	Quality control	Alignment	Normalization
Scotty ► http://scotty.genetics.utah.edu/	FastQC ► http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	Bowtie ► http://bowtie-bio.sourceforge.net/index.shtml	EdgeR ► http://www.bioconductor.org/package/release/bic/html/edgeR.html
RNAtor ► https://github.com/binaypanda/RNAtor	dupRadar ► http://bioconductor.org/packages/devel/bioc/html/dupRadar.html	GNUMAP ► http://dna.cs.byu.edu/gnumap/	DESeq ► http://bioconductor.org/packages/release/bioc/html/DESeq.html
PROspective Power ► https://rdrr.io/bioc/PROPER/	RNA-SeQC ► http://archive.broad ► ituteute.org/cancer/cga/rna-seqc	PerM ► https://code.google.com/archive/p/perm/	BaySeq ► http://www.bioconductor.org/packags/release/bic/html/baySeq.html

Phred scores higher than 33. FastQC can also be used to confirm graphically that the GC content is represented by a nicely bell-shaped form.

If the quality score of the run is low, it may be possible to remove unwanted parts of the raw data. The unwanted parts are either technical contaminations such as low-quality read parts or technical sequences such as the adaptors or biological contamination like poly(A) tails, rRNA, mitochondrial RNA, or tRNA. When these parts are removed, then FastaQC is run again, and it should be decided if you want to proceed with the data or do a rerun.

If you decide that the quality of the reads is sufficient to proceed, the sequences are ready to be aligned to a reference genome.

10.5.2 Alignments of Sequence Reads

Now the reads needs to be aligned to the genome(s) which provided the RNA sample(s); this is called read alignment or mapping. While the term "alignment" describes the process of finding the position of a sequencing read on the reference genome, "mapping" refers to assigning already aligned reads to transcripts which is also called quantification. The general challenge of short-read alignment is to map millions of reads accurately and in a reasonable time, despite the presence of sequencing errors, genomic variation, and repetitive elements. There are several tools to assist in the alignment, depending on whether the reads are aligned to a genome or the reads are assembled de novo, in which the reads need to be assembled first into longer contigs. These contigs can then be considered as the expressed transcriptome to which reads are remapped for quantification.

10.5.3 Normalization of Data

Normalization is a process designed for adjustment for sequencing depth and compositional bias. It is done, to identify and remove systematic technical differences between samples that occur in the data and to ensure that technical bias has minimal impact on the results. The most common need for normalization is related to differences in the total number of aligned reads.

Library size influences read counts. If one library is sequenced to 20 M reads and another to 40 M, in the latter case, most genes will be approximately double in their counts. Also, the library composition has importance as highly expressed genes may be overrepresented at cost of lowly expressed genes. One way to normalize data is by the use of read per kilobase million (RPKM) (for SE-RNA-seq) of fragments per kilobase million (FPKM) (for PE-RNA-seq). RPKM is calculated by dividing the reads for gene A by the length of gene A times the total number of reads.

RPKM = Reads for gene A/(Length of gene A × Total number of reads)

This formula normalizes read counts for (1) the sequencing depth, since sequencing runs with more depth will have more reads mapping to the gene, and (2) the length of the gene, since longer genes will have more reads mapping to them.

An alternative that might be considered as a better solution is TPM (transcripts per kilobase million). TPM is very similar to RPKM/FPKM. The only difference is the order of operations:

1. Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
2. Count up all the RPK values in a sample, and divide this number by 1,000,000. This is your "per million" scaling factor.
3. Divide the RPK values by the "per million" scaling factor. This gives you TPM.

Normalization for the library size may also be done using different software, such as DESeq2 and Edger R (Table 10.1).

Gene length influences the count, as longer transcripts generate more reads. However, the transcript length does not differ between samples. Since it is the relative difference that is of interest, gene length counts do not need to be normalized.

10.6 Differential Gene Expression

RNA-seq is a *relative* abundance measurement technology. The primary goal of the differential gene expression analysis is to quantitatively measure differences in the levels of transcripts between two or more treatments or groups.

Step one in any analysis is always the same: plot the data! You may use a principal component analysis (PCA) or something similar to plot the data. This plot gives a nice overview of how similar the different replicates are in the RNA composition (the closer, the better) and will give a first impression if you can expect to find interesting difference between you samples (Fig. 10.3).

The actual identification of DEG is typically done using the statistical program R with either edgeR or DESeq and will result in a plot similar to Fig. 10.4. In Fig. 10.4, black dots represent genes that are expressed the same, while each red dot is a gene that is expressed differently between the two sample conditions ("C" and "S" in Fig. 10.4). The

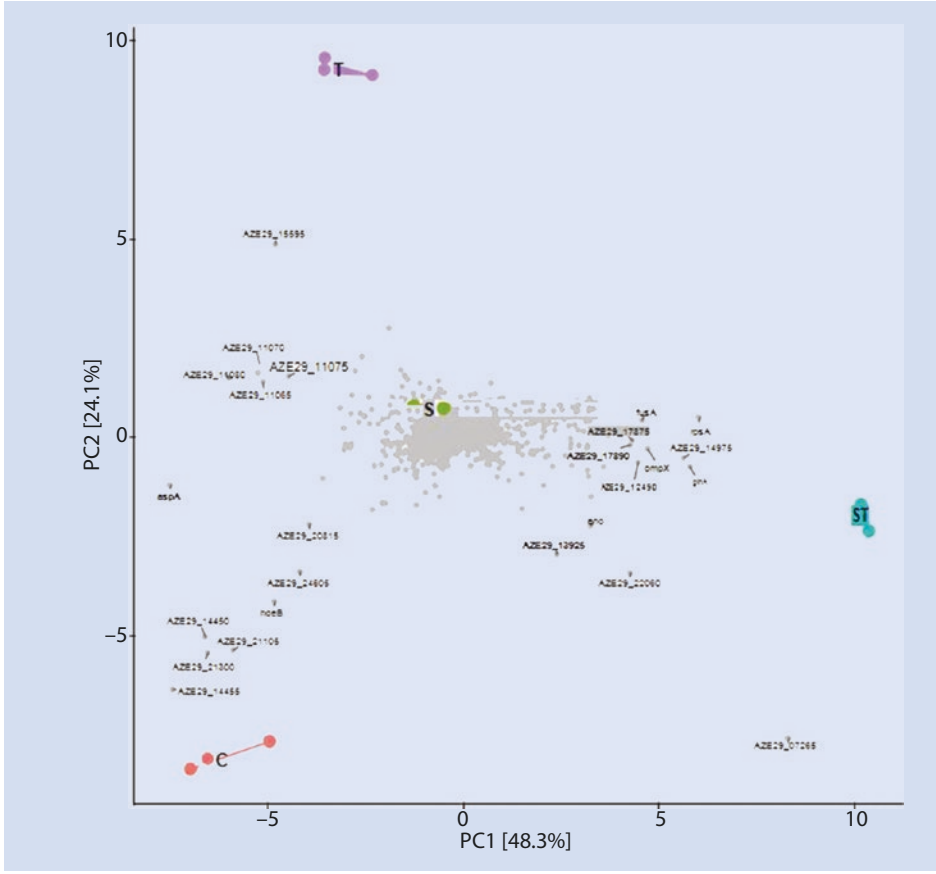


Fig. 10.3 Examples of a principal component analysis (PCA) to visualize the global gene expression of *Escherichia coli* under four different treatment conditions (C, T, S, ST). All conditions were performed in triplicates. Samples located together have similar gene expression. Genes (gray dots) located in the same direction as samples have higher expression in those samples

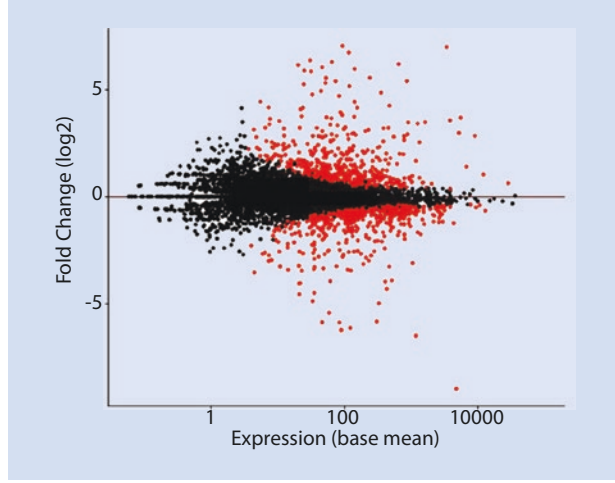


Fig. 10.4 This MA plot shows the fold change between the two treatment conditions of *Escherichia coli* (C and S) compared (log2 scaled) as function of the average expression of each gene. Each point represent a gene; red indicates an adjusted *p*-value <0.01

x-axis tells you how much each gene is transcribed, while the *y*-axis tells you how big the relative difference is between "C" and "S." The red genes are therefore your interesting genes. If you know what you are looking for, you can see if the experiment is validating you hypothesis. If you don't know what you are looking for, you can see if certain pathways are enriched under the different sample conditions using, e.g., KEGG Mapper (► www.genome.jp/kegg/mapper.html) or Cytoscape (Cline et al. 2007).

10.7 Conclusion

Since the first reported studies using RNA-seq were published in 2008, our understanding of gene expression has to be taken to a new level. However, there are still some technical issues awaiting resolution such as the PCR amplification stage of the library construction, which may result in redundant sequence read and bias in the final dataset. Nevertheless, RNA-seq holds the promise to continue to replace other genome-wide expression analysis in the future and will likely add in refining our understanding of gene regulation in prokaryotes.

Take-Home Messages

- It is possible to get insight to the regulation status of all genes of a prokaryotic strain at the same time by sequencing the total amount of mRNA (the transcriptome) and annotating the sequences against the fully sequenced genome of the same strain.
- The primary goal of the differential gene expression analysis is to quantitatively measure differences in the levels of transcripts between two or more treatments.
- RNA sequencing (RNA-seq) is based on high-throughput sequencing which allows a genome-wide detection of transcribed genes.
- The workflow for RNA-seq is that extracted RNA is converted to cDNA, it is sequenced with a next-generation sequencing platform such as Illumina, and finally, the sequence data are matched to annotated genes by sequence alignment.
- Data from sequencing will be provided in FASTQ format, and data management includes assessing data for the quality, aligning of the reads to a reference genome, and normalization of the data, before the differential gene expression analysis can be conducted.

References

Alwine JC, Kemp DJ, & Stark GR. 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. U.S.A. 74 (12): 5350–5354.

Bester-Van Der Merwe A, Blaauw S, Du Plessis J, Roodt-Wilding R. Transcriptome-wide single nucleotide polymorphisms (SNPs) for abalone (Haliotis midae): validation and application using GoldenGate medium-throughput genotyping assays. Int J Mol Sci. 2013 Sep 23;14(9):19341–60. doi: https://doi.org/10.3390/ijms140919341.

Buermans HP, den Dunnen JT. 2014. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta. Oct;1842(10):1932-1941. doi: https://doi.org/10.1016/j.bbadis.2014.06.015.

Chu Y, Corey DR. 2012. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther. Aug;22(4):271–4. doi: 10.1089/nat.2012.0367. Epub 2012 Jul 25

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD. 2007. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc. 2:2366–82.

Cui P, Lin Q, Ding F, Xin C, Gong W, Zhang L, Geng J, Zhang B, Yu X, Yang J, Hu S, Yu J. 2010. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics. Nov;96(5):259-65. doi: https://doi.org/10.1016/j.ygeno.2010.07.010.

Depardieu F, Podglajen I, Leclercq R, Collatz E, Courvalin P. 2007. Modes and modulations of antibiotic resistance gene expression. Clin Microbiol Rev. 2007 Jan;20(1):79–114

Kumar A, Kankainen M, Parsons A, Kallioniemi O, Mattila P, Heckman CA. The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia. 2017. BMC Genomics. Aug 17;18(1):629. doi: https://doi.org/10.1186/s12864-017-4039-1.

Liu Y, Zhou J, White KP. 2014. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 30:301–4.

Malone JH, Oliver B. 2011. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. May 31;9:34. doi: https://doi.org/10.1186/1741-7007-9-34.

Rolfe MD, Rice CJ, Lucchini S, Pin C, Thompson A, Cameron AD, Alston M, Stringer MF, Betts RP, Baranyi J, Peck MW, Hinton JC. 2012. Lag phase is a distinct growth phase that prepares bacteria for exponential growth and involves transient metal accumulation. J Bacteriol. Feb;194(3):686–701. doi: https://doi.org/10.1128/JB.06112-11. Epub 2011 Dec 2.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2010. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014 Feb;15(2):121–32. doi: https://doi.org/10.1038/nrg3642.

Sultan M, Amstislavskiy V, Risch T, Schuette M, Dökel S, Ralser M, Balzereit D, Lehrach H, Yaspo HL.2014. Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC Genomics. 2014; 15(1): 675. Published online 2014 Aug 11. doi: https://doi.org/10.1186/1471-2164-15-675.

Tan SC, Yiap BC. 2009. DNA, RNA, and protein extraction: the past and the present. J Biomed Biotechnol.:574398. doi: https://doi.org/10.1155/2009/574398.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10, 57–63.

Sequenced-Based Typing of Prokaryotes

Henrik Christensen and John Elmerdahl Olsen

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_11

What You Will Learn in This Chapter

You will learn some background of why sequence-based typing of microorganisms is performed. You will learn about sequence-based identification and characterization of populations based on multilocus sequence typing (MLST). You will then learn how the MLST concept has been transferred to whole genomic sequence comparison. You will learn about single-nucleotide polymorphism (SNP) analysis based on the whole genomic sequences as well as organism-specific servers enabling the prediction of serotype, antibiotic resistance gene profile and MLST type.

11.1 Background of Prokaryotic Populations and Population Genetics

In this chapter the background of population genetics mainly related to bioinformatics will be presented (Fig. 11.1). Population genetics is the study of the evolutionary change in the genetic composition of populations (Whittam 1995). According to Whittam (1995), population genetics applies both to how the mechanisms (mutation, natural selection, migration, genetic drift) influence the evolutionary rate of change in the populations (Fig. 11.2) as well as to historical investigations of how and when pathogens have evolved. The outcomes of such investigations will be definition of population structures, knowledge of the nature of allelic variation, and the role of different modes of recombination in generating genotypic variation (Milkman 1973; Selander and Levin 1980; Whittam 1995).

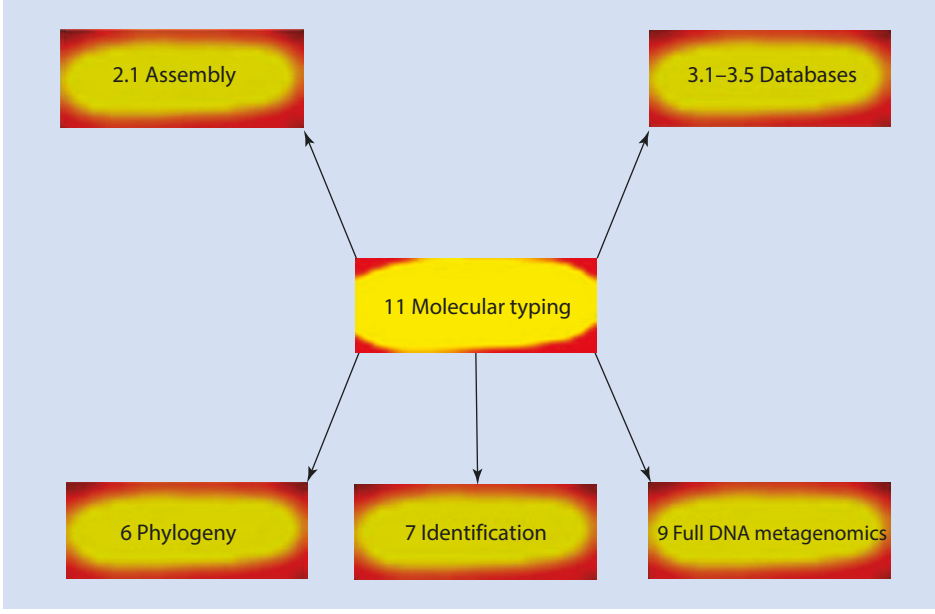


Fig. 11.1 Relationship of the chapter to other chapters of the book. Molecular typing based on sequence comparison relates to many preceding chapters dealing with sequence assembly, databases, phylogeny, and identification. For the future, molecular typing could probably also be linked to full DNA metagenomics

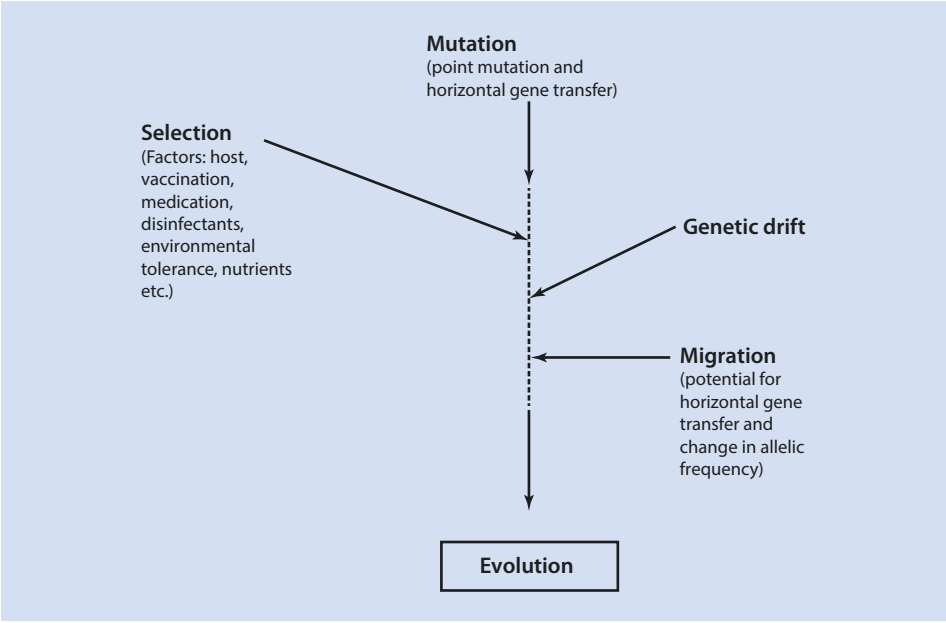
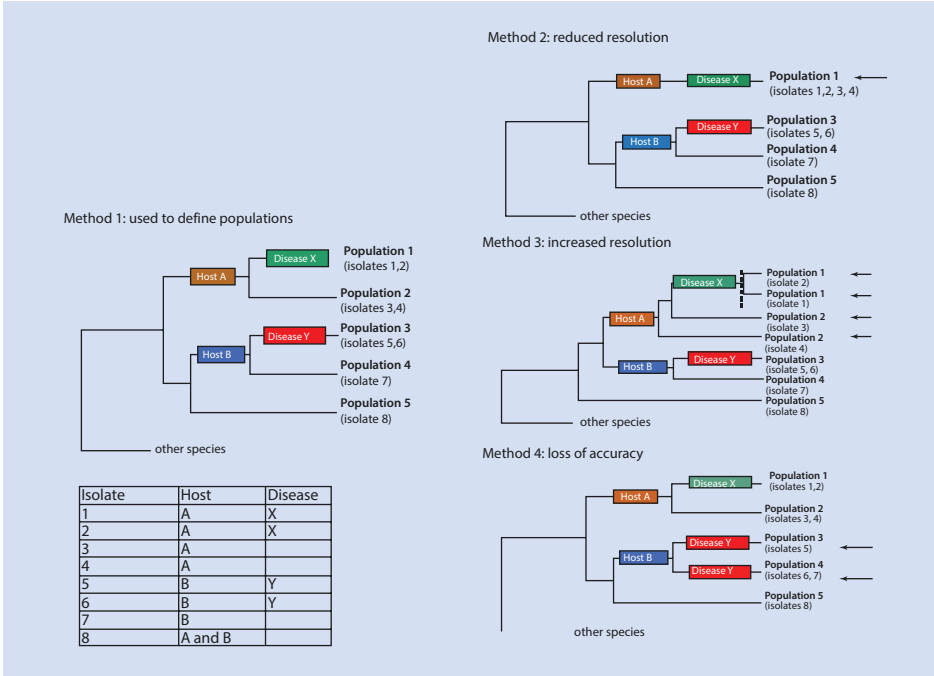


Fig. 11.2 Population genetic mechanisms relevant for the prokaryotes. The mechanisms of selection, mutation, and genetic drift contribute to the evolution of populations by shaping their genetic properties. Migration may also contribute if allelic frequencies are affected

In animals, a population is defined as a group of individuals with a potential for sexual reproduction often limited to a certain geographic region at a certain time point. However, this definition cannot be used for prokaryotes due to the lack of sexual reproduction. Groupings of prokaryotes are better reflected by clonal relationships than physical barriers. A prokaryotic clone was originally defined as a group of prokaryotic isolates "showing so many identical phenotypic and genetic traits that the most likely explanation for this identity is a common origin" (Ørskov and Ørskov 1983). Clonal populations are also called genetic lineages.

The most difficult task with prokaryotes is to set up the limits for the population. A practical solution might be to define the population borders according to the role that their members have in causing disease or other properties useful for epidemiology and for the investigation of pathogenesis. However, populations need to be identified by certain genotyping methods, and their principles are very much determining for how we observe the populations. These methodological problems with identification of clonal populations are illustrated in Fig. 11.3. Method 1 is optimal in linking genotypic clusters with isolate properties in relation to host and disease association. Method 2 has slightly lower resolution than method 1 and results in the lack of separation of populations 1 and 2. The non-disease-related isolates 3 and 4 are therefore included in this population which is confusing compared to method 1 that allowed the separation of population-associated disease from the nonpathogenic. Method 3 has higher resolution than method 1, and this will not contribute with any error as long as it is taken care of that some clusters might belong to the same population. Method 4 has the same resolution as method 1; however, isolates are clustered differently, and methods 4 will give another interpretation about their role in disease.



Isolate	Host	Disease
1	A	X
2	A	X
3	A	
4	A	
5	B	Y
6	B	Y
7	B	
8	A and B	

Fig. 11.3 Model illustrating dilemmas between methodological and biological relevance of populations. Clonal population are defined according to method 1 providing optimal biological interpretation (hosts A, B, disease associations X, Y) for the isolates investigated in the table

The term "strain" is sometimes misused to equal a population or a clone. In the strict sense, a strain is an isolate that has been further characterized, archived, and documented. The term "strain" refers to the culture or subcultures of it. The misuse is related to the fact that two isolates that share phenotype and genotype in principle must be considered as belonging to the same strain.

11.1.1 Mutation

The most likely starting point for an event affecting the genetics of a population is a mutation (Fig. 11.2). The progeny of a prokaryotic cell should in principle be genetically identical to its ancestor. Two genetic mechanisms, point mutations and horizontal gene transfer (HGT), tend to abolish this identity over time. Mutations will increase genetic diversity between individuals. Point mutations accumulate with a nearly constant rate at random positions in the sequence, and we can analyze them by sequence comparison until a certain level of divergence. HGT, however, is more problematic to analyze, since large fragments of sequences are exchanged between individuals and events of transfer cannot easily be traced since the origin of such fragments are often unknown.

Recombination in prokaryotes has a completely different meaning compared to eukaryotes. In eukaryotes, recombination refers to the results of crossing over in a symmetrical way between chromosomes during the zygotene stage of meiosis. This happens between members of the same species and even with members of the same population.

Recombination in prokaryotes includes HGT, and it can take place between prokaryotes by transformation, transduction, and conjugation and usually occurs as an asymmetrical exchange events between the partners. Some bacteria like *Haemophilus influenzae* are naturally transformable meaning that they can take up DNA directly from the environment. In bacteria with a double-wall structure (Gram negative), DNA is taken up as double stranded across the outer membrane and as a linear single-stranded across the inner membrane, and uptake signal sequences (USS) have been found to favor uptake (Maughan et al. 2008). Transduction is the transfer of genetic material by bacteriophages. There is high variability in the ability of bacteriophages to cross-react with strains within the same species, between different species of the same genus and between different genera of the same family (Jones and Sneath 1970). Conjugation is when genetic material is transferred between two prokaryotes on a conjugational plasmid. This is the most frequent mechanism for horizontal transmission of antibiotic resistance genes.

Luria and Delbruck (1943) pioneered the investigation of prokaryotic population genetics by investigating how bacteria become resistant to bacteriophages. They found resistance to develop as a random process since resistance to bacteriophages developed independent on the presence of bacteriophages in cultures of bacteria sensitive to bacteriophages. Furthermore, they analyzed mutations in a statistical framework and found that the distribution of mutational events followed a special distribution (Luria & Delbruck distribution). Mutations rates are still measured based on the principles laid down by Luria and Delbruck (1943). The most common procedure is the fluctuation test. In this test, the distribution of mutants in a number of parallel cultures is used to estimate the mutation rate based on knowledge of the expected number of mutation events, the number of cultures, and the size of initial inoculum. A range of assumptions are taken in this calculation: constant probability of mutations per cell cycle, that the mutation rate is independent of growth phase, no cell death occur and revertants are not formed, only single mutants arise, wild type and mutants have the same growth rate, a negligible number of mutant cells are initially present compared to final numbers in cultures investigated, and that we are able to detect all mutants (Pope et al. 2008). The mutation rate is the probability of a mutation occurring per cell division. Another measure is the mutation frequency being the proportion of mutant prokaryotes present in a culture. The mutation rate is independent on the age of the prokaryotic culture and a more accurate measure of mutations compared to the mutation frequency that is related to the age of the prokaryotic culture. In principle we should be able to trace mutations directly to the DNA sequence level (Bell 2008). In practice this will require an enormous DNA sequencing effort; however, next-generation sequencing technologies might be able to do this job in order to measure mutation rates directly and not via fluctuation tests.

11.1.2 Selection

Periodic selection is the most prominent population-genetic mechanism of prokaryotes. If a specific allele is favored in the population, and this allele results in higher fitness, then all members of the population with this allele will replace other populations without the allele in the given environment by expansion of this clone. Periodic selection is a form of "bottleneck" (Levin 1981). The higher rate of HGT, the less is the effect of periodic selection (Levin 1981).

The degree of selection acting on a specific coding gene can in theory be predicted by calculation of the so-called dN/dS ratio (Nei 2005). The ratio is the ratio of non-synonymous nucleotide substitutions per non-synonymous site (dN) to that of synonymous nucleotide substitutions per synonymous site (dS). A non-synonymous substitution at DNA level is a change in nucleotide leading to a change in the amino acid translated from the codon where it occurs. It follows that a synonymous substitution is not affecting the amino acid translated from the codon. The theory is based on the "neutral theory": without positive selection, dN/dS = 1. Negative selection (dN/dS <1) occurs when deleterious alleles (recognized at amino acid level) are eliminated from the population by purifying selection and leaving only the synonymous changes to be observed. In positive selection (dN/dS >1), polymorphism is assumed to be maintained at the amino acid level, and the changes here will be relatively higher than those observed at synonymous sites. However, high polymorphism at the amino acid level might not necessarily be observed for positive selection to occur (Nei 2005). Textbook examples of positive selection are with peptide-binding sites of MHC genes from humans and mice as well as the antigenic genes of influenzae virus (Nei 2005). To carry out the dN/dS calculations, at least two closely related nucleotide sequences need to be compared pairwise. Computer programs can evaluate the substitutions in regard to all combinations of nucleotides that are legal for each codon. The program DnaSP (Librado and Rozas 2009) (► http://www.ub.edu/dnasp/index_v5.html) can be used to calculate selection and many other parameters from DNA sequence data. MEGA7 (Kumar et al. 2016) introduced in ► Chap. 6 can also calculate the dN/dS parameter and will evaluate if the value is statistically significantly different from neutral.

11.1.3 Genetic Drift

"Genetic drift is a random process that can cause gene frequencies in a population to change over time causing evolution without natural selection" (Madigan et al. 2018). The random distribution of certain genotypes within a population may lead to genetic drift if the population is small. The limit has been suggested as 10^8 cells, and with high population size, this effect may not be important. With low population size, then clones developed by periodic selection may be wiped out more or less by chance. The effect is linked to selection in the way that the weaker and more unpredictable the periodic selection, the higher the effect of genetic drift.

11.1.4 Migration

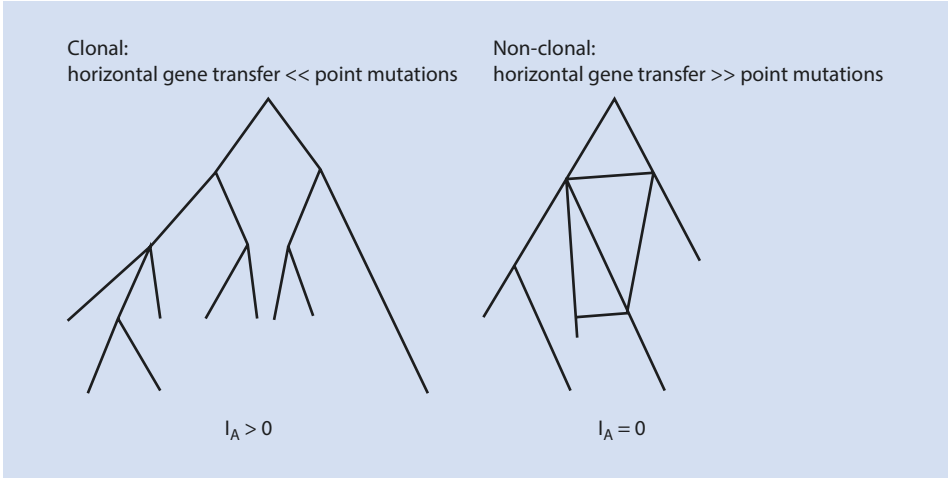
Migration is the assimilation of new individuals into a population from another population. For migration to take effect on evolution, the introduction of individuals should have consequences on population-genetic processes like allelic frequencies and mutation rates. Investigation of the spread of populations including bacterial spread between animals is not population genetics as long as such spread is not affecting the evolution of the organisms. Investigation of such spread is part of epidemiology.

11.1.5 The Biological Consequences of Population Genetics of Prokaryotes

In theory, the population structure of a prokaryote species is determined by the "ratio of genetic changes caused by recombination relative to de novo mutation" (Spratt and Maiden 1999) meaning that if HGT is relatively higher than point mutation rates, the population structure will be very complex and diverse, whereas low degree of recombination relative to point mutation will result in well-defined populations at the sequence level. Only in the last case are we able to recognize clonal populations (■ Fig. 11.4). The ratio between "recombination and point mutations" can be estimated by comparing recombination between the alleles of the genes used in the MLST (▶ Sect. 11.2) analysis to actual point mutations observed with the same genes.

The distribution of alleles investigated by MLST (▶ Sect. 11.2) can be used to predict if the population structure is clonal or non-clonal (panmictic). With equal assortment of alleles, there should be an equal random distribution of alleles between populations of a species. The expected variance V_E should equal the observed variance V_o. If the populations have evolved like clones, the alleles will be identical or highly related within clones and very different between clones, and V_o will be higher than V_E. To compare V_E and V_o, the index of association is calculated ($I_A = ((V_o/V_E) - 1)$, and it follows that I_A is not significantly different from 0 with a non-clonal (panmictic) population structure but significantly different with a clonal population structure (■ Fig. 11.4).

Some clonal lineages of a species seem to have adapted to specific hosts. In addition, these lineages often cause disease to a higher extent than other lineages, and for epidemiological investigations, it is therefore of great importance to identify and understand prokaryotes at the population level.



■ **Fig. 11.4** Population structures are related to the ratio between recombination and point mutations. The expected variance of allele frequency, V_E should equal the observed variance V_O. If the populations have evolved like clones, the alleles will be identical or highly related within clones and very different between clones, and V_O will be higher than V_E. The index of association is calculated as $I_A = ((V_O/V_E) - 1)$

One outcome of population genetics investigations is more realistic diagnostic methods for prokaryotic populations involved in disease. If population structures reflect disease patterns and hosts, we will be able to determine the populations that are really responsible for disease and not their commensal sister groups (Fig. 11.2).

Ecotypes are groups of prokaryotes playing ecological distinct roles defined based on DNA sequences. Ecotypes are analyzed by construction of phylogenetic trees based on housekeeping genes of isolates representing populations of a species. At a certain level of depth of a cluster in the tree, a group of populations that equal an ecotype is defined. Simulations are used to select this level of cutoff for the ecotype as well as to estimate periodic selection and genetic drift. Ecotypes are as a consequence one or more clonal populations if we refer to the current species concept of prokaryotes. Ecotypes can be regarded as species if we redefined the prokaryotic species concept; however, multiple ecotypes are usually recovered within the traditional species (Koeppel et al. 2008).

11.2 Multilocus Sequence Typing (MLST)

11.2.1 MLST

Multilocus sequence typing (MLST) is based on the comparison of DNA sequences of conserved genes in strains of a species (Maiden et al. 1998). For each gene (locus), all different versions of sequences are scored as alleles and designated a random number. This number is not reflecting the quantitative difference between the sequences but just stating that the sequences are different. An allele profile, for instance, 1, 3, 4, 5, 2, 3, 1 is defined as a sequence type (ST). If there only is one difference in the allelic profiles between two STs, these STs are single-locus variants (SLV). This will happen if, for instance, the one ST is 1, 1, 1, 1, 1, 1, 1 and the other 1, 1, 1, 1, 1, 1, 2. Double-locus variants (DLV) share five alleles and have two variants, and as the name indicates then triple-locus variants (TLV) have three different alleles. If a group of STs are linked as SLVs or DLVs, they form a clonal complex (CC). The clonal complexes are named after the ST estimated to be the ancestor of the complex – the founder.

MLST typing has mainly been performed on two servers ► http://www.mlst.net/ and ► https://pubmlst.org/general.shtml, respectively, each dedicated to a series of species. The databases on these servers have been maintained by curators who have uploaded new sequence types and information of isolates. Users are able to upload sequences to the servers and to obtain a sequence type.

ClonalFrame (► http://www.xavierdidelot.xtreemhost.com/clonalframe.htm) (Didelot and Falush 2007) is a program that can analyze population-genetic parameters based on MLST sequence data. ParseCF is used to do the final calculations of the population-genetic parameters that can be extracted from ClonalFrame such as the relative effect of recombination compared to mutation.

11.2.2 Multilocus Sequence Analysis

The information in the DNA sequences compared for MLST can also be directly analyzed by phylogeny (► Chap. 6) and is then called multilocus sequence analysis (MLSA). For MLSA the DNA sequences of the genes (loci) used for each ST are concatenated meaning

that they are joined end by end. For instance, if a seven-locus scheme includes partial regions of 500 nt., the concatenated sequence will be 3500 nt. For most of the MLST servers, concatenated datasets can be downloaded for this type of phylogenetic analysis. The information gained from such an analysis differs from the MLST analysis since the phylogeny will show the actual evolutionary diversity of the STs. The MLSA analysis works best for prokaryotes with a clonal population structure (Fig. 11.4). For a true clonal population structure, the phylogeny of all seven genes should be the same (congruent). More frequent horizontal gene transfer will lead to more linkage equilibrium and will result in a different phylogeny for each gene. In this case a concatenation of all genes will make little sense.

11.3 Whole-Genome-Based Typing

The relative low cost of generating whole genomic draft sequences has enabled the use of the most common typing methods based on the direct comparison of the whole genomic sequences. Even for MLST typing which only relies on a few kilo bases of sequence, it is now cheaper to sequence the whole genome and use that to determine the allelic profile than to sequence the seven genes by the traditional Sanger method. For the more common species of clinical importance, servers have become available that can predict serotypes, antibiotic resistance gene profiles, and virulence gene profiles.

11.3.1 Whole Genomic Multilocus Sequence Typing (wgMLST)

Whole genomic MLST (wgMLST) is an extension of the MLST concept to more than seven conserved genes of the species. In principle all conserved gene sequences can be included in such wgMLST scheme. We are aware that the term wgMLST in some other texts is used for the pangenome and cgMLST reserved only for the conserved part of the pangenome. For simplicity we will use wgMLST here and only consider to include the conserved genes of a species in such a scheme. However, if all conserved genes of the species are included, the number of variations in alleles and sequence types will be very high, and there will be a risk of too high resolution of the typing system (Fig. 11.2 method 3). To focus on clones with certain properties, a set of genes can be selected which includes genes predicted to encode for virulence or other important functional factors. A wgMLST typing system can be set up on the BIGSdb platform ▶ https://pubmlst.org/software/database/bigsdb/ (Jolley and Maiden 2010). This platform can be used both for MLST and for wgMLST. To set up the platform, you will need assistance from a professional system administrator. The platform allows users to upload whole genomic sequences and extract the sequence type. Curators of the database associated with BIGSdb are able to add new sequence types and information of isolates to the database. The platform has, for instance, been set up for various microorganisms: ▶ http://199.133.98.43/dbase_list.html.

ClonalOrigin is a parallel to ClonalFrame for whole genomic sequences (Didelot et al. 2010). This program should be able to model the source of specific recombination events.

11.3.2 Single-Nucleotide Polymorphisms (SNP)

Single-nucleotide polymorphisms (SNP) are point mutations that only occur as single-nucleotide changes in sequences. The SNP concept will be introduced here in relation to the comparison of raw reads of isolates obtained by high-throughput sequencing to a reference sequence. The reference can be a fully assembled genome of a well-characterized strain of the bacterial species investigated. It is important that this reference is closely related to the isolates that are being typed. The isolates for SNP typing are whole genome sequenced, and the reads are mapped to the reference. All positions where SNPs are identified between the reads and reference are then scored. The positions can be scored as a multiple alignment, and it can be used to construct a phylogeny as described in ▶ Chap. 6. To construct such a phylogeny, many isolates are compared at one time to the same reference.

The analysis can be performed at Center for Genomic Epidemiology (▶ http://www.genomicepidemiology.org). The files with all reads of a genome can be uploaded to ▶ https://cge.cbs.dtu.dk//services/CSIPhylogeny/ (Leekitcharoenphon et al. 2012). A reference genome needs to be defined, and then files with reads each representing isolates for typing can be uploaded. The programs on the server will extract all SNPs identified between the isolates compared to the reference. The SNPs can be downloaded in the form of a multiple alignment, and a phylogeny can be visualized by MEGA7 (Kumar et al. 2016). The benefit with this typing method is that is relatively easy to perform and has high resolution (Schürch et al. 2018). A database is not needed for this type of SNP like for wgMLST. Unfortunately, there are at least two drawbacks with the SNP concept. One is that rates of SNPs differ in different species. When *Listeria monocytogenes* was subcultured from a frozen stock every 3 months during a 3-year period, only one SNP was detected (Kwong et al. 2016) and almost at the same level of SNPs was found in *Mycobacterium tuberculosis* with four SNPs over 4 years, whereas *Helicobacter pylori* accumulated 30 SNPs per year (Schürch et al. 2018). The other problem is that SNP results are only valid in comparison with a specific reference. If another reference is selected, another result is obtained. For species of *Salmonella*, wgMLST was preferred compared to SNP in a large comparative study (Alikhan et al. 2018).

11.3.3 Typing of Virulence, Serotype, and Antibiotic Resistance Based on the Whole Genomic Sequence

This approach is only possible for certain well-characterized species where all or most virulence genes and antibiotic resistance genes have been identified and characterized and where the genetic background for the antigenic profile is well characterized. With this information, a typing system can be established on a server with Internet access. Such a tool is available Center for Genomic Epidemiology (DTU) (▶ http://www.genomicepidemiology.org/) (Larsen et al. 2012; Joensen et al. 2014, 2015). Here the prediction of virulence genes, MLST (multilocus sequence type), and serotype on the basis of the whole genomic sequences can be done for selected prokaryotic species. The whole genomic sequences can be uploaded in different formats including assembled genomes or raw reads.

11.4 Organism-Specific Platforms for Whole-Genome Sequence-Based Typing

SuperPhy (► https://lfz.corefacility.ca/superphy) (Whiteside et al. 2016) is dedicated for genomics investigation of *E. coli* that aims to link genomic information to phenotypic knowledge.

Enterobase is dedicated work with *Escherichia*, *Salmonella*, *Clostridioides* (*Clostridium difficile*), and *Yersinia* (► https://enterobase.warwick.ac.uk/) (Alikhan et al. 2018). At this server the raw Illumina reads can be uploaded and compared to traditional MLST (legacy MLST) and different variants of wgMLST databases. It is also possible only to perform MLST based on genes encoding for the ribosomal proteins (rMLST) (Jolley et al. 2012).

At NCBI, the pathogen detection portal (► https://www.ncbi.nlm.nih.gov/pathogens/) is available for identification and typing of the 22 most common human prokaryotic pathogens.

11.5 Activity

11.5.1 MLST Typing of *Pasteurella multocida*

This species is mainly an important animal pathogen; however, it can also affect humans if bitten by animals. A well-curated MLST scheme has been set up at ► https://pubmlst.org/bigsdb?db=pubmlst_pmultocida_seqdef (Davies et al. 2003; Jolley and Maiden 2010; Subaaharan et al. 2010). Select **Sequence query all loci,** and upload a genome of *P. multocida*. To get a genome, follow the instruction in Activity 3.8.2, and use acc. no. LT906458. The result should be ST13 in the **RIRDC MLST** scheme and ST3 in the **Multiple host scheme**. The reason for two sequence types is that there exist two MLST typing systems for this species. For most species, only one MLST scheme is available.

11.5.2 Graphics

PHYLOViZ (phylogenetic inference and data visualization) is used for visualization of results generated by sequence-based typing (► http://www.phyloviz.net/) (Francisco et al. 2012). It is a free-ware program that can visualize population genetics structures as a system of lines and circles with sizes of circles proportional to number of isolates in each ST (◘ Figs. 11.5 and 11.6). Similar facilities are only available from costly programs like Bionumerics (AppliedMaths) (► http://www.applied-maths.com/bionumerics).

The program can both be used online from the URL and as a downloaded program. Two files are needed, one with the MLST profiles and one with the isolate information (auxiliary data). At least one of the columns in the isolate files needs to be labeled the same way as in the profiles. This can, for example, be "ST" (◘ Fig. 11.5).

The program can be used online, but to obtain graphics for publications, you need to download and install the program from the URL. For Windows, save the zip file to your computer, unzip to a folder on c:, for instance, **c:/phyloviz**. Open the folder and the "**bin**" folder, and press the **phyloviz64.exe** icon which will activate the program.

Profiles file

ST	gene1	gene2	gene3	gene4	gene5	gene6	gene7
1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	2
3	3	1	1	1	1	1	2
4	3	1	1	2	1	1	2
5	3	1	2	1	2	1	1
6	1	2	1	3	2	2	1
7	3	3	2	3	2	2	1
8	3	3	1	2	1	2	3
9	3	3	2	3	2	2	3
10	3	2	2	1	3	2	1

Isolates file (note that the ST column links this file to the Profiles file)

ST	isolate	country	host
1	A1	USA	Mouse
2	B2	Germany	Mouse
3	C4	Australia	Mouse
3	D5	Germany	Mouse
3	E6	Germany	Mouse
4	G2	Australia	Rat
5	H3	Netherlands	Mouse
6	I1	Netherlands	Mouse
7	J6	Netherlands	Mouse
8	K9	Denmark	Mouse
9	L3	Germany	Mouse
9	M7	Netherlands	Mouse
10	N4	Australia	Rat

Fig. 11.5 Input files for PHYLOViZ. This example is showing ten MLST types in the "profiles file" and 13 isolates linked to the STs in the "isolates file"

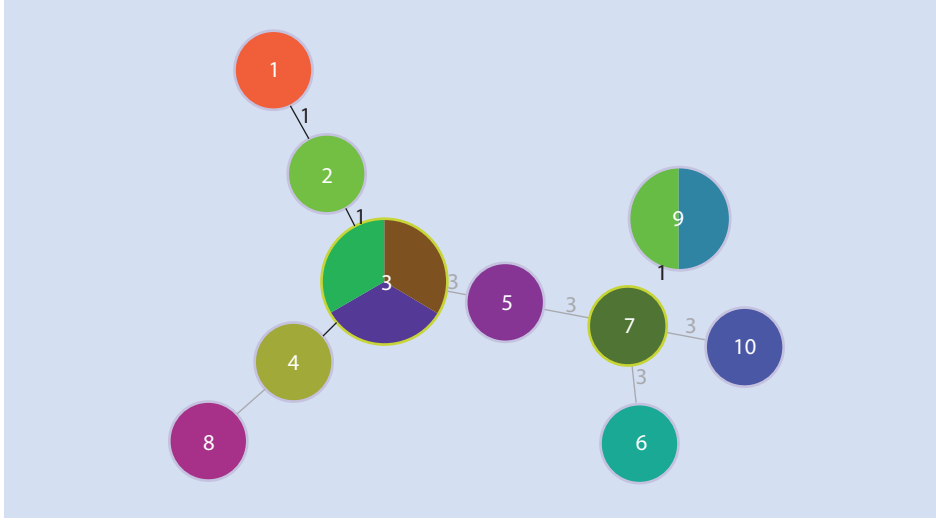


Fig. 11.6 Output from PHYLOViZ based on the data in Fig. 11.5. The nodes are labeled with the STs, and the numbers at the branches are the number of allele differences between STs. The size of nodes reflect the number of isolates in each ST, and the color can be used for a legend indicating the origin of the isolates

Prepare the two input files (◘ Fig. 11.5) with a text editor like **WordPad**. Separate the columns with one tab stroke and save each file in txt format. Start the program, and select **File | Load dataset | Dataset name** where you can select any name. Select dataset type: **MLST**, select **Next,** and upload the profiles file, press Open and **Finish,** and then the isolate files, and again press **Open** and **Finish**. At **Datasets** in the top left corner, you will now see the name of the dataset just uploaded. Double-click on it, and you will see the name of the two files with the isolate and profile (MLST) information. Right-click on the MLST set and select **Compute | geoBURST** and select **Next**. Select the level of links between STs, SLV, DLV, or TRV, for instance, TRV. Click on the key to left to the MLST name which will show the file **geoBURST**. Click on it, and you will see the graphics. Now go the upper left corner of the window, double-click on isolates which will show the table with isolates information, and press **Select** and **View**. Format the information from the **Options** panel below left to the graphics by unselecting **Group,** and select all other options. Use **control** to adjust the line lengths and node sizes. You can arrange the nodes and lines by dragging with the mouse. When done press the pause (||) sign. It should result in the output like ◘ Fig. 11.6. If there are problems, then control the input files with respect to profiles and isolate information, and make sure that columns have been separated by exactly one tab field. If the nodes are not nice pie-shaped, try to open the isolate table from the program, and press **Select** and **View** again. When the graphics is satisfactory, pause the viewer, save the result by clicking on the camera, and select appropriate format, for instance, jpg. Insert the jpg in power point as a picture (◘ Fig. 11.6).

You can now use this setup with the fictive dataset as a learning tool by changing the allelic profiles in the profile file and see the effect on the graphics. If there too few common alleles between the STs (less than 4), they will no longer be linked in the graphic structure.

Take-Home Messages

- Sequence-based typing of prokaryotes is performed to group isolates by common properties like generic genotypic relatedness or specific genotypic properties linked to virulence, serotype, or antimicrobial resistance.
- Multilocus sequence typing (MLST) is based on the comparison of nucleotide sequences of seven or more conserved genes in a set of related isolates.
- In MLST analysis the nucleotide differences within each gene (locus) of the isolates compared is scored as alleles, and a combination of alleles (profile) for all loci is designated a sequence type (ST).
- MLST information can be compared between different laboratories and analyzed in well-curated databases on servers via the Internet.
- Whole genomic multilocus sequence typing (wgMLST) is extending the MLST concept by the analysis of the core genome which may include hundreds of genes.
- In single-nucleotide polymorphism (SNP) analyses of whole genomes, the reads from high-throughput sequencing of isolates are compared to a common references sequence allowing a very detailed comparison at the single-nucleotide level in order to trace single isolates.
- Typing of virulence, serotype, and antibiotic resistance based on the whole genomic sequence can be performed for certain, mainly human pathogenic bacteria like *Escherichia coli* and *Salmonella enterica*, on dedicated servers that can be accessed via the Internet.

References

Alikhan, N.-F., Zhou, Z., Sergeant, M.J. & Achtman, M. (2018) A genomic overview of the population structure of *Salmonella*. PLoS Genet 14: e1007261.

Bell, G. 2008. Selection. The mechanism of evolution. 2nd ed. Oxford Univ. Press.

Davies, R.L., MacCorquodale, R., Baillie, S., Caffrey, B., 2003. Characterization and comparison of Pasteurella multocida strains associated with porcine pneumonia and atrophic rhinitis. J. Med. Microbiol. 52, 59–67.

Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence dta. Genetics 175:1251–66.

Didelot X, Lawson D, Darling A, Falush D. 2010. Inference of homologous recombination in bacteria using whole-genome sequences. Genetics. 186:1435–1449.

Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. 2012. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics 13:87.

Joensen, K.G., Scheutz, F., Lund, O., Hasman, H., Kaas, R.S., Nielsen, E.M., Aarestrup, F.M., 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J. Clin. Microbiol. 52, 1501–1510.

Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M., Scheutz, F., 2015. Rapid and easy in silico serotyping of Escherichia coli using whole genome sequencing (WGS) data. J. Clin. Microbiol. 5, 2410–2426.

Jolley, K. A. & Maiden, M. C. J. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **11:595**

Jolley, K.A., Bliss, C.M., Bennett, J.S., Bratcher, H.B., Brehony, C., Colles, F.M., Wimalarathna, H., Harrison, O.B., Sheppard, S.K., Cody, AJ. & Maiden, M.C. 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology 158, 1005–1015.

Jones, D. and Sneath, P. H. 1970. Genetic transfer and bacterial taxonomy. Bacteriol Rev. 34, 40–81.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, Nevo E, Cohan FM. 2008. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci U S A. 105: 2504–9.

Kwong, J.C., Mercoulia, K., Tomita, T., Easton, M., Li, H.Y., Bulach, D.M., Stinear, T.P., Seemann, T., Howden, B.P., 2016. Prospective whole-genome sequencing enhances national surveillance of Listeria monocytogenes. J Clin Microbiol. 54, 333–342.

Larsen, M.V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D.W., Aarestrup, F.M., Lund, O., 2012. Multilocus sequence typing of total genome sequenced bacteria. J. Clin. Microbiol. 50, 1355–1361.

Leekitcharoenphon, P., Kaas, R.S., Thomsen, M.C.F., Friis, C., Rasmussen, S., Aarestrup, F.M., 2012. snpTree-a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics 13 Suppl 7:S6

Levin, B. R. 1981. Periodic selection, infectious gene exchange and the genetic structure of E. coli populations. Genetics 99, 1–12.

Librado P & Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25:1451–1452.

Luria, S.E. & Delbrück M. 1943. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28, 491–511.

Madigan, M. T., Bender, K. S., Buckley, D. H., Sattley, W. M. & Stahl, D. A. 2018. Brock biology of microorganisms. 15th ed. Global Edition. Pearson, New York.

Milkman R. 1973. Electrophoretic variation in Escherichia coli from natural sources. Science182, 1024–6.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. & Spratt, B.G. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 95, 3140–3145.

Maughan, H., Sinha, S., Wilson L. and Redfield, R. 2008. Competence, DNA Uptake and Transformation in the Pasteurellaceae. In Kuhnert, P and Christensen, H. Pasteurellaceae, Biology, genomics and molecular aspects. Caister Acad. Press. In press.

Nei, M. (2005). Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22, 2318–42.

Pope et al. 2008. A practical guide to measuring mutation rates in antibiotic resistance. Antimicrob. Agents Chemotherapy 52, 1209–14.
Schürch, A.C., Arredondo-Alonso, S., Willems, R.J.L., Goering, R.V. 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin. Microbiol. Infect. 24, 350–354.
Selander RK, Levin BR. Genetic diversity and structure in Escherichia coli populations. 1980. Science 210, 545–7.
Spratt, B. G., M. C. J. Maiden (1999): Bacterial population genetics, evolution and epidemiology. Phil. Trans. R. Soc. Lond. B, 354, 701–10.
Subaaharan, S., Blackall, L.L., Blackall, P.J., 2010. Development of a multi locus sequence typing scheme for avian isolates of Pasteurella multocida. Vet. Microbiol. 141, 354–361.
Whiteside MD, Laing CR, Manji A, Kruczkiewicz P, Taboada EN, Gannon VP. 2016. SuperPhy: predictive genomics for the bacterial pathogen *Escherichia coli*. BMC Microbiol. 16:65.
Whittam, T. S. 1995. Genetic population structure and pathogenicity in enteric bacteria. In Population genetics of bacteria. Eds. Baumberg, S., Young, J. P. W., Wellington, E. M. H. and Saunders, J. R. 1995. 52nd Symposium of the Society for General Microbiology. pp. 217–245. Cambridge Univ. Press.
Ørskov, F., I. Ørskov (1983): Summary of a workshop on the clone concept in the epidemiology, taxonomy, and evolution of the Enterobacteriaceae and other bacteria. J. Infect. Dis. 148, 346–357.

Supplementary Information

H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8

Appendix

Abbreviation of Amino Acids

A	Alanine	Ala
C	Cysteine	Cys
D	Aspartic acid	Asp
E	Glutamic acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
P	Proline	Pro
R	Arginine	Arg
Q	Glutamine	Gln
S	Serine	Ser
T	Threonine	Thr
V	Valine	Val
W	Tryptophan	Trp
Y	Tyrosine	Tyr

Ambiguity Table Symbols

G or A	R	puRine
T or C	Y	pYrimidine
A or C	M	aMino
G or T	K	Keto

G or C	S	Strong interaction (3 H bonds)
A or T	W	Weak interaction (2 H bonds)
A or C or T	H	not-G, H follows G in the alphabet
G or T or C	B	not-A, B follows A
G or C or A	V	not-T (not-U), V follows U
G or A or T	D	not-C, D follows C
G or A or T or C	N	aNy

Codon Tables

The standard codon table (1) used for most animals and plants. Codons labelled with green are start codons and those labelled red are the stop codons.

All codon tables are listed at

► https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

1. The Standard Code (transl_table=l)
By default all transl_table in GenBank flatfiles are equal to id 1, and this is **not** shown.
When transl_table is not equal to id 1, it is shown as a qualifier on the CDS feature.
All code tables
http:/fmltw.ncbl.nlm.nlh.gov/Taxonomy!Utilslwprlntgc.cgi?mode=t#SGI

	pos. 2				
pos. 1	A	C	G	T	pos. 3
A	AAA K Lys	ACA T Thr	AGA R Arg	ATA I Ile	A
A	AAC N Asn	ACC T Thr	AGC S Ser	ATC I Ile	C
A	AAG K Lys	ACG T Thr	AGG R Arg	ATG M Met i	G
A	AAT N Asn	ACT T Thr	AGT S Ser	ATT I Ile	T
C	CAA Q Gln	CCA P Pro	CGA R Arg	CTA L Leu	A
C	CAC H His	CCC P Pro	CGC R Arg	CTC L Leu	C
C	CAG Q Gln	CCG P Pro	CGG R Arg	CTG L Leu i	G
C	CAT H His	CCT P Pro	CGT R Arg	CTT L Leu	T
G	GAA E Glu	GCA A Ala	GGA G Gly	GTA V Val	A
G	GAC D Asp	GCC A Ala	GGC G Gly	GTC V Val	C
G	GAG E Glu	GCG A Ala	GGG G Gly	GTG V Val	G
G	GAT D Asp	GCT A Ala	GGT G Gly	GTT V Val	T
T	TAA * Ter	TCA S Ser	TGA * Ter	TTA L Leu	A
T	TAC Y Tyr	TCC S Ser	TGC C Cys	TTC F Phe	C
T	TAG * Ter	TCG S Ser	TGG W Trp	TTG L Leu i	G
T	TAT Y Tyr	TCT S Ser	TGT C Cys	TTT F Phe	T

Codon table 11 used for prokaryotes. Codons labelled with green are start codons and those labelled red are the stop codons. Compared to codon table 1, table 11 has more alternative start codons; however, ATG is most frequently used as start codon.

All codon tables are listed at

► https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

11. The Bacterial and Plant Plastid Code (transl_table=11)

	pos. 2				
pos. 1	A	C	G	T	pos. 3
A	AAA K Lys	ACA T Thr	AGA R Arg	ATA I Ile	A
A	AAA N Asn	ACC T Thr	AGC S Ser	ATV I Ile	C
A	AAG K Lys	ACG T Thr	AGG R Arg	ATG M Met i	G
A	AAT N Asn	ACT T Thr	AGT S Ser	ATT I Ile	T
C	CAA Q Gln	CCA P Pro	CGA R Arg	CTA L Leu	A
C	CAC H His	CCC P Pro	CGC R Arg	CTC L Leu	C
C	CAG Q Gln	CCG P Pro	CGG R Arg	CTG L Leu i	G
C	CAT H His	CCT P Pro	CGT R Arg	CTT L Leu	T
G	GAA E Glu	GCA A Ala	GGA G Gly	GTA V Val	A
G	GAC D Asp	GCC A Ala	GGC G Gly	GTC V Val	C
G	GAG E Glu	GCG A Ala	GGG G Gly	GTG V Val	G
G	GAT D Asp	GCT A Ala	GGT G Gly	GTT V Val	T
T	TAA * Ter	TCA S Ser	TGA * Ter	TTA L Leu	A
T	TAC Y Tyr	TCC S Ser	TGC C Cys	TTC F Phe	C
T	TAG * Ter	TCG S Ser	TGG W Trp	TTG L Leu i	G
T	TAT Y Tyr	TCT S Ser	TGT C Cys	TTT F Phe	T

Index

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

W

X

Z