Kang Ning · Yi Zhan *Editors*

Synthetic Biology and iGEM: Techniques, Development and Safety Concerns

An Omics Big-data Mining Perspective

Springer

Synthetic Biology and iGEM: Techniques, Development and Safety Concerns

Kang Ning • Yi Zhan
Editors

Synthetic Biology and iGEM: Techniques, Development and Safety Concerns

An Omics Big-data Mining Perspective

Springer

Editors
Kang Ning
College of Life Science and Technology
Huazhong University of Science and Technology
Wuhan, China

Yi Zhan
College of Life Science and Technology
Huazhong University of Science and Technology
Wuhan, China

ISBN 978-981-99-2459-2 ISBN 978-981-99-2460-8 (eBook)
https://doi.org/10.1007/978-981-99-2460-8

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Synthetic biology is a new research area, while it originated from the long-established area, including biological engineering, metabolite engineering, and systems biology. So what makes synthetic biology unique? Systematic design, rational engineering, and design sustainability might represent three aspects that make synthetic biology an outstanding research area.

Such an exciting area has undoubtedly drawn attendees of all students worldwide, thus making iGEM one of the biggest synthetic competitions on the globe. Every year, several hundred or more than thousand undergraduate and graduate teams have gathered in Boston for the iGEM competition. Thus, iGEM has both educational and research purposes.

However, in the age of omics, the potential of synthetic biology as a research area and iGEM as competition have not yet been fully excavated in all of the three aspects of systematic design, rational engineering, and design sustainability. Firstly, in the age of omics, we have faced paramount multi-omics data from various sources, which is good since these omics data would provide clues for several biomedical or clinical applications. However, it has also created hurdles for omics data integration, data-mining, and in-depth understanding. Therefore, how to best utilize omics data for systematic design of biological systems (most are microbes) that are both efficient and useful are usually critical for the healthy development of synthetic biology. Secondly, rational engineering is crucial since if not rationally designed, a combinatorial explosion would make the synthetic system impossible. On the other hand, rational design would also lead to possible safety issues that are disastrous. Thus, rational engineering is also an essential procedure in which both choices of chassis and parts should be rationally made. Thirdly, the design sustainability ensures that both synthetic biology itself and iGEM competition would have enough people devoted to real applications.

As such, in this book, we will feature not only biological engineering techniques, multi-omics big-data integration, and data-mining techniques, but also focus on cutting-edge researches in principles and applications of several synthetic biology applications:

(1) Introduction to synthetic biology and iGEM, mainly focusing on the systematic design, rational engineering, and design sustainability in the omics ages.
(2) Synthetic biology-related multi-omics data integration and data-mining techniques.
(3) The technical issues, development issues, and safety issues of synthetic biology.
(4) Data resources, web services, and visualizations for synthetic biology.
(5) Advancement in concrete research on synthetic biology with several case studies shown.

As a reference on synthetic biology research and education in the omics age, this book focuses on systematic design, rational engineering, and design sustainability for synthetic biology, which will explain in detail and with supportive examples the "What," "Why," and "How" of the topic.

This book attempts to bridge the gap between synthetic biology's research and education side, for best practice of synthetic biology and iGEM, and for in-depth insights towards related questions. In reading this book, we hope that the readers could not only gain information about the biological resources, databases, and tools available for them to conduct the synthetic biology project, but also learn how to conduct a rational synthetic biology study, and be reminded about technical, ethic, and safety concerns.

So, let's begin.

Wuhan, China Kang Ning
Wuhan, China Yi Zhan

About the Book

This book focuses on the synthetic biology and iGEM competition, mainly focusing on techniques, development, and safety concerns from the following aspects: (1) Introduction to synthetic biology and iGEM, primarily focusing on systematic design, rational engineering, and design sustainability in the omics ages. (2) Synthetic biology-related multi-omics data integration and data-mining techniques. (3) The technical issues, development issues, and safety issues of synthetic biology. (4) Data resources, web services, and visualizations for synthetic biology. (5) Advancement in concrete research on synthetic biology, with several case studies shown.

Synthetic biology and iGEM represent very new research fields, which could be profoundly boosted by omics technologies. As a reference on synthetic biology and iGEM in the age of omics, this book focuses on systematic design, rational engineering, and design sustainability for synthetic biology, which will be explained in detail and with supportive examples of the "What," "Why," and "How" of omics on synthetic biology and iGEM-related researches. It attempts to bridge the gap between synthetic biology and iGEM, and the data-mining techniques, for the best practice of contemporary bioinformatics and in-depth insights for the synthetic biology and iGEM-related questions.

Contents

1 Introduction to Synthetic Biology 1
Dan Zhao and Kang Ning

2 iGEM: The Competition on Synthetic Biology 23
Yi Zhan, Kang Ning, and Dan Zhao

3 Synthetic Biology-Related Multiomics Data Integration and Data Mining Techniques 31
Kang Ning and Yuxue Li

4 Synthetic Biology: Technical Issues 39
Bohan Wang, Xiunan Huo, Xianglei Zhang, Yuanhao Liang, Yingying Yang, Jiacheng Shi, Xinyu Huan, Xilin Hou, Weilin Lv, and Yi Zhan

5 Synthetic Biology: Development Issues 63
Kang Ning, Yi Zhan, and Dan Zhao

6 Synthetic Biology: Safety Issues 71
Xue Zhu, Dan Zhao, and Kang Ning

7 Synthetic Biology: Data Resources, Web Services, and Visualizations 81
Yuzhu Zhang and Yi Zhan

8 Synthetic Biology: Case Studies 99
Pei Liu, Yi Zhan, and Kang Ning

9 Concluding Remarks 107
Dan Zhao and Kang Ning

Appendix 111

About the Editors

Kang Ning Professor of Microbial Bioinformatics Group and Director of the Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology. Dr. Ning obtained his BS in Computer Science from USTC, and PhD in Bioinformatics from NUS. He obtained his Post-Doc training in Bioinformatics at the University of Michigan. He has been devoting himself to bioinformatics research for more than 20 years focusing on omics big-data integration, microbiome analyses, and single-cell analyses. His current research interests include AI methods for multi-omics, especially metagenomics data-mining and their applications. He is also interested in synthetic biology and high-performance computation. Dr. Ning as the corresponding author, published over 100 research articles and reviews on leading journals including *PNAS*, *Gut*, *Genome Biology*, *Nucleic Acids Research*, and *Bioinformatics*, with more than 5,000 citations in total. He is the committee member of several national bioinformatics and biology big-data committees in China. He serves as an editorial board member of the journal *Genomics Proteomics and Bioinformatics* and *Scientific Reports* and served as reviewers for several international funding agencies including UK-BBSRC and UK-NERC.

Yi Zhan Associate Professor and Deputy Chair of the Committee of College of Life Science and Technology, Huazhong University of Science and Technology. Dr. Zhan obtained his BS in Biotechnology and PhD in Biophysics from Huazhong University of Science and Technology. He obtained his Post-Doc training in Molecular Neuroscience at Hong Kong University of Science and Technology. He has been devoting himself to molecular biology research for more than 20 years focusing on the structure and function of proteins, especially membrane proteins in neuroscience. He is also interested in synthetic biology and its application in the environment, healthcare, and manufacturing. Dr. Zhan served as judge in iGEM jamborees for several years and also served as team PI and instructors for iGEM team HUST-China for over 10 years. He has published several research articles in leading journals including *Nature Communications*, *PNAS*, etc. In some of these articles

published on *Small*, *ACS Synthetic Biology*, *BMC Biology*, and *Scientific Reports*, except for the corresponding author Dr. Zhan and his faculty colleagues, the other authors are all undergraduate students from competition teams. He is now devoting his efforts to the education and training of undergraduate students studying life science.

Chapter 1
Introduction to Synthetic Biology

Dan Zhao and Kang Ning

Abstract Synthetic biology is a multidisciplinary field of research that aims to develop new biological components, devices, and systems, as well as redesign existing natural systems. Synthetic biology is involved in many aspects of human life, such as processing information, manipulating chemicals, providing food, fabricating materials and structures, producing energy, as well as maintaining and enhancing human health and our environment. Advances in omics technologies have profound effects on synthetic biology research, pushing synthetic biology research toward rational data-driven and model-driven science. The ultimate goals of synthetic biology are to design and build engineered biological systems. Furthermore, synthetic biology has a broad area for applications, including sustainable food, sustainable energy, sustainable environment, etc. In this chapter, we will introduce the history, techniques, applications, and future challenges of synthetic biology.

Keywords Synthetic biology · History · Techniques · Applications · Frontiers · Future challenges

Advances in omics technologies have profound effects on synthetic biology research [1], pushing synthetic biology research toward rational data-driven and model-driven science.

1.1 The History and Importance of Synthetic Biology

Synthetic biology is a multidisciplinary field of research that aims to develop new biological components, devices, and systems, as well as redesign existing natural systems [2]. It is a field of study that includes a wide range of approaches from other

D. Zhao · K. Ning (✉)
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: ningkang@hust.edu.cn

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8_1

disciplines, including biotechnology, genetic engineering, molecular biology, molecular engineering, systems biology, membrane science, biophysics, and chemical and biological engineering, etc. Due to improved genetic engineering techniques and lower costs for DNA synthesis and sequencing, it is expanding quickly and has the potential to transform biotechnology and medicine [3].

The ultimate goals of synthetic biology are to design and build engineered biological systems. Synthetic biology is involved in many aspects of human life, such as processing information, manipulating chemicals, providing food, fabricating materials and structures, producing energy, and maintaining and enhancing human health and our environment. Additionally, synthetic biology is an inalienable component of systems biology [1], as well as biological engineering [4], serving as a link between these two more mature research areas. Furthermore, synthetic biology has a broad area for applications, including sustainable food, sustainable energy, sustainable environment, etc.

1.1.1 Timeline for the Development of Synthetic Biology (Fig. 1.1)

1910 *Theorie physico-chimique de la vie et générations spontanées* by Stephane Leduc has the earliest recorded use of the term "synthetic biology." Additionally, he made note of this phrase in 1912's *La Biologie Synthétique* [5].

1970–1980 In 1973, the first molecular cloning and amplification of DNA in a plasmid was published in *P.N.A.S.* by Cohen, Boyer et al., constituting the dawn of synthetic biology. Arber, Nathans, and Smith won the 1978 Nobel Prize in Physiology or Medicine for discovering restriction enzymes. Szybalski published the following editorial in *Gene*: This finding ushered in a new era of synthetic biology where novel gene arrangements may be made and examined, making it easier to create recombinant DNA molecules and investigate individual genes [6].

1980–1990 Mullis et al. published the first polymerase chain reaction (PCR)-based DNA amplification utilizing a thermostable DNA polymerase in *Science* in 1988 [7]. As a result, DNA mutagenesis and assembly were considerably facilitated by eliminating the need to introduce fresh DNA polymerase following each PCR cycle.

2000–2010 Two publications published in *Nature* in 2000 explored the development of synthetic biological circuit devices, including a genetic toggle switch and a biological clock, by fusing genes in *Escherichia coli* cells. The Massachusetts Institute of Technology in the United States hosted Synthetic Biology 1.0 (SB1.0), the first international conference on synthetic biology, in 2004. Researchers created a light-sensing circuit in *E. coli* in 2005. Another team developed multicellular pattern-forming circuits. Researchers created a synthetic circuit in 2006 that encourages bacterial invasion of tumor cells as a new therapeutic technique for cancer treatment. *M. mycoides* JCVI-syn1.0, the first synthetic bacterial genome, was

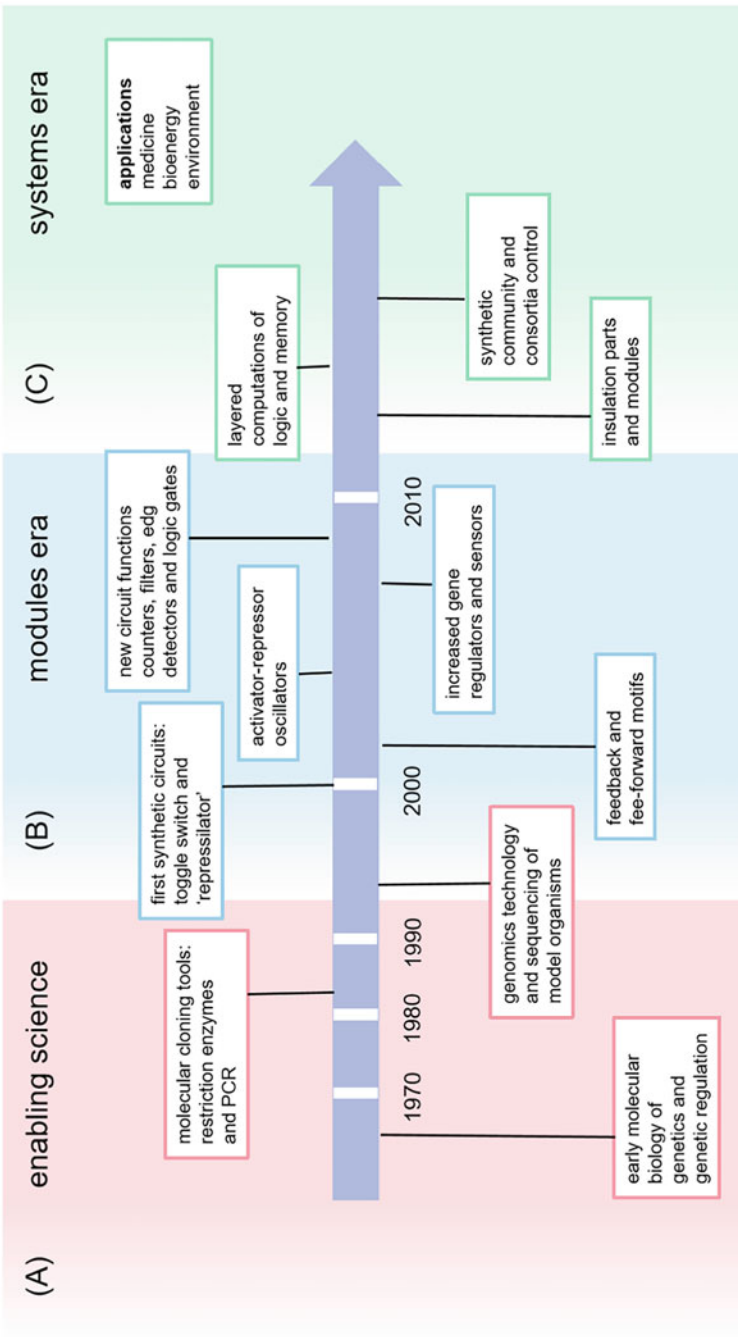

Fig. 1.1 Brief timeline for the development of synthetic biology. The development of systems biology includes (A) enabling science, (B) modules era, and (C) systems era. The development of various science and technology supports the development of systems biology, which can be applied to medicine, bioenergy, the environment, and many other aspects

published by researchers in *Science* in 2010. Using yeast recombination, chemically produced DNA is used to create the genome [8].

2010–2020 Functional synthetic chromosomal arms were created in yeast in 2011. The programming of CRISPR–Cas9 bacterial immunity for targeting DNA breakage was published in *Science* in 2012 by the Charpentier and Doudna laboratories [9]. Eukaryotic gene editing has been substantially facilitated and broadened by this technology. A related viable variant of *Caulobacter ethensis*-2.0 was not yet discovered when the first bacterial genome, *Caulobacter ethensis*-2.0, was announced to have been created in 2019 by researchers at ETH Zurich. A variant of the bacteria *E. coli* was created 1 month later by lowering the normal number of 64 codons in the bacterial genome to 59 codons to encode 20 amino acids. This resulted in the formation of a new synthetic (perhaps artificial) type of viable life. This is seen as a turning point for synthetic biology [10].

1.2 Synthetic Biology Techniques

Synthetic biology requires knowledge from multiple disciplines (Fig. 1.2), such as chemistry, microbiology, and bioinformatics.

The development of synthetic biology was largely dependent on a number of cutting-edge enabling technologies, in which DNA editing and reading are basic technologies (sequencing and fabrication).

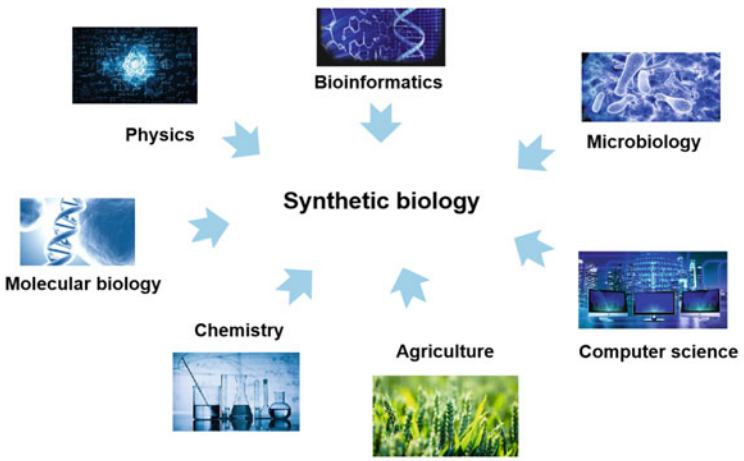

Fig. 1.2 Synthetic biology requires knowledge from multiple disciplines and has applications in a broad area. Synthetic biology requires knowledge from multiple disciplines, such as chemistry, microbiology, and bioinformatics, etc.

1.2.1 DNA and Gene Synthesis

The sizes of DNA constructs made from oligonucleotides ("oligos") have expanded to the genomic level, driven by a sharp decline in the cost of oligonucleotide synthesis. The CRISPR/Cas system is also known as a promising method of gene editing. The phrase "the most significant innovation in the field of synthetic biology in over 30 years" was used to describe it. The time for gene editing is sped up to weeks via CRISPR. However, because of its accessibility and ease of usage, it has generated ethical questions, particularly about its use in biohacking [11].

1.2.2 Sequencing

The order of the nucleotide bases of a DNA molecule is determined through DNA sequencing. DNA sequencing is a tool that synthetic biologists employ in a variety of ways. First, extensive attempts at genome sequencing continue to offer knowledge about naturally existing species. This knowledge offers synthetic biologists a rich base from which to build components and apparatus. Sequencing can also confirm that the manufactured system functions as planned. Third, quick detection and identification of synthetic systems and creatures can be made possible by inexpensive, dependable, and fast sequencing [12].

1.2.3 Microfluidics

A developing technique for creating novel components as well as characterizing and analyzing them is microfluidics, particularly droplet microfluidics. It is frequently used in screening tests [13].

1.2.4 Modularity

Tom Knight designed the BioBrick plasmids, which are the most widely used standardized DNA components, in 2003. Thousands of students from all around the world have participated in the international Genetically Engineered Machine (iGEM) competition using the BioBrick standard. Molecular motifs are a part of a larger network containing upstream and downstream elements in a live cell. These elements could change how well the modeling module can signal. When using ultrasensitive modules, a module's sensitivity contribution may be different from the sensitivity it maintains when used alone [14].

1.2.5 Modeling

By more accurately forecasting system behavior before fabrication, models assist in the design of engineered biological systems. Better models of biological processes, such as DNA encoding the information required to define the cell, biological molecules' ability to bind substrates and catalyze reactions, and the behavior of multicomponent integrated systems, all improve synthetic biology. Applications in synthetic biology are the main focus of multiscale models of gene regulation networks. All biomolecular interactions in transcription, translation, regulation, and induction of gene regulatory networks may be modeled using simulations [15].

1.2.6 Synthetic Transcription Factors

The elements of the DNA transcription mechanism have been the subject of research. Synthetic biological circuit designers want to be able to regulate how synthetic DNA is expressed in both single-celled creatures (prokaryotes) and multicellular organisms (eukaryotes). One experiment looked at how synthetic transcription factors (sTFs) may be tweaked in terms of their transcriptional output and their capacity to cooperate with other transcription factor complexes. Researchers were able to reduce the site-specific activity of sTFs by mutating their zinc fingers, the DNA-specific component of sTFs, to reduce their affinity for certain operator DNA sequence sites (usually transcriptional regulation) [16].

1.3 Synthetic Biology Applications

In combination with systems biology, synthetic biology provides a framework for applying recent discoveries in genomics, proteomics, and molecular biology. Applying engineering design ideas to biological systems is central to this production-oriented strategy. New genetic constructs that express a target product can be designed, constructed, and tested with the use of molecular biology techniques in the context of synthetic biology. The concept behind synthetic biology is that by applying engineering principles to the life sciences, we can significantly boost the validity of the "design–build–test–learn cycle" (Fig. 1.3) [17]. Any field of engineering must have a high degree of predictability and reproducibility.

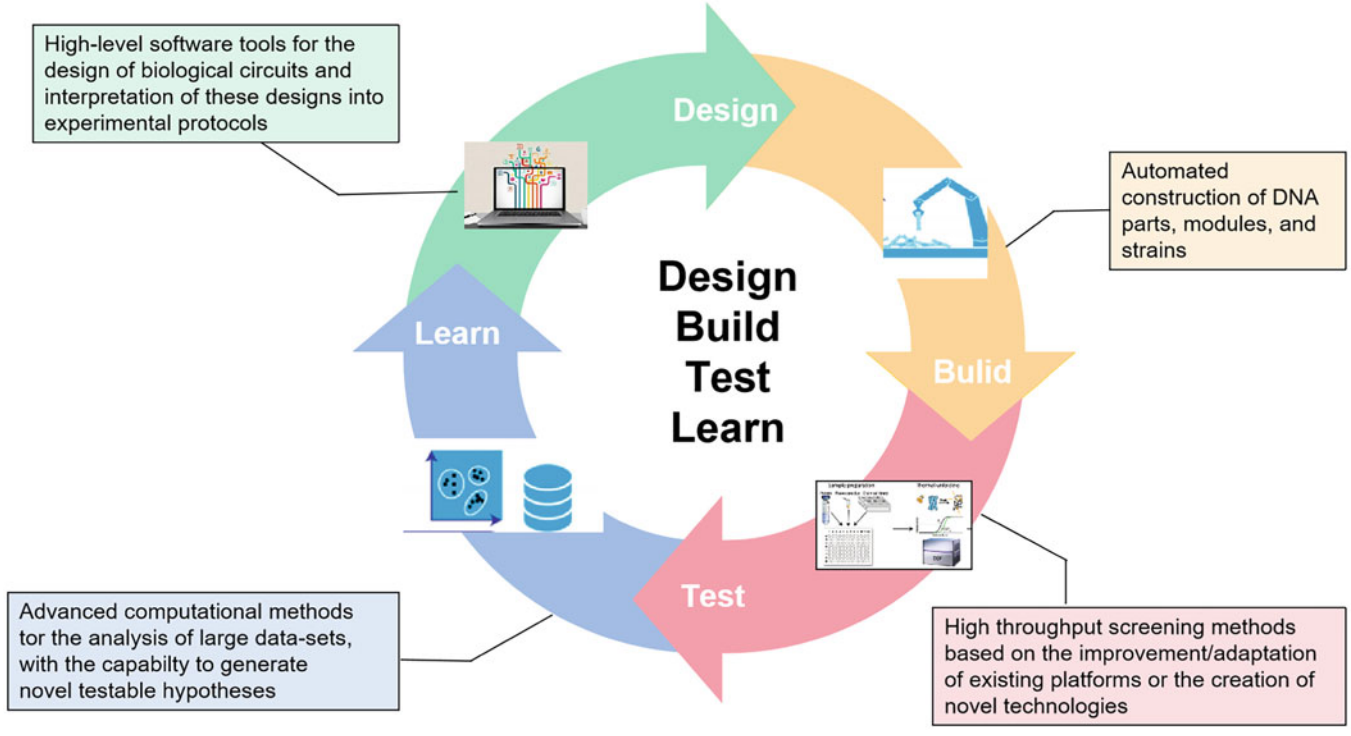

Fig. 1.3 Design–build–test–learn cycle at the heart of synthetic biology. The design relies on fully represented biological components and computer-aided methods, various advanced technologies and various combined library data to simplify the construction process of recombinant vectors, then relies on some analytical techniques for testing, and finally uses various computer modeling methods for simulation and optimization. The cycle was repeated

1.3.1 Biological Computers

The concept of the "biological computer" as a constructed biological system capable of performing computer-like functions is the current paradigm in the field of synthetic biology. Researchers have constructed and analyzed a wide range of logic gates across several species, demonstrating both analog and digital processing in real time within individual cells. That bacteria can be programmed to conduct analog and digital computing was proven [18].

1.3.2 Biosensors

Typically, bacterial biosensors are designed organisms with the ability to detect and report on environmental phenomena such as the presence of poisons or metals [19].

1.3.3 Cell Transformation

Gene circuits are networks of genes and proteins that allow cells to perform a wide range of tasks, including sensing and responding to their surroundings, making decisions, and exchanging information with one another. There are three main factors: gene circuits constructed from DNA, RNA, and synthetic biology that regulate gene expression at the transcriptional, post-transcriptional, and translational levels. Combinations of foreign genes and optimization through directed evolution have bolstered conventional metabolic engineering. Yeast and *E. coli* have been engineered to produce Artemisinin's precursor for commercial use. Furthermore, cells may be used as microscopic molecular foundries to make materials whose features are genetically encoded if synthetic biology and materials science are integrated [20].

1.3.4 Designed Proteins

Using directed evolution, we can create novel protein structures with the same or enhanced functionality as already existing proteins, demonstrating the potential for protein engineering to improve upon nature. Computational methods can also be used to create proteins with novel capabilities or specificities. Methods include computational enzyme design methods and molecular modeling techniques for sequence database mining. Experiments aimed at enlarging the amino acid pool beyond the standard set of 20 are also rather prevalent. It is possible to modify some codons so that they code for noncanonical amino acids such as O-methyl tyrosine or

for exogenous amino acids such as 4-fluorophenylalanine. Others have studied protein structure and function by narrowing the amino acid pool down from the typical 20. It is possible to create small libraries of protein sequences by producing proteins in which many consecutive amino acids have been substituted with single ones. In proteins, for instance, numerous nonpolar amino acids can be interchanged for a single nonpolar amino acid [21]. Researchers and companies practice synthetic biology to synthesize industrial enzymes with high activity, optimal yields, and effectiveness. These synthesized enzymes aim to improve products such as detergents and lactose-free dairy products, as well as make them more cost-effective [22].

1.3.5 Designed Nucleic Acid Systems

Scientists can digitally encode information onto a single strand of synthetic DNA. In a broader sense, the development of new genetic systems is facilitated by algorithms such as NUPACK [23], ViennaRNA [24], Ribosome Binding Site Calculator [25], Cello [26], and Non-Repetitive Parts Calculator, enabling the design of new genetic systems. The incorporation of synthetic nucleotides and amino acids into nucleic acids and proteins has also progressed greatly through the use of various in vitro and in vivo technologies.

1.3.6 Space Exploration

Recently, NASA became interested in synthetic biology because it may be used to assist in creating resources for astronauts using only a small selection of molecules shipped from Earth. Synthetic biology holds great promise for the establishment of self-sufficient human settlements in space, and on Mars in particular, it could lead to production processes based on local resources. Methods comparable to those used to promote crop tolerance to environmental conditions have been applied to the development of plant strains that could survive in Mars's harsh climate [27].

1.3.7 Synthetic Life

Synthetic life is a very active area of study in synthetic biology, and it focuses on the creation of hypothetical organisms in vitro from biomolecules and/or chemical analogs thereof. Experiments with synthetic life aim to learn more about where life came from, examine its characteristics, or, more ambitiously, generate life from inanimate (abiotic) parts. The goal of synthetic life biology is to design and build organisms that can perform useful tasks, such as producing drugs or cleaning up contaminated environments. New classes of medicines and diagnostic tools may be

developed with the help of designer biological parts, which have exciting potential in the field of medicine. The international Build-a-Cell large-scale research partnership for the manufacture of synthetic living cells began in 2017 and has since been followed by national synthetic cell organizations in numerous countries, including FabriCell, MaxSynBio, and BaSyC. In 2019, European efforts to synthesize cells were consolidated under the SynCellEU program [28].

1.3.8 Drug Delivery Platforms

Engineered Bacteria-Based Platform In the past, bacteria were employed to treat cancer. Both *Bifidobacterium* and *Clostridium* can selectively colonize tumors, reducing their volume. Researchers in synthetic biology have recently reprogrammed microorganisms to detect and react to a specific form of cancer. Bacteria are commonly utilized to transport therapeutic molecules to the site of a tumor, where they can have the fewest possible off-target effects. Bacterial surfaces were engineered to express peptides capable of recognizing tumor cells, allowing the bacteria to then target those cells. An affibody molecular target of human epidermal growth factor receptor 2 and a synthetic adhesin are employed [29]. The alternative strategy involves incorporating a logic gate into bacteria, which would then be able to detect changes in the tumor microenvironment, such as hypoxia [30]. The bacteria either lyse or secrete only the therapeutic compounds of interest to the tumor. The benefits of lysis include enhancement of the immune system and regulation of growth. A wide variety of secretion systems and other methods can be employed. These therapies employ a wide variety of species and strains.

Cell-Based Platform A cancer patient's immune system can be used against tumor cells. Immunotherapies are the primary focus of cell-based therapeutics, with a primary emphasis on manipulating T cells. Scientists have "taught" T-cell receptors to recognize cancer epitopes through genetic engineering. To activate and stimulate cell proliferation, chimeric antigen receptors (CARs) [31] are constructed from an antibody fragment linked to intracellular T-cell signaling domains. However, there are restrictions, such as the difficulty in inducing extensive DNA circuits into the cells and the hazards involved with introducing foreign components, notably proteins, into the cells.

1.4 The Synthetic Biology Frontiers

Several technical advances and application frontiers are worth mentioning.

1.4.1 Gene Editing

The appearance of gene-editing tools based on the CRISPR system quickly revolutionized the entire field of life sciences. In the past decade, from CRISPR/Cas9 [32] to Prime Editing [33] based on reverse transcription, efficient and precise editing tools will provide medical services, agriculture, and inroad prospects. However, gene-editing technology is not without weaknesses. First, several problems remain in gene editing: off-target effects (editing tools modify unexpected DNA/RNA sites), antibody problems (immune system resists foreign CRISPR protein), and delivery problems (lack of efficient, accurate, and heavy-loaded human delivery tools). These three major problems of gene editing still hinder its application. Additionally, the ethical and sociological challenges brought by gene editing, such as gene-editing babies and gene drives, need further attention.

1.4.2 Genome Engineering

In the past decade, there have been numerous extraordinary findings, ranging from the chemical synthesis of the *E. coli* genome to the simplest life Synthia 3.0 [34], from the manual replacement of 321 codons on the *E. coli* genome using MAGE + CAGE to the complete synthesis of the *E. coli* genome with only 61 codons, and from yeast gene combination plan Sc2.0 to artificial single chromosome yeast, the capabilities of genome synthesis, modification, and design continuously improve. In the future, we need less expensive and faster technology for gene assembly and software accompanied by automation, analysis, and design capabilities to eventually reveal all the mysteries of the genome.

1.4.3 De Novo Protein Design

Using computational tools to explore the sequential space that has never been explored before, thus regulating protein, which is the executor of life functions, has always been. In the past decade, the coming of age of de novo protein design by David Baker indicated that the age of entire artificial protein design is coming. From symmetrical protein design to artificially designed functional proteins to the biological regulatory network composed of artificially designed proteins, we are actually learning from nature and designing life execution tools that have never been involved in evolution [35].

1.4.4 Synthesis of Unnatural Materials

Twenty years ago, Peter Schultz formally introduced synthetic amino acids into cells. In the past decade, we have witnessed how to introduce synthetic substances into organisms using evolutionary and synthetic chemistry. We have also realized complete pathways from synthetic DNA to synthetic RNA and finally proteins containing synthetic amino acids. Meanwhile, many findings are currently breaking the barriers between different species and different life mechanisms, acting as a solid foundation to create more abundant biological tools [36].

1.4.5 Design of the Gene Circuit

Ten years ago, Gibson cloning technology arose and revolutionized the entire field of molecular cloning construction. A more predictable design guided by models has always been one of the primary goals that synthetic biology is pursuing. In the past decade, we have developed various regulatory tools, applied a variety of principles used for biological circuit design, and designed much more complicated genetic circuits. However, calling for novel design concepts of gene circuits is ongoing due to several hurdles, such as cell burden and toxicity [37].

1.4.6 Directed Evolution

Evolution is the most unique, magical, and powerful tool of the life system. In the past decade, the evolutionary events that have drawn the most attention are newly invented evolutionary tools such as MEGA11 (Molecular Evolutionary Genetics Analysis) [38] and the abundant biological tools and evolutionary achievements beyond imagination brought by directed evolution.

1.4.7 Cell Factory/Therapy

Microbes are great cell factories. In the past decade, we have witnessed the potential of cell factories to synthesize high-value-added and market-competitive products from heterogeneous sources. We have also witnessed the biological production of chemical molecules such as artemisinin, farnesene, opioids, and cannabis. More capital is concentrating on artificial meat. Synthetic biology seems to need more socially valuable industrial products to bring cell-based industrial production lines back to the public's vision. Meanwhile, therapies based on microbes (https://www.synlogictx.com/), immune cells, and stem cells are bringing new opportunities to the

medical field. By introducing engineering design concepts into cell therapy, we could design smarter and less expensive cell machines to solve numerous disease challenges that humans are facing [39].

1.4.8 Big Data-Assisted Biological Design

In the past decade, along with the revival of machine learning and neural network algorithms and the rapid accumulation of high-throughput biological data, an increasing number of data analysis algorithms are assisting synthetic biologists in design, such as obtaining interpretable biological models via data analysis or using machine learning to assist protein design [40]. High-throughput, high-quality data and highly interpretable algorithm models are becoming the first choice of researchers [41].

1.4.9 Education of Synthetic Biology

In the past decade, the International Genetically Engineered Machine competition (iGEM) has set off a new round of synthetic biology craze in universities worldwide. Every year, thousands of outstanding students gather in Boston to fight for gold medals. The concept of synthetic biology brought by iGEM is bringing new possibilities to STEM education and training more talents for the development of synthetic biology [42].

1.4.10 Synthetic Biology Community

International cooperation projects represented by Sc2.0 (http://syntheticyeast.org/), Global Biofoundries Alliance [43], etc., are promoting rapid progress in the field of synthetic biology. The values of sharing and cooperation advocated by the community of synthetic biology will offer more development opportunities in the future.

1.5 The Future Challenges for Synthetic Biology

After 20 years of development, synthetic biology is now a fully developed discipline that is propelling major advancements in the bioeconomy and the fields of medicine and biology. From large-scale efforts to create synthetic life, cell simulators, and custom genomes to novel approaches to engineering biology through automation, deep learning, and evolutionary control, ten technological advances are discussed

that are expected and hoped to come from the next generation of research and investment in synthetic biology.

1.5.1 Automation and Industrialization

Because of its emphasis on reusability and simple DNA assembly, synthetic biology has long advocated for the standardization of biological components [44]. The downstream benefit of such uniformity is the ability to automate DNA creation and scale up this procedure to produce vast numbers of modified cells that can be evaluated simultaneously. The field has been working toward this "industrialization" of the constructing and testing process for quite some time, although it is still not common practice among most research groups. Most of the necessary tools for automating the design–build–test–learn (DBTL) cycle are already in place, particularly in biofoundries and at large corporations. The selection of parts and the creation of genetic constructs are made easier by repositories of characterized parts and various computer-aided design (CAD) tools. High-precision liquid-handling robots simplify the complex combination of reagents needed for a single experiment by transferring volumes of liquid as small as a few microliters and as large as a few picoliters. Experimental reproducibility is demonstrated by sandim. Computer programs for managing data standardize laboratory procedures. Big data are processed using statistical analysis programs to reveal patterns, and the DBTL loop can be completed with the assistance of DoE tactics, which aim to improve subsequent experiments [45].

1.5.2 Deep Learning for DNA Design

Some of deep learning's earliest successes in bioscience include its application to the classification of microscope images for various purposes, including predicting protein structure and the prediction of the molecular structures of drugs that function as antibiotics. On the other hand, DNA design is where deep learning is most likely to have the most impact on synthetic biology. The foundation of synthetic biology rests on the ability to design DNA. Synthetic biology has its own benefits in DNA design compared to earlier sciences. Because of developments in genome editing and DNA synthesis, it is no longer necessary to rely just on reading in gathered data; instead, reading and writing are both possible. This allows for the production of more informative training data, which may be used to stress-test a model's internal representation of a system and embed a more complete understanding. Active learning is a machine learning technique that may be easily included in automated workflows to find the optimal next set of perturbations to supplement a learning model. Thus, deep learning models may one day assist us in abandoning the practice of viewing DNA sequences as priceless natural artifacts to be standardized and

employed as building blocks in biofoundry-created assemblies. Instead, we should be able to use deep learning to generate optimal DNA sequences for certain permutations of genetic components and genetic contexts by simply issuing high-level commands. What we know about the gradual improvement of DNA sequences through natural selection may also help deep learning. Machine learning mimics biological evolution by emulating the process of trial-and-error mutation leading to progress. However, what about other elements of biological evolution, such as error detection and correction and sequence recombination? As we learn more about how to steer DNA evolution (described below), we can also anticipate design improvements as a result of this feedback [46].

1.5.3 Designing with Whole-Cell Simulations

A simulation of the behavior of all the genes and proteins in the bacteria *Mycoplasma genitalium*, which has a genome of approximately 500 genes, was created using such data for the first time nearly a decade ago by researchers focusing on natural minimum cells. They created mathematical models for each essential cellular activity, parameterized them with omics data, and then determined how to include the models in a running simulation of the cell cycle. This systems biology accomplishment aided in illuminating cellular resource utilization and, perhaps most excitingly for synthetic biology, it predicted the effects of deleting or adding genes to the genome on the organism [47].

1.5.4 Biosensing: Detecting Anything, Anywhere

The incorporation of natural sensing and response mechanisms into created cells to create biosensors has been a mainstay of synthetic biology efforts for over 20 years. When you think about all the sensing that biology is already doing all across the earth, it is hard not to feel like biosensing has untapped potential. The ability of living cells to sense their environment is a capability that humanity should do everything it can to tap into. There are countless ways in which the ability to detect anything, anyplace, will change our world for the better, including research, pandemic preparedness, and gauging our own and the planet's health [48].

1.5.5 Real-Time Precise Control of Evolution

To date, the majority of synthetic biology efforts have focused on designing and building a cell or strain for use in an application (such as biosynthesis or biosensing). Once we put our "finished product" into use, we either cross our fingers and hope

that natural processes such as mutation and selection do not have any effect on it, or we try to minimize the likelihood of this happening as much as possible throughout the design process. However, living things are constantly changing as a result of natural processes such as mutation and natural selection. Therefore, it is not possible to ignore or fully block evolution, even on tiny scales in controlled conditions. Rather, we need to figure out how to harness evolution while engineering biology, learning to regulate and direct the fate of designed genes, cells, and organisms in varied contexts and for various purposes [49].

The biotechnological resources essential to this development are currently in the process of emerging. We can now design, direct, and control mutations to specific sites and genes within live systems using CRISPR-based systems and related innovations such as MAGE. Predictions of how changes in DNA and amino acids influence genes, gene regulation, and the form and function of proteins are also becoming possible thanks to deep-learning models. Hundreds of enhanced and unique enzymes have been produced by directed evolution, helping to better adapt hosts for metabolic biosynthesis and even improve gene circuits. In addition, there are currently available molecular tools for continuous in vivo regulated evolution of target genes.

1.5.6 Cellular Communities and Multicellularity

The effective execution of the "division of labor" in synthetic multicellular systems is a significant barrier for synthetic biology. One of the main aims of synthetic biology thus far has been to build a specific function into each cell of a population. After completing research on simple multicellular functions such as programmed feature extraction in bacteria and population-based oscillators, groups have announced progress toward more complicated multicellular functions such as consortia-based computation, molecular computing designs, and synthetic morphogenesis [50].

There is no question about the tremendous potential in synthetic biology, which was offered by engineering systems with numerous interacting cells, but they do not have to be confined to merely co-cultures or co-dependent communities. Opportunities in systems in which cells physically join together and differentiate to accomplish specialized functions in tissues, organs, and bodies are already evident in the complexity, robustness, and multifunctionality of natural multicellular creatures such as plants and animals. We are getting better at programming cells to adhere to one another and grow in specific patterns by writing synthetic differentiation programmers out of DNA components. It would be great to advance the rational engineering of prototissues and organoids in tandem with the advancement toward predictable engineering of co-cultures and consortia.

1.5.7 Custom and Dynamic Synthetic Genomes

Synthetic genomes are possibly the next logical step for large-scale synthetic biology research. The DNA costs and labor needed, while still enormous, have dropped sufficiently for individual teams to make megabase chromosomes now that a common approach to manufacturing synthetic genomes has begun to emerge. Within the next decade, synthetic genomics will begin to tackle the considerably larger genomes of multicellular creatures such as mammalian cells and plants. In the near future, bacterial and yeast chromosome synthesis projects will become the objective of various organizations. Current synthetic genome studies attempt to provide new information, such as insights into the genome's coding, content, and organization, all of which are difficult to dissect using conventional approaches. Future advances in DNA synthesis and DNA assembly automation and scalability will make it easier to create synthetic genomes for specific purposes [51].

Current synthetic genome studies attempt to provide new information, such as insights into the genome's coding, content, and organization, all of which are difficult to analyze using traditional techniques. The difficulty will change to creating synthetic genomes with specific applications in mind when the cost of DNA synthesis decreases and automated, scalable DNA assembly becomes the standard. Reduced-codon-usage genomes are already optimized for use in cases of genetic code extension, such as the incorporation of noncanonical amino acids into proteins. The next step would likely be the creation of genomes designed for certain industrial purposes, such as metabolism optimized for biosynthesis research. Customizing these to focus just on the genes essential for specific activities may be beneficial when the research approaches the enormous genomes of multicellular organisms.

1.5.8 Artificial Cells

While extensive work has gone into developing new genomes for existing cells, a more fundamental strategy for creating synthetic cells is emerging from various fields of chemistry and biochemistry research that aims to replicate the essential characteristics of a living cell purely by connecting biochemical elements. The objective is to create synthetic, programmable replicas of the fundamental cells found in biology by creating artificial cells from the bottom up. This may sound like a long-term goal, but the groundbreaking in artificial cell research over the past 10 years has resolved a number of practical problems, creating a fertile field for initiatives that are both more intricate and exciting [52].

Artificial cell research is primarily motivated by the concept of understanding by creation, which has been a guiding principle in synthetic biology. By building a self-replicating organism out of inert molecules, we may establish a synthetic cell that will assist us in defining the distinction between the nonliving and the living. However, on a more theoretical level, it will also aid in our understanding of how

natural mechanisms are effective. We can already examine difficult-to-assess aspects of molecular biology using artificial cells, such as lipid vesicle protocells [53], such as the minute impacts of often confusing factors in a cell, such as molecular concentration fluctuations and macromolecular crowding. It is an exciting time to conduct synthetic biology in protocells, and this field promises to mature and help us better investigate the engineerability of simplified biological systems. Supporting fields such as microfluidics, chemical compartmentalization, and DNA synthesis are advancing at an accelerated rate.

1.5.9 Materials with DNA-Encoded Properties

Engineered living materials (ELMs) are a rapidly growing area that is a logical progression from cellular communities [54]. Cells being used as chemical factories have been a key element in some of synthetic biology's largest breakthroughs and encouraged growth in areas where they intersect with other businesses, such as the food and fashion sectors. Advanced architecture actively uses materials such as bacterial cellulose, spider silk, and mushrooms as the foundation for sustainable fabrics, furniture, and even building materials. Furthermore, innovative iGEM initiatives are helping to expand the variety of materials that may be generated at scale by bacteria.

We can eventually reach a position where the characteristics and uses of materials generated in this way can be specified and programmed by modular DNA-encoded programmers inserted inside the cells that make them, assuming synthetic biology can approach this enormous advancement with a uniform framework. We can already observe this in plants, where each cell's genome includes the DNA programs that allow the cell and the tissue around it to develop into a variety of different materials, from delicate flower petals to rigid nut shells [55].

1.5.10 Engineered Organisms for Sustainability Goals

Biotechnology has the power to improve our living environment and health as the global population continues to grow, the environment is seriously polluted, and the climate is constantly changing. Under these circumstances, human food and environmental security face great challenges. The use of synthetic biology techniques to create cell factories suitable for the food industry to transform renewable raw materials into important food components is seen as a viable way to solve the food problem. In agricultural production, the problems of soil compaction and acidification caused by the large increase in the use of nitrogen fertilizer can be effectively solved by synthetic biology "microbial nitrogen fixation" technology. In the field of environmental governance, "customized" microorganisms can be used to remove organic pollutants that are difficult to degrade. Artificial microbial sensors can also

be developed to help humans monitor the environment, and microorganisms capable of identifying and enriching heavy metal pollutants such as cadmium, mercury, and arsenic in soil or water can be designed and constructed to greatly improve the efficiency of pollution control.

Synthetic biology also has broad applications in the field of life and health, not only in the production of natural products and other medical products but also in the development of disease research models, biomarker monitoring, stem cells, and regenerative medicine and other fields.

1.6 Synthetic Biology Case Studies

Several successful cases of synthetic biology are worth mentioning. One of the primary goals of synthetic biology is to empower cells with new functions via the design of gene circuits, while the current development of the entire field mainly focuses on gene-level regulatory design. However, cell functions are realized not only by gene-level regulation but also by protein-level interaction. Cells use the interaction between proteins to regulate many parameters, such as the activity, location, and stability of a specific protein.

Compared with traditional transcription regulatory circuits, synthetic protein circuits have more advantages, such as faster regulatory processes and more direct coupling with endogenous signal pathways. Circuits of synthetic biology at the protein level could help us design and monitor cellular behaviors. The composable protein–protein regulatory system will promote the precise design of protein circuits, in which various protein components can form diverse regulatory circuits via regulation.

Micheal Elowitz [56] demonstrated that engineered viral proteases could function as composable protein components. This combination could realize several functions in mammalian cells. This system is called CHOMP (circuits of hacked orthogonal modular proteases), in which the protease treated as the input signal could structurally dock with and cleave the target protease, thereby inhibiting the expression of the target protease. This combination of protein components could realize cascaded control, binary logic gates, and dynamic simulation of signal processing. To prove the utility of this system, the researchers designed a synthetic biological circuit that can induce cell death in response to the upstream activator of the Ras oncogene. They pointed out that CHOMP could perform complex functions, be encoded as a single transcript and be integrated into cells without genome integration, which will strongly promote the application of protein regulatory circuits in the area of biotechnology.The calculation process of modern computers is mainly based on sequential logic, in which the state of the circuit depends on the current input signal and the past input signal (memory). Implanting sequential logic into living cells enables them to perform corresponding biological processes in the form of discrete states. However, there is a huge challenge, that is, the realization of sequential logic requires feedback regulation implanted into the gene circuit, which is difficult to design and scale-up.

A study from Christopher Voigt [57] proposed a quantitative method for designing sequential logic gene circuits. This method uses the "NOT" logic gate as the core unit of regulation, which refers to a series of reactions: the input promoter drives the expression of the repressor protein, and the repressor protease is used to close the output promoter. The researchers characterized each NOT logic gate in detail by measuring the response function. At the same time, tools from nonlinear dynamics are applied to predict how combining gates leads to multiple steady states and dynamics. In this work, the researchers applied sequential logic to the regulation of cell checkpoints; only when the correct signal appears would the cell turn to the next state. The designed gene circuit in this research could guide E. coli to switch between linear and cyclic states. Based on simple rules, this study shows us a quantitative method for realizing sequential logic gene circuits in cells by combining reliable regulation logic. This method is beneficial to the design of the corresponding automated software, and in return, software could use these rules to build a larger-scale gene circuit composed of different logic gates. This provides a feasible approach for establishing regulatory gene networks with feedback loops.

References

1. Patra, P., et al.: Recent advances in systems and synthetic biology approaches for developing novel cell-factories in non-conventional yeasts. Biotechnol. Adv. **47**, 107695 (2021)
2. Fletcher, L., Rosser, S., Elfick, A.: Exploring synthetic and systems biology at the University of Edinburgh. Biochem. Soc. Trans. **44**(3), 692–695 (2016)
3. The future is synthetic biology. Cell. **175**(4), 895–897 (2018)
4. Toda, S., et al.: Engineering synthetic morphogen systems that can program multicellular patterning. Science. **370**(6514), 327–331 (2020)
5. Arrabito, G., et al.: Artificial biosystems by printing biology. Small. **16**(27), e1907691 (2020)
6. Saiki, R.K., et al.: Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. **239**(4839), 487–491 (1988)
7. Green, M.R., Sambrook, J.: The basic polymerase chain reaction (PCR). Cold Spring Harb Protoc. **2018**, 5 (2018)
8. Sleator, R.D.: The story of mycoplasma mycoides JCVI-syn1.0: the forty million dollar microbe. Bioeng Bugs. **1**(4), 229–230 (2010)
9. Ma, Y., Zhang, L., Huang, X.: Genome modification by CRISPR/Cas9. FEBS J. **281**(23), 5186–5193 (2014)
10. Joshi, K.K., Chien, P.: Regulated proteolysis in bacteria: Caulobacter. Annu. Rev. Genet. **50**, 423–445 (2016)
11. Roberts, T.C., Langer, R., Wood, M.J.A.: Advances in oligonucleotide drug delivery. Nat. Rev. Drug Discov. **19**(10), 673–694 (2020)
12. Mardis, E.R.: Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. **9**, 387–402 (2008)
13. Liu, Z., et al.: Microfluidics for production of particles: mechanism, methodology, and applications. Small. **16**(9), e1904673 (2020)
14. Matsumura, I.: Methylase-assisted subcloning for high throughput BioBrick assembly. PeerJ. **8**, e9841 (2020)
15. Lemire, S., Yehl, K.M., Lu, T.K.: Phage-based applications in synthetic biology. Annu. Rev. Virol. **5**(1), 453–476 (2018)

16. Pandelakis, M., Delgado, E., Ebrahimkhani, M.R.: CRISPR-based synthetic transcription factors in vivo: the future of therapeutic cellular programming. Cell. Syst. **10**(1), 1–14 (2020)
17. Jessop-Fabre, M.M., Sonnenschein, N.: Improving reproducibility in synthetic biology. Front. Bioeng. Biotechnol. **7**, 18 (2019)
18. Emani, P.S., et al.: Quantum computing at the frontiers of biological sciences. Nat. Methods. **18**(7), 701–709 (2021)
19. Dalila, R.N., et al.: Current and future envision on developing biosensors aided by 2D molybdenum disulfide (MoS2) productions. Biosens. Bioelectron. **132**, 248–264 (2019)
20. Nettesheim, P., Barrett, J.C.: Tracheal epithelial cell transformation: a model system for studies on neoplastic progression. Crit. Rev. Toxicol. **12**(3), 215–239 (1984)
21. Kamtekar, S., et al.: Protein design by binary patterning of polar and nonpolar amino acids. Science. **262**(5140), 1680–1685 (1993)
22. Gibney, B.R., Rabanal, F., Dutton, P.L.: Synthesis of novel proteins. Curr. Opin. Chem. Biol. **1**(4), 537–542 (1997)
23. Zadeh, J.N., et al.: NUPACK: analysis and design of nucleic acid systems. J. Comput. Chem. **32**(1), 170–173 (2011)
24. Lorenz, R., et al.: ViennaRNA Package 2.0. Algorithms Mol. Biol. **6**, 26 (2011)
25. Salis, H.M.: The ribosome binding site calculator. Methods Enzymol. **498**, 19–42 (2011)
26. Bernstein, M.N., et al.: CellO: comprehensive and hierarchical cell type classification of human cells with the cell ontology. iScience. **24**(1), 101913 (2021)
27. Sleator, R.D., Smith, N.: Terraforming: synthetic biology's final frontier. Arch. Microbiol. **201**(6), 855–862 (2019)
28. Fernau, S., Braun, M., Dabrock, P.: What is (synthetic) life? Basic concepts of life in synthetic biology. PLoS One. **15**(7), e0235808 (2020)
29. Cao, Z., Lin, S., Liu, J.: Bacteria-based microdevices for the oral delivery of macromolecules. Pharmaceutics. **13**, 10 (2021)
30. MacIntyre, N.R.: Tissue hypoxia: implications for the respiratory clinician. Respir. Care. **59**(10), 1590–1596 (2014)
31. Keane, J.T., Posey Jr., A.D.: Chimeric antigen receptors expand the repertoire of antigenic macromolecules for cellular immunity. Cell. **10**, 12 (2021)
32. Jinek, M., et al.: A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science. **337**(6096), 816–821 (2012)
33. Anzalone, A.V., et al.: Search-and-replace genome editing without double-strand breaks or donor DNA. Nature. **576**(7785), 149–157 (2019)
34. Peng, D., et al.: Global and local texture randomization for synthetic-to-real semantic segmentation. IEEE Trans. Image Process. **30**, 6594–6608 (2021)
35. Quijano-Rubio, A., et al.: De novo design of modular and tunable protein biosensors. Nature. **591**(7850), 482–487 (2021)
36. Yang, Z.P., Freas, D.J., Fu, G.C.: Asymmetric synthesis of protected unnatural alpha-amino acids via enantioconvergent nickel-catalyzed cross-coupling. J. Am. Chem. Soc. **143**(23), 8614–8618 (2021)
37. Wu, F., Zhang, Q., Wang, X.: Design of adjacent transcriptional regions to tune gene expression and facilitate circuit construction. Cell Syst. **6**(2), 206–215 e6 (2018)
38. Tamura, K., Stecher, G., Kumar, S.: MEGA11: molecular evolutionary genetics analysis version 11. Mol. Biol. Evol. **38**(7), 3022–3027 (2021)
39. Dong, Y., et al.: Genome annotation of disease-causing microorganisms. Brief. Bioinform. **22**(2), 845–854 (2021)
40. Yang, K.K., Wu, Z., Arnold, F.H.: Machine-learning-guided directed evolution for protein engineering. Nat. Methods. **16**(8), 687–694 (2019)
41. Zhu, H.: Big data and artificial intelligence modeling for drug discovery. Annu. Rev. Pharmacol. Toxicol. **60**, 573–589 (2020)

42. Lu, Y., et al.: Bringing scientific research education closer to undergraduates through international genetically engineered machine competition. Sheng Wu Gong Cheng Xue Bao. **34**(12), 1923–1930 (2018)
43. Hillson, N., et al.: Building a global alliance of biofoundries. Nat. Commun. **10**(1), 2040 (2019)
44. Schneider, G.: Automating drug discovery. Nat. Rev. Drug Discov. **17**(2), 97–113 (2018)
45. Fenster, J.A., Eckert, C.A.: High-throughput functional genomics for energy production. Curr. Opin. Biotechnol. **67**, 7–14 (2021)
46. Routhier, E., Bin Kamruddin, A., Mozziconacci, J.: keras_dna: a wrapper for fast implementation of deep learning models in genomics. Bioinformatics. **37**(11), 1593–1594 (2021)
47. Rees-Garbutt, J., et al.: Designing minimal genomes using whole-cell models. Nat. Commun. **11**(1), 836 (2020)
48. Kim, Y., Gonzales, J., Zheng, Y.: Sensitivity-enhancing strategies in optical biosensing. Small. **17**(4), e2004988 (2021)
49. Garcia De Viedma, D., Perez-Lago, L.: The evolution of genotyping strategies to detect, analyze, and control transmission of tuberculosis. Microbiol. Spectr. **6**, 5 (2018)
50. Parsek, M.R., Tolker-Nielsen, T.: Pattern formation in Pseudomonas aeruginosa biofilms. Curr. Opin. Microbiol. **11**(6), 560–566 (2008)
51. Choi, Y.R., et al.: A genome-engineered bioartificial implant for autoregulated anticytokine drug delivery. Sci. Adv. **7**(36), eabj1414 (2021)
52. Rhodes, K.R., Green, J.J.: Nanoscale artificial antigen presenting cells for cancer immunotherapy. Mol. Immunol. **98**, 13–18 (2018)
53. Omidvar, R., Romer, W.: Glycan-decorated protocells: novel features for rebuilding cellular processes. Interface Focus. **9**(2), 20180084 (2019)
54. Srubar 3rd, W.V.: Engineered living materials: taxonomies and emerging trends. Trends Biotechnol. **39**(6), 574–583 (2021)
55. Song, Y., Li, X.: Evolution of the selection methods of DNA-encoded chemical libraries. Acc. Chem. Res. **54**(17), 3491–3503 (2021)
56. Gao, X.J., et al.: Programmable protein circuits in living cells. Science. **361**(6408), 1252–1258 (2018)
57. Voigt, C.A.: Synthetic biology. ACS. Synth. Biol. **1**(1), 1–2 (2012). https://doi.org/10.1021/sb300001c

Chapter 2
iGEM: The Competition on Synthetic Biology

Yi Zhan, Kang Ning, and Dan Zhao

Abstract The international Genetically Engineered Machine (iGEM) competition is one of the most popular competitions in the area of synthetic biology, with a focus on students' involvement in the field of synthetic biology. The competition is now organized by the iGEM foundation, a nonprofitable organization that is independent of universities. iGEM is not only a competition but also a resource: one side of such resource is about education for students, and another side is about the fact that the plasmids from iGEM teams would normally be deposited to registry ("Registry of Standard Biological Parts") for others' utilization. These efforts would be very beneficial toward a standardized platform for either top-down or bottom-up synthetic biology simply because these are the building blocks of more complex synthetic biology systems. These plasmids have been stored in the iGEM foundation's Registry of Standard Biological Parts sample bank. In this chapter, we will introduce iGEM competition and its relation to synthetic biology. We will also show some typical cases of iGEM competition.

Keywords The international Genetically Engineered Machine (iGEM) · Synthetic biology

The international Genetically Engineered Machine (iGEM) competition is one of the most popular competitions in the area of synthetic biology, with a focus on students' involvement in the field of synthetic biology [1].

Y. Zhan · K. Ning · D. Zhao (✉)
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: zhaodann@hust.edu.cn

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8_2

2.1 iGEM's History and Current Status

The international Genetically Engineered Machine (iGEM) is a synthetic biology competition mainly for undergraduate. It started in January 2003 and has received substantial attention since then. Every year, iGEM has been held around autumn in Boston since it started, and in 2014, more than 200 teams joined this event. The competition is now organized by the iGEM foundation, a nonprofitable organization that is independent of universities.

iGEM is not only a competition but also a resource: one side of such resource is about education for students, and the other side is about the fact that the plasmids from iGEM teams would normally be deposited to registry ("Registry of Standard Biological Parts") for others' utilization. These efforts would be very beneficial toward a standardized platform for either top-down or bottom-up synthetic biology simply because these are the building blocks of more complex synthetic biology systems. These plasmids have been stored in the iGEM foundation's Registry of Standard Biological Parts sample bank.

iGEM represents one of the symbols for the synthetic biology research area, especially in the minds of many students around the world, with an increasing number of teams every year. Therefore, iGEM would serve well as a testbed for competition and, more importantly, collaborations in the area of synthetic biology.

2.2 iGEM and Synthetic Biology Research

iGEM was born at the peak of the last frenzy of synthetic biology. Thus, it also bears the hope of many synthetic biologists. During its more than 10 years of running, it has produced some novel ideas for synthetic biology research, as well as many potential biological parts that have been stored in the registry.

Synthetic biology has quite a long history from back to the 1960s, with a quicker pace and wider scale ever since. It originally came from genetic engineering, and with the help of next-generation sequencing, the development of GFP proteins, as well as the advancement of genetic engineering techniques, itself has advanced greatly in recent years. Currently, not only synthetic circuits but also synthetic networks and synthetic whole genomes have been designed. Moreover, the application of synthetic biology in biofuel, agriculture, and translational medicine has also advanced. Some hallmark synthetic biology works include first-generation gene circuit development in 2000; synthesis of bacterial photosynthesis circuit in 2005; synthetic circuits for biofuel production in 2008; creation of a bacterial cell with synthetic genome "Synthia"; a complete set of logic gates in *Escherichia coli* in 2011; commercial production of artemisinin using engineered yeast strain; as well as the use of CRISPR–Cas9 system for synthetic biology from 2013. In 2014, a special issue on synthetic biology was published in *Science* that summarized the cutting-edge development in the area [1].

In China, there are already national projects that support synthetic biology research, and the internal newsletters, which began in 2013, have already covered a wide area of synthetic biology frontiers. Chinese synthetic biologists have already developed several genetic circuits for algae, plant, and even mammalian cells. China's teams have joined iGEM since 2007, and in 2014, more than 1/4 of iGEM teams came from China. During its years of journey in iGEM, China won more than 10 golden and more than 20 silver and copper medals, representing national university students' interests in iGEM.

Several representative cases have also proven the applicability of incubating a synthetic biology project from iGEM on to a real research project. iGEM was founded in 2003 and is hosted by the Massachusetts Institute of Technology (MIT) every year. It is an international academic competition in the field of synthetic biology and an interdisciplinary competition involving cross-cooperation in the fields of mathematics, computers, and statistics.

iGEM competition was primarily aimed at undergraduates in school at the beginning and then gradually expanded to graduate students and high school students. The multidisciplinary iGEM team needs to use standard Biobricks to construct gene circuits, establish effective mathematical models, and realize the prediction, manipulation, and measurement of delicate and complex artificial biosystems to complete the competition. The long-term goal of iGEM competition is to use the model of academic competition: to realize the systemization and engineering of biology, promote the open source and transparent development of biological tools, and help build a project that can safely and effectively apply biotechnology system. iGEM competition expects to answer the core question in synthetic biology through a competition format—whether interchangeable standardized components can be used in living cells to build simple biological systems and manipulate them. Each team tried to use a standardized library of Biobricks, using standardized genetic engineering methods, to assemble artificial biosystems for specific purposes and manipulate and measure them [2].

The iGEM competition requires students to independently select topics and use their spare time to collaborate to complete the corresponding experimental work, which fully exercises the students' independent academic ability and teamwork ability and cultivates students' enthusiasm for science. Participating students can submit the useful results of the research to the MIT competition organizing committee for scientists around the world to share the research results of the students.

The competition is divided into three groups: graduate students, undergraduates, and high school students. According to the field of the participating teams, they are divided into multiple categories, such as medical health, measurement, environment, and software. Gold medals, silver medals, and bronze medals will be selected; a single award will be set up at the same time. The best project awards in various categories, including the best modeling award, the best new biological module award, etc. The champion, runner-up, and third runner-up will be selected from the final list of finalists from the team that won the gold medal.

2.3 iGEM's Representative Projects

2.3.1 iGEM Project by HUST-China 2021

In daily life, many people are keen to perm or dye their hair, but traditional chemical reagents will cause adverse effects on human health [3]. Therefore, the HUST-China team decided to use natural pigments and short peptides, which are safe and harmless materials, to perm and dye their hair. In addition, they design a fermentation bottle and a dyeing comb, providing users with a convenient way to produce the pigments and dye their hair entirely on their own.

This project "Mr.Tony" goal is to utilize synthetic biology methods to generate natural pigments and short peptides and use them in the hairdressing industry, reducing the harm caused by chemical reagents. In our project, they chose *Pichia pastoris* as our chassis organism and produced three natural pigments: indigo, curcumin, and lycopene. In addition, they produce several peptides to perm the hair. In addition to the health benefits of this method of perming and dyeing hair, there is another advantage, that is, it can quickly recover. In the future, they hope that this product can be used by more people and has broader application prospects.

2.3.2 iGEM Project by HUST-China 2019

Bananas are grown in large numbers all over the world, resulting in an annual output of 8.8 billion tons of banana stalks, but most banana stalks can only be used as fertilizer, which is a waste of resources. Therefore, this team decided to use biological fermentation to extract fine fiber from banana stalks, maximize its use, and ensure that the method is environmentally friendly and harmless. The goal of this project is to design engineered bacteria that can degrade lignin and pectin in banana straws and extract cellulose from them.

To make better use of banana stalks, HUST-China designed an engineered yeast this year, which contains two systems: a degradation system and a pH-responding system. There are three enzymes in the degradation system that can work together to degrade lignin and pectin. The pH response system can regulate pH so that the promoters can be better expressed. Through the fermentation of their designed yeasts under certain conditions, cellulose can be extracted from banana straw, which effectively solves the problem of excess banana straw resources.

2.3.3 iGEM Project by HUST-China 2018

As one of the important members of new energy, photovoltaic power generation has many advantages, but it also has problems such as high pollution and high energy

consumption. To convert optical energy into electric energy in a clean and sustainable way, this team has conducted research on microbial fuel cells and designed a photovoltaic system called Optopia through synthetic biology methods. They selected and modified the strains in the MFC and constructed an optimized version of the MFC that can generate electricity more efficiently.

2.3.4 iGEM Project by XMU-China 2020

Tea is deeply rooted in Chinese culture. For a long period, a large amount of glyphosate has been used as an herbicide, which raises a severe problem of pesticide residues in tea food. XMU-China 2020 decided to engineer strains that can rapidly detect and effectively degrade glyphosate. This year, XMU-China 2020 aimed to develop an efficient glyphosate detection and degradation system by using synthetic biology technology. It is hoped that this project could provide new ideas for the detection and degradation of pesticide residues.

2.4 iGEM Needs a Great Leap Forward

iGEM is the abbreviation of International Genetically Engineered Machine, whose major aim is undergraduate synthetic biology competition [4]. It started in January 2003 and has received substantial attention since then. Every year, iGEM is held in autumn in Boston, and in 2019, more than 300 teams joined this event. The competition is now organized by the iGEM foundation, a nonprofitable organization that is independent of universities. iGEM is not only a competition but also a resource: the plasmids from iGEM teams would usually be deposited to a registry ("Registry of Standard Biological Parts") for others' utilization. Such a registry effort would be very beneficial toward a standardized platform for either top-down or bottom-up synthetic biology simply because these are the building blocks of more complex synthetic biology systems. With the increasing number of teams every year, as well as the ever-increasing size of the iGEM registry, iGEM represents one of the symbols for the synthetic biology research area, especially in the minds of many students around the world. Therefore, iGEM is an excellent competition and testbed for collaboration in the area of synthetic biology.

Currently, synthetic biology research is at an exciting stage where more parts and more easily accessible platforms are available for all interested in synthetic biology. Many synthetic biology platforms, such as BIOFAB [5] and Synberc, have been set up, and their popularity has been increasing steadily. Moreover, the data-driven approach for synthetic system modeling has been developed rapidly, largely due to the advancement of biocuration and machine learning. Therefore, as a leading competition in synthetic biology, iGEM truly needs a leap forward to keep up the pace. The following are some strategies that iGEM might need to take to cope with the rapid development in this field:

An international host scheme should be established: As increasingly more iGEM terms are from developing countries, such as China and those from South America, sending students to Boston has been an increasing burden for students in those countries, with the average cost per team reaching approximately 20,000–50,000 dollars per year for travel and stay during the competition [6]. Letting some of these countries host iGEM once in a few years would make it financially feasible for more students from diverse backgrounds to take part in the competition. Thus, the basic design and advanced technologies in the synthetic biology field can be easily accessible to more students worldwide. It will also help to establish valuable meet-ups and possible collaborations among students through on-site discussions and possible exchanges.

For easy and confidential access to the registry of parts, interconnected sample repositories and authorized laboratories should be established and associated with each other. The current centralized registry would undoubtedly bear an increasingly heavier service burden for sample distribution, and the distributed registry scheme would largely solve this problem. Establishment of such a scheme would be technically difficult as registry mirrors are different from electronic database mirrors, which could control data distribution through a data-encryption strategy, but it needs to consider real plasmids and possibly strain storage and distribution. However, such registry mirrors would undoubtedly improve the effective realization of synthetic biology projects. Additionally, with the current large investment in biological research in developing countries such as China and Brazil, it is not unrealistic to set it up outside Europe and the United States. Furthermore, with the development of "Internet of Things" and blockchains that could support smart organization and synchronization of parts in various repositories, this should also be technically feasible. Moreover, recent advancements in microbiome research have revealed millions of novel genes from microbial communities, which could potentially serve as more diverse functional parts or modules for synthetic biology. Thus, functional element mining and adaptation from microbiome datasets would become more popular, and the accessibility of these new resources would also be a plus for iGEM.

The industrial sector should be incorporated into the iGEM's operation: It has been common practice for other synthetic biology platforms (such as BIOFAB and Synberc) to invite industrial scientists and researchers to be involved in their programs. The involvement of engineers from industry could bring in fresh ideas and different perspectives on synthetic biology, which provides a much-needed connection to the production of commercially viable products. For example, Synberc has a special program for industrial applications, and the service at BIOFAB is available to commercial users. These programs are set up to put the ideas and techniques of synthetic biology into real applications. To date, iGEM is mainly dedicated to competition and education, and the financial support from the biotech industry could also serve as a good boost for their sustainability as well. More importantly, this could help to bridge the gap of ideas to real applications for some promising iGEM teams through longer development periods [7].

More judges from industry, art, and science sectors should be included. Currently, iGEM judges' backgrounds are becoming more diverse, but not enough [6]. Judges from more diverse areas would definitely provide more options for iGEM's long-term strategic development for those that we can foresee (such as biodiversity, biosafety, etc.) and those we cannot foresee as of now. Different backgrounds of judges would also make iGEM competition more lively by means of keeping up to date in the broader area of synthetic biology.

Although obstacles may exist for a long time, we can still see a bright future ahead for iGEM, largely because of the novel (sometimes wild) ideas that could come from the students who have learned about the cutting-edge techniques. We anticipate them to have more proposals and tests on some frontier topics. First, synthetic circuits could be of real application value, such as for biofuel or other light-harvesting modules and valuable small molecules for biomedical use, etc. Second, mammalian synthetic biology, which has become accessible, is waiting for students to propose novel and easier ideas. Third, "synthetic communities," namely, the synthesis of whole communities for biomedical or industrial utilization, would also be of great importance for the development of iGEM, yet with a broader definition of "synthesis" [8]. Fourth, DNA computing and DNA storage have also emerged as potentially important areas for synthetic biology and iGEM competition to explore, among others. Finally, with the rapid development of machine learning and deep learning techniques, the weights of computational modeling techniques and affiliated databases will become increasingly important [9].

Overall, based on students unified in the iGEM competition, who can provide numerous novel ideas, together with iGEM's well-organized platform including parts and to-be-developed easily accessible platform, many creative synthetic biology projects could be carried out. To make the best use of these ideas and projects, iGEM's evolving strategies would be critical, both for attracting more students and realizing more promising projects, in the years ahead. In this sense, we believe iGEM will have a bright future in the years to come.

2.5 iGEM: Bright Futures Toward Better Synthetic Biology

Although obstacles may exist for a long time, we can still see a bright light ahead for iGEM, largely because of the novel (sometimes wild) ideas that could come from the common students who have learned about the cutting-edge techniques to have more proposals and tests on these frontier topics: First, synthetic circuits that could be of real application value, such as for biofuel or other light-harvesting modules [10], semiauto objects such as concretes or coral reefs, valuable small molecules for biomedical use, etc. Second, mammalian synthetic biology [11], which has readily become accessible, is waiting for students to propose novel and easier ideas. Third, "synthetic communities" [8], namely, the synthesis of whole communities for biomedical or industrial utilization, would also be of great importance for the development of iGEM, yet with a broader definition of "synthesis." Fourth, DNA

computing and DNA storage have also emerged as great areas for synthetic biology and iGEM competition to explore, among others [12, 13].

As more gene-editing techniques emerge and mature, it is anticipated that more synthetic biology projects will emerge, which can solve more emergent problems in biology as well as in our society.

References

1. Zhao, X., et al.: Development of international genetically engineered machine competition in China. Sheng Wu Gong Cheng Xue Bao. **34**(12), 1915–1922 (2018)
2. Hillson, N., et al.: Building a global alliance of biofoundries. Nat. Commun. **10**(1), 2040 (2019)
3. Sun, Y., et al.: Bioinspired polymeric pigments to mimic natural hair coloring. RSC Adv. **11**(3), 1694–1699 (2021)
4. Cameron, D.E., Bashor, C.J., Collins, J.J.: A brief history of synthetic biology. Nat. Rev. Microbiol. **12**(5), 381–390 (2014)
5. Saviranta, P., et al.: In vitro enzymatic biotinylation of recombinant fab fragments through a peptide acceptor tail. Bioconjug. Chem. **9**(6), 725–735 (1998)
6. Vilanova, C., Porcar, M.: iGEM 2.0—refoundations for engineering biology. Nat. Biotechnol. **32**(5), 420–424 (2014)
7. Kunjapur, A.M., Tarasova, Y., Prather, K.L.: Synthesis and accumulation of aromatic aldehydes in an engineered strain of Escherichia coli. J. Am. Chem. Soc. **136**(33), 11644–11654 (2014)
8. Grosskopf, T., Soyer, O.S.: Synthetic microbial communities. Curr. Opin. Microbiol. **18**, 72–77 (2014)
9. Sharafeldin, I.M., et al.: Computational modeling for biomimetic sensors. Methods Mol. Biol. **2027**, 195–210 (2019)
10. Zhang, Y., et al.: Development of synthetic DNA circuit and networks for molecular information processing. Nanomaterials (Basel). **11**, 11 (2021)
11. Katayama, K., Mitsunobu, H., Nishida, K.: Mammalian synthetic biology by CRISPRs engineering and applications. Curr. Opin. Chem. Biol. **52**, 79–84 (2019)
12. Emanuelson, C., Bardhan, A., Deiters, A.: DNA computing: NOT logic gates see the light. ACS Synth. Biol. **10**(7), 1682–1689 (2021)
13. Xu, C., et al.: Uncertainties in synthetic DNA-based data storage. Nucleic Acids Res. **49**(10), 5451–5469 (2021)

Chapter 3
Synthetic Biology-Related Multiomics Data Integration and Data Mining Techniques

Kang Ning and Yuxue Li

Abstract The success of a synthetic biology project is heavily dependent on omics studies, while multiomics data integration and data mining techniques have served as the foundation for rational synthetic biology work. Multiomics combines multiple types of omics data, including genomics, transcriptomics, proteomics, epigenomics, and microbiomics. It is foreseeable that multiomics research will be widely used in many biological problems to reveal more profound omics patterns. In this chapter, we have described and explained the basics of multiomics data integration and data mining techniques, which could be helpful for conducting synthetic biology studies, especially those that focus on the utilization of microbes.

Keywords Synthetic biology · Multiomics · Data integration · Data mining techniques

The success of a synthetic biology project is heavily dependent on omics studies, while multiomics data integration and data mining techniques have served as the foundation for rational synthetic biology work [1].

In this chapter, we will describe and explain the basics of multiomics data integration and data mining techniques, which could be helpful for conducting synthetic biology studies, especially those that focus on the utilization of microbes.

3.1 Introduction to Multiomics

Omics studies typically include genomics, transcriptomics, proteomics, microbiomics, and metabolomics, resulting in massive datasets [2]. With the accumulation of data, higher requirements are placed on technologies and informatics

K. Ning · Y. Li (✉)
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: liyuxue@hust.edu.cn

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8_3

tools to generate and process large biological datasets (omics data) to better handle biomedical problems [3]. In recent years, the rapid development of omics research has solved many biological problems, but to understand and solve complex biological problems more deeply, multiomics research is considered to be one of the most promising research areas [4].

Multiomics combines multiple types of omics data, including genomics, transcriptomics, proteomics, epigenomics, and microbiomics [5]. Since multiomics contains rich digital genetic materials for research objectives, it can provide multifaceted information for the target of interest, which can overcome the limitations of single-omics research to a certain extent. Specifically, genomics obtains omics data from DNA materials and is generally used to elucidate the structure of the genome, the relationship between structure and function, and the interactions between genes [3]. Transcriptomics mainly researches gene expression from RNA materials [6]. Proteomics refers to omics data obtained from protein materials to study the protein composition and changes in cells, tissues, or organisms [7]. The above three omics are prevalent research directions in the field of biology, and they are also the basis for solving many biological problems. In addition, epigenomics, phenomics, metabolomics, and microbiomics [8] are also important parts of multiomics research, and even data from bioimaging, biosensors, and social networks can be used as omics [7–9]. In general, these omics all represent one aspect of the research goal, and they consistently reveal the regulatory patterns and principles of the process of how genetic material regulates genotypes. However, there is only one type of data analysis that has limited relevance, mainly reflecting the reaction process rather than causality. The integration of different omics data types is usually used to clarify the potential pathogenic changes or treatment targets that cause the disease, which can then be further tested. Based on different research goals, combining multiple types of omics data can explain biological phenomena more comprehensively, and multiomics studies can better understand the basic information flow of diseases than single-type omics studies. Due to the importance of multiomics in basic research and clinical applications, multiomics has become a research focus in recent years [9]. For multiomics data, a range of disease-related differences are often generated that can serve as markers of the disease process, as well as insights into biological pathway or process differences between disease and control groups.

3.2 Timeline of Omics Development

The development of DNA-related knowledge has promoted the development of omics to a certain extent. DNA was first isolated in 1869, and two years later, Friedrich Miescher identified the presence of nuclear and related proteins in the nucleus, which is what we now call DNA, and formed the basis of the field of genomics. In 1904, the theory of chromosomal inheritance, in which chromosomes appear in pairs inherited from the mother and father, respectively, was proposed. In 1910, Albert Kossel discovered five bases: adenine (A), cytosine (C), guanine (G),

thymine (T), and uracil (U). After 40 years, Erwin Chargaff found that thymine and adenine or cytosine and guanine are always equal in DNA samples. Therefore, he reasoned that adenosine and thymine form a pair of chromosomes, and cytosine and guanine form a pair of chromosomes. This discovery advances the understanding of the base pairing of adenosine, cytosine, guanine, and thymine nucleotides. In 1952, Alfred Hershey and Martha Chase proved through a series of experiments that DNA, not protein, is responsible for carrying genetic features that may be inherited. The following year, the double helix structure of DNA was discovered by James Watson and Francis Crick [10].

Since DNA was isolated, DNA-related knowledge has continued to develop and improve, but Sanger invented the first sequencing technique in 1958 [11], making genomics a relatively new subject. The principle of Sanger sequencing is to use a DNA polymerase to extend the primers bound to the template of the undetermined sequence until a chain terminating nucleotide is incorporated. In 1977, Frederick Sanger developed a DNA sequencing technology to sequence the first complete genome, called the phiX174 virus, which opened the door to possibilities in the field of genomics. Sanger's enzymatic sequencing technology is the basis of today's large-scale genome sequencing.

In 1983, Dr. Kary Mullis developed a technique for amplifying DNA called polymerase chain reaction (PCR) [12]. In 1990, a significant scientific project, the Human Genome Project, was launched to sequence all 3 billion letters of the human genome. The project, completed in 2003, confirmed that humans have 20,000–25,000 genes. During this period, the first bacterial genome sequence, *Haemophilus influenzae* [13], and the yeast genome were completed [14]. In 2007, there was a breakthrough in the technology used for DNA sequencing, which increased the output of DNA sequencing by 70 times in 1 year. This led to the launch of the 1000 Gene Project in 2008, which aims to sequence the genome of a huge population of 2500 people.

In general, the progress of DNA technology has promoted the development of omics, and the deepening of omics research in turn provides new opportunities for the development of DNA technology.

3.3 Databases and Tools for Omics Studies

With the development of sequencing technology, a large amount of multiomics data has been generated on a global scale. The growth rate of multivariate data was unimaginable a decade ago, and large public databases already use cloud facilities to store these data. Moreover, the cost of generating multiple sets of data dropped rapidly, leading to a further increase in the amount of multiomics data [20]. In the face of multiomics data, computational methods for the reasonable integration and accurate analysis of heterogeneous multiomics data are increasingly demanding, and auxiliary means such as databases and analysis tools are also crucial. Tables 3.1 and 3.2 show representative databases and analytical tools.

Table 3.1 Representative databases for multiomics researches

Database	Functionality	Web link	Reference
ChEBI	Metabolomics database and ontology	http://bigd.big.ac.cn/databasecommons/database/id/364	[15]
GenBank (database)	Proteomics database open access annotated collection of all publicly available nucleotide sequences and their protein transitions	https://www.uniprot.org/database/DB-0028	[16]
Human Metabolome Database (HMDB)	Human metabolite and pathway database	https://hmdb.ca/	[17]
KEGG	Collection of databases dealing with genomes biological pathways, disease, drugs, and chemical substances	https://www.kegg.jp/	[18]

Table 3.2 Representative analytical tools for multiomics researches

Tool/method	Tool/method approach	Tool/method link	Reference
PARADIGM	Probabilistic graphical models using directed factor graphs	http://paradigm.five3genomics.com/	[19]
iCluster	Joint latent variable model-based clustering method	https://cran.r-project.org/web/packages/iCluster/index.html	[20]
iClusterPlus	Generalized linear regression for the formulation of the joint model	http://www.bioconductor.org/packages/release/bioc/html/iClusterPlus.html	[21]
LRAcluster	Probabilistic The model with low-rank approximation	http://lifeome.net/software/lracluster/	[22]

3.4 Multiomics Applications

Multiomics research is increasingly widely used in the field of biology, and many achievements have been made in cancer, inflammatory diseases, and microbes. In 2018, Hasin et al. [23] proposed the clinical cancer genome atlas with triple platform sequencing of the whole genome, whole exome, and transcriptome. They used genome, exome, and transcriptome sequencing to investigate the genetic structure of 78 clinical tumor samples and identified tumor-associated structural variants (SVs), somatic mutations, and pathogenic mutations to evaluate the potential of a comprehensive clinical trial to detect different types of somatic and germline mutations associated with childhood oncology. The results of the study show that the sensitivity of this method is greatly improved compared to the previous combination of WES and RNA-Seq to detect pathogenic variants. This work highlights the need to include WGS in pediatric cancer testing and illustrates the importance of multiomics studies in disease detection.

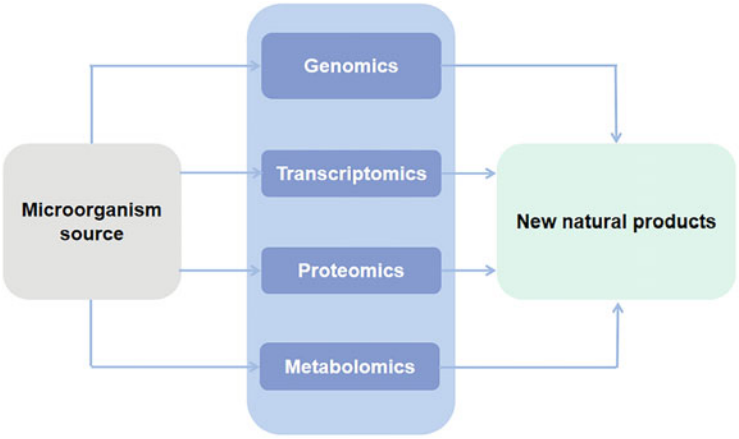

Fig. 3.1 Omics workflow for discovering natural products. Based on microbe sources, genomics, transcriptomics, proteomics, and metabolomics have been combined for the discovery of natural products from microbes

Multiomics approaches have also been applied to study the biosynthesis of microbial secondary metabolites [24]. Combining genomics, transcriptomics, proteomics, and metabolomics to study the development of natural metabolites. Figure 3.1 shows the omics workflow for discovering natural products.

Multiomics is widely used in disease, and a review published in 2017 outlined the idea of multiomics research in disease [25]. This review provides an overview of the genome, epigenome, transcriptome, proteome, metabolome, and microbiome, as well as the interrelationships between omics, with an emphasis on the methods for their integration across multiple omics layers. In general, compared to studies of a single omics type, multiomics offers an opportunity to understand the flow of information behind disease.

Multiomics also has many interesting applications, such as the NASA twin study, a multidimensional analysis of a year-long human spaceflight [26]. To study the changes that take place in humans' bodies during long-term survival in space, when NASA astronaut Scott Kelly embarked on a year-long mission to the international space station, researchers collected genomic, molecular, and physiological data from both him and his twin brother Mark Kelly (former astronauts) and compared them through a multiomics longitudinal analysis process (Fig. 3.2). This study is mainly divided into four parts: (1) human physiology: changes in the heart, muscles, brain, and other organs in space and their causes. (2) Behavioral health: cognitive reasoning, decision-making, and alertness in the space environment. (3) Microbiology: dietary differences and stress differences and how they affect the gut microbiome. (4) Molecular/omics: effects of space environment, radiation, claustrophobic conditions, and space environment on gene expression. Those changes in the data may reveal how the human body responds to extreme environments.

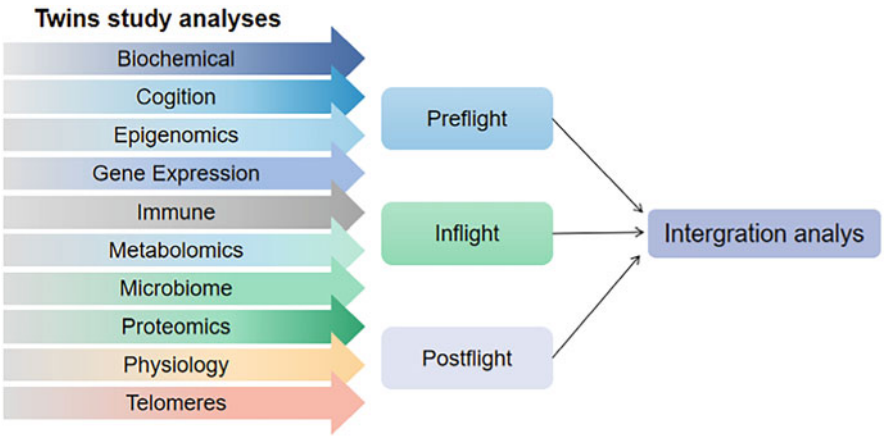

Fig. 3.2 Multidimensional, longitudinal assays of the NASA Twins Study. Characteristics of 10 generalized biomedical models were collected preflight, in-flight, and postflight over a 25-month period

3.5 Future Perspectives

Currently, multiomics data are rapidly accumulating and more accessible, and related databases and analysis tools are becoming increasingly mature. It is foreseeable that multiomics research will be widely used in many biological problems to reveal more profound omics patterns. For example, benefitting from the importance of multiomics analysis in causal inference, mediation analysis, and risk prediction, epidemiological studies of chronic diseases have achieved many results [27]. A longitudinal multiomic study of host–microbe dynamics in prediabetes provides greater insight into the multiomic signatures of its early state [28]. Similarly, in another study, deep multiomics measures were used to identify clinically relevant T2D molecular pathways, exploring the ability of deep longitudinal analysis in health-related findings to identify clinically relevant molecular pathways and provide relevant information for accurate health information [29]. In pharmaceutical research, epidemiological, pharmacological, genetic, and gut microbiology data are integrated into drug metabolite maps. Applying the research map to targeted experimental drug research and clinical trials improves the effectiveness and safety of the drug [30]. In general, studies at the single omics level lack multilevel integration and have limited value in inferring the etiology of complex diseases. Multiomics studies have greatly expanded the depth of etiology research. Multiomics provides new ideas for traditional observational epidemiological research to infer the cause of chronic diseases, provides valuable resources in the integration of systemic epidemiology to explore disease mechanisms, and will become an important reference for subsequent further experimental verification studies [27].

However, there are also many challenges in multiomics research. Obtaining disease-related multiomics data requires maintenance of long-term follow-up and

laboratory testing, which is expensive. Furthermore, the combined effects of multiple factors and the high variability of a single dataset can lead to erroneous findings, which makes multiomics analysis results difficult to interpret, especially the identification of biologically relevant molecules [27]. In general, although multiomics studies have advantages over many biological questions that cannot be achieved by single omics and the current development trend is very promising, there is still a long way to go to fully understand how genetic material modulates the overall and dynamic patterns of the target phenotype.

References

1. Keshava, R., et al.: Chapter 4—Synthetic biology: overview and applications. In: Barh, D., Azevedo, V. (eds.) Omics Technologies and Bio-Engineering, pp. 63–93. Academic Press (2018)
2. Osier, N.D., et al.: Symptom science: repurposing existing omics data. Biol. Res. Nurs. **19**(1), 18–27 (2017)
3. Manzoni, C., et al.: Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief. Bioinform. **19**(2), 286–302 (2018)
4. Sun, Y.V., Hu, Y.-J.: Chapter three - integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. In: Friedmann, T., Dunlap, J.C., Goodwin, S.F. (eds.) Advances in Genetics, pp. 147–190. Academic Press (2016)
5. Chung, R.-H., Kang, C.-Y.: A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. GigaScience. **8**, 5 (2019)
6. Dong, Z., Chen, Y.: Transcriptomics: advances and approaches. Sci. China Life Sci. **56**(10), 960–967 (2013)
7. Marchesi, J.R., Ravel, J.: The vocabulary of microbiome research: a proposal. Microbiome. **3**(1), 31 (2015)
8. Kumar, P.S.: Microbiomics: were we all wrong before? Periodontology.**85**(1), 8–11 (2021)
9. Chakraborty, S.A.-O.X., et al.: Onco-Multi-OMICS approach: a new frontier in cancer research. Biomed. Res. Int.**2018**, 9836256 (2018)
10. Portin, P.: The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA. J. Genet.**93**(1), 293–302 (2014)
11. Heather, J.M., Chain, B.: The sequence of sequencers: the history of sequencing DNA. Genomics. **107**(1), 1–8 (2016)
12. García-Quesada, A.A.-O., et al.: Seroprevalence and prevalence of Babesia vogeli in clinically healthy dogs and their ticks in Costa Rica. J. Genet.**93**, 293–302 (2014)
13. Fraser, C.M., Rappuoli, R.: Application of microbial genomic science to advanced therapeutics. Annu. Rev. Med. **56**(1), 459–474 (2004)
14. Zhang, M.Q.: Promoter analysis of co-regulated genes in the yeast genome. Comput. Chem. **23**(3), 233–250 (1999)
15. Degtyarenko, K., et al.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res.**36**(Database issue), D344–D350 (2008)
16. Benson, D.A., et al.: GenBank. Nucleic Acids Res. **46**(D1), D41–D47 (2018)
17. Wishart, D.S., et al.: HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. **46**(D1), D608–D617 (2017)
18. Kanehisa, M., et al.: KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. **45**(D1), D353–D361 (2016)

19. Gluth, S., Rieskamp, C., Fau-Büchel, J., Büchel, C.: Deciding not to decide: computational and neural evidence for hidden behavior in sequential choice. PLoS Comput. Biol.**9**(10), e1003309 (2013)
20. Shen, R., Olshen, M., Fau-Ladanyi, A., Ladanyi, M.: Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics.**25**(22), 2906–2912 (2009)
21. Pierre-Jean, M., et al.: Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. Brief. Bioinform.**21**(6), 2011–2030 (2020)
22. LRAcluster. http://lifeome.net/software/lracluster/
23. Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease. Genome Biol. **18**(1), 83 (2017)
24. Palazzotto, E., Weber, T.: Omics and multi-omics approaches to study the biosynthesis of secondary metabolites in microorganisms. Curr. Opin. Microbiol.**45**, 109–116 (2018)
25. Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease. Genome Biol.**18**(1), 83 (2017)
26. Garrett-Bakelman Francine, E., et al.: The NASA twins study: a multidimensional analysis of a year-long human spaceflight. Science. **364**(6436), eaau8650 (2019)
27. Pang, Y.J., et al.: A multi-omics approach to investigate the etiology of non-communicable diseases: recent advance and applications. Zhonghua Liu Xing Bing Xue Za Zhi. **42**(1), 1–9 (2021)
28. Zhou, W., et al.: Longitudinal multi-omics of host–microbe dynamics in prediabetes. Nature. **569**(7758), 663–671 (2019)
29. Schüssler-Fiorenza Rose, S.M., et al.: A longitudinal big data approach for precision health. Nat. Med. **25**(5), 792–804 (2019)
30. Liu, J., et al.: Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug–metabolite atlas. Nat. Med. **26**(1), 110–117 (2020)

Chapter 4
Synthetic Biology: Technical Issues

Bohan Wang, Xiunan Huo, Xianglei Zhang, Yuanhao Liang, Yingying Yang, Jiacheng Shi, Xinyu Huan, Xilin Hou, Weilin Lv, and Yi Zhan

Abstract As a frontier of biological research with engineering design as the main methodology, synthetic biology integrates the most cutting-edge and important technical means in current life science research to shape and construct novel organisms. It explores new scientific and technological knowledge and achieves the purpose of serving the development needs of human society at the same time. From the perspective of the technology involved, the field of synthetic biology mainly includes the following technical applications in its research process: DNA synthesis and assembly, gene editing technology, directed evolution and artificial design of living organisms, and technical details that are easily ignored: the selection and operation of the chassis, which involves all living organisms in biological research, from in vitro cell-free systems to living organisms such as microorganisms, animals, and plants. They are all the targets that synthetic biology attempts to transform. This chapter will mainly discuss the details and latest developments to offer readers a preliminary understanding of synthetic biology technologies.

Keywords Synthetic biology technologies · DNA synthesis and assembly · Gene editing technology · Directed evolution · Artificial design

4.1 Introduction and Background

Synthetic biology has developed for more than 100 years as a genetic engineering research discipline based on systems biology and molecular biology. Since Watson and Crick discovered the double helix structure of DNA in 1953 and started the era of molecular genetics, synthetic biology has entered a new stage of rapid development. In 2000, at the annual meeting of the American Chemical Society, synthetic

B. Wang · X. Huo · X. Zhang · Y. Liang · Y. Yang · J. Shi · X. Huan · X. Hou · W. Lv · Y. Zhan (✉)
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: zhanyi@hust.edu.cn

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8_4

biology was again introduced to describe the synthesis of unnatural organic molecules that function in living systems [1].

As a frontier research of biology with engineering design as the main methodology, synthetic biology integrates the most cutting-edge and important technical means in current life science research to shape and construct novel organisms. It explores new scientific and technological knowledge and achieves the purpose of serving the development needs of human society at the same time. From the perspective of the technology involved, the field of synthetic biology mainly includes the following technical applications in its research process: DNA synthesis and assembly, gene editing technology, directed evolution and artificial design of living organisms, and technical details that are easily ignored: the selection and operation of the chassis, which involves all living organisms in biological research, from in vitro cell-free systems to living organisms such as microorganisms, animals, and plants. They are all the targets that synthetic biology attempts to transform. This chapter will mainly discuss the details and latest developments to offer readers a preliminary understanding of synthetic biology technologies.

4.2 DNA Synthesis and Assembly

4.2.1 DNA Synthesis

DNA synthesis technology is the foundation of synthetic biology and modern molecular biology. Only natural DNA fragments can be obtained by polymerase chain reaction (PCR) and restriction endonuclease digestion, while artificially designed DNA fragments can be obtained by DNA synthesis from scratch. At present, the main methods of DNA synthesis can be classified as oligonucleotide column synthesis based on chemical synthesis, chip DNA synthesis, and DNA synthesis based on enzymatic reactions.

The column synthesis of oligonucleotides began in the 1950s using the phosphodiester method [2]. In the 1980s [3], Beaucage and Caruthers developed the DNA synthesis method based on phosphoramidite, which is also the main method used in automatic production currently. The synthesis of phosphoramidite consists of a cycle of four chemical reactions: deprotection, coupling, capping, and oxidation. The oligonucleotide chain is synthesized one by one on the solid phase carrier by activating the chemically active protective groups bound to the 3′ and 5′ positions of nucleotides step by step [4].

Column DNA synthesis technology can easily extract any oligonucleotide fragments needed for the synthesis of a certain gene, which can meet the requirements of general experiments. However, with the decrease in chemical reaction efficiency, purity, and yield caused by chain extension, the length synthesized by this method is generally no more than 200 nucleotides. Column synthesis is gradually unable to meet the needs of large-scale gene and genome synthesis in the field of synthetic biology because of its weak ability of sustainable synthesis, high consumption of

chemical reagents, many side reactions, and low flux [5]. The current research is mainly focused on high-throughput synthesis technology based on chip carriers and enzymatic oligonucleotide synthesis technology based on template-free single-stranded DNA synthesis.

Chip DNA synthesis uses the chip as a solid phase carrier to synthesize at specific sites on its surface in a high-density and integrated manner. The synthesis principle is the same as that of the phosphoramidite method, but different "deprotection" fixed-point control methods are adopted, which can be divided into three types: lithography synthesis, electrochemical synthesis, and ink-jet printing synthesis [6]. However, due to the unique heterogeneity and edge effect of the chip, chip synthesis has decreased in length and accuracy [7].

In 2010, Kosuri [8] achieved a larger scale and higher precision synthesis through a multiple PCR strategy combined with enzymatic error correction to make the synthesis length more than 200 nt. Matza [9] selected the correct oligonucleotide products by sequencing and then amplified a large number of these products to reduce the error rate of the sample.

Enzymatic DNA synthesis has attracted increasing attention because of its advantages in the accuracy of synthesis and the need for no use of toxic compounds. There are two important problems in enzymatic synthesis. First, conventional DNA polymerase depends on the template and cannot be used in the ab initio synthesis of DNA. Second, how to make the enzyme add specified nucleotides one by one under controlled conditions.

In the 1960s, Bollum first discovered terminal deoxynucleotidyl transferase [10] (TdT) and proposed in 1962 [11] that TdT can be used in the synthesis of single-stranded oligonucleotides. TdT has little preference for four nucleotides and high coupling efficiency. Continuous synthesis and extension of single-stranded DNA can produce homopolymers as long as 8000 nt [10, 12–14]. An effective reversible termination method is needed for TdT to be used in controllable enzyme-catalyzed DNA synthesis.

In 2018, Keasling's team connected a single nucleotide with a cleavable connector on a single-molecule TdT. After adding nucleotides to the primer chain using TdT, it remained connected to the DNA chain, effectively preventing the further extension of the DNA chain, and after releasing TdT from the disconnect, DNA can carry out a new round of nucleotide addition cycles.

4.2.2 Biobricks and Standardization

From a single-cell organism, such as *Escherichia coli*, to a blue whale or the human brain, nature uses the same set of building blocks of (deoxy-)ribonucleic acids, amino acids, sugars, and lipids. To a certain extent, biological compatibility can play a role at the genetic level by inserting genes from one organism into another.

To realize the free assembly and expression of genes, it is necessary to establish a standard. Standardized components have long been widely used in mechanical,

electronic, computer engineering, and other industrial fields. Due to the application of standardized components, components with different functions and products of different companies can be easily integrated, thus enabling the industry to produce complex and reliable products [15].

Similarly, synthetic biologists proposed the concept of a standardized biological module, the biobrick. Biobricks can be small pieces of DNA with a certain function, called parts, such as promoters and RBSs; can be slightly larger gene regulatory circuits composed of several parts, called devices; or even systems formed by the cooperative operation of devices with different functions.

The use of standardized biobricks allows standard parts constructed by different laboratories to be assembled according to the same rules, thus avoiding much repetitive work and shortening the time required to synthesize complex devices or systems.

Biobrick standardization uses a standard assembly method or layered assembly to add a unified interface to both ends of the biobrick [16]. The standard assembly method mainly uses restriction endonuclease for restriction endonuclease digestion and ligation: each biobrick contains two specified restriction sites at both ends, and the viscous ends produced by these isocaudamers can connect two different biobricks. Layered assembly mainly uses gateway technology to produce assembly carriers with biological elements and then uses ab technology to assemble two assembly carriers with different biological elements.

The standardization of synthetic biology is reflected not only in the assembly method but also in the quantitative mechanism. The iGEM Registry provides standards for measuring and representing input and output signals: PoPS and RIPS. PoPS (RNA polymerase per second) means the number of RNA polymerase molecules passing through a certain point on the DNA molecule per second, which is used to measure the transcriptional level of a gene [17]. RIPS (ribosomal initiations per second) refers to the number of ribosomal molecules passing through a certain point on the mRNA molecule per second, which is used to measure the translation level of mRNA.

Industrial biotechnology and synthetic biology have combined to enable the manufacture of a wide variety of bulk and fine compounds from renewable resources. Nevertheless, the creation of the necessary microbial cell factories is still a protracted, labor-intensive project with an unknown end. A fundamental barrier to predicted strain engineering in this regard is the absence of dependable, defined, and standardized biological components.

4.2.3 Homologous Recombination

Efficient gene assembly and editing techniques have great significance for basic biological research and biotechnology. Homologous recombination is the most widely used among different kinds of recombination. DNA recombination involves the exchange, rearrangement, and transfer of dNTPs within DNA or between DNA

molecules and alters the arrangement of existing genetic materials. Homologous recombination is a form of direct exchange that occurs between two homologous sequences. Through homology, organisms can not only produce new genes or allelic combinations but also improve the diversity of genetic material. Today, we use homologous recombination for genetic mapping, gene diagnosis, and gene editing. This section will introduce the basic knowledge and principles of homologous recombination and how these principles are applied to synthetic biology as an important part of modern molecular biology.

Different Models of Homologous Recombination

In 1964, American molecular biologist Robin Holliday proposed the Holliday model. It contains three key steps: strand invasion, the process of branch migration, and Holliday junction resolution. The main contents of the model are as follows [18].

First, the two homologous chromosomes are close to each other, and chromatid synapsis, the two single strands in the same direction of the two DNA molecules, were cut at the same position under the action of DNA endonuclease. The double helix is slightly loosened at each cut. Then, the released single strand is complementary to the strand on another chromatid to form a Holliday junction. Finally, the Holliday junction is separated in two different ways to form recombinant or nonrecombined products.

Meselson and Radding proposed the Meselson–Radding model [19] to improve the Holliday model. They proposed that breaks occur in only one chain. After the repair and synthesis of the missing DNA strand, the free end of the single strand invades into the adjacent DNA double strand and binds with the homologous region. Then, the replaced DNA strand forms a D-loop, a structure that keeps growing with the process of recombination. The end of the single strand will cross into the adjacent DNA double strand when the D-loop is released. The free end of DNA will form the Holliday structure by covalent connection.

Further research found that vector DNA double-strand breaks could promote homologous recombination, but both Holliday and Meselson–Radding models could not interpret this situation. Therefore, Szostak et al. proposed a double-strand break repair model (DSBR) [20]. Recombination occurs in a specific region called the DNA double-strand break region. Because the repair of this region uses the corresponding homologous DNA strand as the template, the repaired fracture region essentially becomes the same as its homologous sequence. Then, the Holliday junction will separate and form either a crossover product or non-crossover product.

Application of Homologous Recombination: Gene Knockout

In the 1980s, based on the principle of homologous recombination, Capecchi and Smithies realized the site-directed integration of foreign genes of ES (embryonic stem) for the first time [21]. This technique is called “gene targeting” or “gene

knockout." Gene knockout mice also benefit from the development of microinjection technology for this kind of ES. For this work, Capecchi and Smithies shared the Nobel Prize in Medicine with Evans in 2007. To achieve this goal, they designed a recombinant vector with similar sequences at both ends of the target gene. The recombinant vector that is introduced into the ES will cause homologous recombination between homologous sequences. In this way, the target gene can be selectively knocked out or replaced. Gene knockout is an effective method to study the basic function of a gene and understand the influence of different genes on embryonic development. However, there is a complex relationship between the structure and function of the gene, and the mutation of a single nucleotide may lead to a change in function. Subtle mutation [22, 23] of genes developed on this basis makes it possible to study the relationship between fine structure and function of genes in smaller units.

4.2.4 Golden Gate and Gibson Assembly

The classical cloning method has some disadvantages, such as complicated procedures and retaining binding sites of restriction enzymes. Golden Gate and Gibson Assembly avoid these disadvantages well and have become popular cloning technologies with their own advantages.

Golden Gate

Golden Gate is a cloning method that relies on IIS restriction enzymes [24, 25], also known as IIS cloning. Unlike commonly used type 2 restriction endonuclease enzymes, type IIS restriction enzymes (e.g., BsaI, BbSI, BSMBI) cleave DNA to recognize nonpalindromic sequences and cleave outside the recognition sequence to produce a sticky end. Therefore, this method also allows multiple fragments to react at the same time and improves the accuracy. The IIS restriction enzyme recognition site was designed outside the target gene cleavage site. Enzyme digestion and ligation were carried out in the same reaction system. PCR products could be used as substrates, or PCR products could be cloned into plasmid vectors and then plasmids could be used as substrates. After ligation, the recognition site was eliminated and did not appear in the final vector. The target segments containing the same protruding ends are connected to each other or to the carrier. The vector contains protruding ends complementary to the cleavage site of the target gene and can also direct ligation. Under suitable experimental conditions and design, more than 20 DNA fragments can be assembled with Golden Gate Assembly.

Gibson Assembly

Gibson Assembly is known for its ability to easily assemble multiple linear DNA segments. Using this technique, a 583 kb artificial genome can be successfully assembled [26]. At the same time, the whole reaction system can work at the same temperature, and the reaction can be completed in 1 h or less [27]. In Gibson Assembly, the vector and DNA fragment with homologous terminals were obtained by PCR. Then, a reaction system was prepared by T5 Exonuclease, Phusion DNA Polymerase, and Taq DNA Ligase. Exonucleases that hide double-stranded DNA from the 5′ end do not compete with the activity of the polymerase. Every enzyme for DNA assembly in this method will be active simultaneously in one isothermal reaction. T5 exonuclease, which is not heat-resistant, removes the 5′ end nucleotides of double-stranded DNA molecules and complements the annealed single-stranded DNA overhangs. Phusion DNA polymerases fill gaps through DNA synthesis, producing circular DNA with only nicks. Taq DNA ligase repairs the defect by forming phosphodiester bonds and obtains complete double-stranded DNA.

4.2.5 Transformation-Associated Recombination (TAR)

Transformation-associated recombination (TAR) is a cloning technique that can quickly and accurately isolate specific chromosomal regions or genes directly from genomic DNA. The traditional method to obtain the target fragment is to first build the genome library and then select the target fragment from the gene library. However, TAR overcomes this problem. Therefore, it was widely used soon after it was put forward. In particular, yeast transformation-associated recombination can be used to selectively isolate large genomic regions from mammalian total genomic DNA to form linear or circular yeast artificial chromosomes (YACs) under certain conditions [28, 29].

TAR exploits a high level of recombination between homologous DNA sequences during transformation in the yeast *Saccharomyces cerevisiae* [30]. In addition, this method is the only way to isolate the same genome region from multiple individuals simultaneously, segments up to 300 kb in length [31]. Under optimized conditions, any desirable chromosomal fragment up to 300 kb can be isolated in yeast from multiple samples with a yield of gene/genomic fragment-positive clones as high as 32% [32].

The components of the TAR vector generally include the yeast centriole sequence (CEN), selective marker M (His3, Ura3, etc.), and two unique targeting sequences (hook1 and hook2) that are homologous to the 5′ and 3′ ends of a gene of interest. The hook sequences may either be both unique or one of the hook sequences may be a common repeat. TAR vectors containing telomere sequences (TELs) will form linear YACs after homologous recombination. In the absence of telomeres, they will form circular YACs. The vector should replicate autonomously in yeast, be

genetically stable, and contain yeast's initial replication site (ARS). There is an ARS-like sequence every 20–40 kb in the human and mouse genomes. TAR vectors usually do not contain yeast ARS sequences and can be replicated autonomously in yeast cells by capturing ARS-like sequences in genomic DNA [33].

The basic process is to introduce the linear TAR clone vector and the segmented genomic DNA into yeast cells at the same time and make the target gene fragment and TAR vector recombine by using the efficient recombinant system of yeast cells themselves. Although only a 35 bp homologous sequence is needed to ensure that recombination occurs [34], if necessary, genomic DNA can be treated by CRISPR/Cas9 endonuclease before yeast transformation to improve conversion efficiency. However, for some chromosomal regions, such as the heterosomal region, centriole region, and telomere region, the ARS-like sequence is not contained or the frequency of the ARS-like sequence is very low, so it is necessary to use a TAR vector containing the ARS sequence for cloning, but such background interference is not easy to remove, which brings great difficulties for screening and identification [35]. For such cases, a modified TAR method has been developed [36, 37]. In this system, positive clones were selected by positive and negative genetic selection and could eliminate the background caused by vector recircularization that results from end-joining during yeast transformation.

Applications of TAR cloning are wide. It can be used as a tool to selectively isolate specific chromosome segments or genes from individuals and can also be used to isolate rearranged chromosomal regions, such as translocations and inversions, from patients and model organisms. It can isolate gene homologs for evolutionary studies or chromosomal regions that are unclonable in bacterial vectors, separate gene alleles and long-range molecular haplotyping, construct HACs for gene functional studies, and so on.

4.3 Gene-Editing Technology

4.3.1 Cre-loxP and Conditional Editing

The Cre/loxP recombinant enzyme system, found in Phage P1 in 1981 [38], is a highly efficient site-specific recombination system that has been widely used in gene editing. The Cre/loxP recombinant enzyme system allows researchers to perform deletion, insertion, translocation, and inversion at specific sites of DNA, modifying DNA in cells. This system can be used in both eukaryotic and prokaryotic cells [39].

Cre recombinase is derived from a DNA sequence encoded by bacteriophage P1. It is a 38 kD protein encoded by the cre gene [40] and can recognize a specific sequence on DNA and catalyze the cleavage and recombination of DNA at this site. Cre recombinase does not need the participation of cofactors or additional energy when acting on nucleotide sequences. The Loxp site is the specific recognition site of Cre recombinase. It is a DNA sequence of 34 bp in length that consists of two reverse repeat sequences of 13 bp and an asymmetric spacer of 8 bp. Experiments show that

this sequence is a necessary and sufficient condition for Cre recombinase-mediated recombination [41].

The Cre/loxP recombinant enzyme system has many excellent properties as a tool. In the recombination process mediated by the Cre/loxP recombinase system, Cre recombinase can complete DNA recombination in vivo or in vitro without cofactors, and the recombination efficiency is much higher than that of other recombinant enzymes because the Loxp site can accurately label the target gene, and Cre recombinase cannot mediate specific recombination outside the Loxp site. On the other hand, Cre recombinase-mediated DNA recombination is not limited by the length and position of excised fragments, and DNA recombination can be carried out accurately as long as the target gene is marked by two recognition sites.

Cre/loxP recombinant enzyme systems have been widely used in the field of bioscience. In molecular biology research, the Cre/loxP recombinant enzyme system is used for site-directed deletion of genes. As long as the Loxp site is inserted at the predetermined site and the Cre recombinase coding gene is under the control of specific regulatory genes, the expression of Cre recombinase can be controlled, and a gene can be deleted in a specific time range [42]. At the same time, the Cre/loxP recombinase system can also be used for the integration of foreign genes. As long as a Loxp site is introduced into the genome, the foreign gene can be integrated into the genome by using the Cre/Loxp system linked to the foreign gene [43]. In the field of medicine, the Cre/loxP recombinant enzyme system is used to establish animal models of diseases. The principle of this application is similar to the targeted deletion of genes, creating functionally defective disease animal models by the precise deletion of specific genes in experimental animals.

4.3.2 ZFN and TALEN

Since the beginning of the twenty-first century, brand-new gene-editing technologies have emerged. In the development time sequence, ZFN (zinc finger nuclease) technology appeared first, followed by TALEN (transcription activator-like effector nuclease), and the latest is CRISPR. Every time a new technology appears, gene-editing technology becomes easier and more popular. Although the emergence of the CRISPR technology that we will introduce in the next section has greatly simplified gene-editing operations, ZFN and TALEN technologies are still considered to have significant advantages in applications such as gene therapy due to their accuracy and other factors.

Gene-editing technologies such as ZFN and TALEN are mainly composed of specific DNA recognition domains and endonucleases. The DNA recognition domain of ZFNs is composed of a series of Cys2-His2 zinc finger proteins (zinc fingers) in tandem (generally 3 to 4), and each zinc finger protein recognizes and combines a specific triplet base. The amino acid sequence of the zinc finger domain is TGEKPYKCPECGKSFSXXXXXXXHQRTH, and the X position is a variable amino acid. By changing these amino acids, the zinc finger domain can recognize

different bases. Researchers can process and modify the zinc finger DNA-binding domain of ZFN to target different DNA sequences so that ZFN can bind to the target sequence in the complex genome and specifically cut with the DNA cleavage domain. By combining with the intracellular DNA repair mechanism, the genome can be edited in vivo. Therefore, for any target sequence that needs to be identified, we can use a similar method corresponding to the codon to modularize the zinc finger structure to obtain a zinc finger protein structure that can recognize a specific DNA sequence. The most widely used DNA cleavage domain in ZFNs comes from the type IIS restriction enzyme FokI. It functions in the form of a dimer. When constructing ZFN, two ZFNs should be designed for the adjacent regions on each strand of DNA to cut the same position of the double strand. There is a 5–6 bp spacer structure between the two ZFNs [44]. ZFN cuts DNA specifically to form a DNA double-strand break region and inactivates the target gene by breaking nonhomologous end-joining or repairing DNA connections by homologous recombination to complete general genome-editing operations [45]. Researchers have used ZFN technology in the treatment of HIV, Duchenne muscular dystrophy, 21 trisomy syndrome, and other genetic disease gene therapies, producing promising applications [46].

TALEN was first discovered in the plant pathogen as an invasion strategy for bacterial infection of plants. Similar to ZFN, it also includes a DNA recognition domain and endonuclease. The DNA recognition domain of TALENs consists of a series of TAL proteins (usually 14–20). The difference in amino acids at positions 12 and 13 of the 34 aa TAL protein allows each TAL protein to recognize and bind a corresponding DNA base. The constructed pair of TAL target recognition modules are connected with the N-terminal nuclear localization sequence and the C-terminal FokI enzyme to obtain a complete TALEN element. The endonuclease also forms a dimer to cut double-stranded DNA in the 12–20 bp spacer region [47]. DNA self-repair or homologous recombination caused by base deletion or insertion finally completes gene editing. TALEN technology has no genetic sequence or cell or species restrictions but has a higher success rate and efficiency and no obvious off-target effects. Many research groups around the world have verified the specific cleavage activity of TALENs using in vitro cultured cells, yeast, Arabidopsis, rice, fruit flies, zebrafish, and other animal and plant systems and even established a genome-wide scale TALEN system [48]. Through the use of microinjection and other technologies, TALENs have been applied in stem cell research, gene therapy, neural network research, animal and plant breeding, and other fields.

4.3.3 CRISPR/Cas

CRISPRs (clustered regularly interspaced short palindromic repeats) are genetic elements, and bacteria can use CRISPRs to protect against viruses as a kind of acquired immunity [49]. CRISPRs have been integrated into the bacterial genome and consist of short sequences that originate from viral genomes [50, 51]. These

sequences are processed by the CRISPR-associated protein Cas, which then cuts identical viral DNA sequences.

In general, there are three distinct phases in CRISPR–Cas immunity: adaptation, expression, and interference. During the adaptation stage, short fragments of DNA homologous to invading virus or plasmid sequences are integrated into the CRISPR locus [52, 53]. The second stage in CRISPR–Cas-mediated immunity is expression, during which the CRISPR array is transcribed and processed to generate short CRISPR RNAs (crRNAs). Upon subsequent infection, processed crRNAs associate with Cas nucleases to guide the ribonucleoprotein complex to cleave complementary sequences, termed protospacers, in the foreign nucleic acids, which results in both the destruction of the invader's genome and immunity [54].

In terms of the assortment of *cas* genes and the nature of the interference complex, CRISPR/Cas systems are currently classified into two major classes. In this classification, class 1 systems possess multisubunit crRNA–effector complexes, whereas in class 2 systems, all functions of the effector complex are carried out by a single protein, such as Cas9. Each class can be divided into multiple types according to their signature proteins: types I, III, and IV belong to class 1, with Cas3, Cas10, and Csf1 as their respective signature proteins [53], while type II (Cas9), type V (Cas12a–e, Cas12g–i, and Cas14a–c), and type VI (Cas13a–d) belong to class 2 [55–58].

Although all CRISPR–Cas types can recognize specific DNA sequences, not all CRISPR–Cas types are suitable for genetic engineering. Considering the DNA cleavage activity of ribonucleoprotein complexes and the complexity of the systems, type II CRISPR–Cas systems are a simpler mechanism that relies exclusively on the crRNA-guided nuclease Cas9 and its cofactor, trans-encoded crRNA (tracrRNA). There is no requirement for dedicated, repeat-specific endoribonuclease to process the precursor crRNA. In contrast, the type II CRISPR locus produces a short tracrRNA, which is complementary to the repeat sequence [59]. The Cas9 nuclease binds the tracrRNA, and then this dsRNA structure is cleaved by RNase III and generates a complex that contains the Cas9 enzyme, the tracrRNA, and the crRNA guide [60].

Unlike the ZEN and TALEN discussed above, CRISPR/cas9 nuclease does not need to design new and specific proteins for each DNA target site. Because it is easy to locate new sites and the short regions of gRNA only need to be designed and synthesized, this system is considered an effective method for introducing site-specific double-strand breaks (DSBs) [61]. In addition, multiple gRNAs can be simultaneously employed to induce DSBs at various loci. Because the Cas9 protein is not directly connected to the gRNA, this method lends itself well to multiplexing [62].

The CRISPR/Cas system is driving a biotechnology revolution due to its efficacy, modularity, and convenience. As a useful tool to manipulate the genomes of cultured cells, animals, and plants, RNA-guided Cas enzymes vastly speed up the pace of fundamental research and achieve clinical and agricultural breakthroughs.

4.3.4 MAGE/CAGE

Gene editing used to be limited to laborious and serial manipulation of single genes and was not adapted for parallel and continuous directed evolution of gene networks or genomes. In this section, two gene-editing methods, MAGE and CAGE, are introduced to enhance these aspects.

Multiplex automated genome engineering (MAGE) is an oligo-based genome-editing tool that enables large-scale programming and evolution of cells with considerable efficiency. MAGE can carry out multiple modifications simultaneously from a single cell to many cells, as well as across different length scales, from the nucleotide to the genome level [63].

To modify a single location by MAGE, single-stranded oligonucleotides (oligos) containing the desired mutations are inserted into the cell. These oligos are guided to the lagging strand of the replication fork during DNA replication, with the assistance of the bacteriophage λ-Red ssDNA-binding protein β [64], creating new alleles. As the bacteria split, these alleles are distributed throughout the population. Modification of multiple locations simultaneously could be achieved by simply inserting multiple corresponding oligos into the cell.

Note that the efficiency of MAGE is relevant to the scale of modification. Small modifications (a few base pairs) could be made with appreciable efficiency ($>30\%$ for <4 base pairs (bp)) [63]. However, the efficiency of larger modifications in each MAGE cycle substantially drops ($<2\%$ for >20 bp) [65]. As a result, integration of larger genomic sequences such as regulatory elements is limited.

To cope with this deficiency, MAGE was further developed to enable insertion of larger sequences into the genome, namely, coselection MAGE (CoS-MAGE). CoS-MAGE is performed by introducing switchable coselection markers (antibiotic-resistance genes, fluorescent proteins, or metabolic genes) near the targeted location [65]. The enhancement of the method probably takes advantage of the transiently open state of replication forks near the coselection marker, which is more available for oligo-mediated allelic replacement [65].

MAGE features the capability of producing combinatorial genome diversity by repeating the MAGE cycle or activating different coselection markers in CoS-MAGE. MAGE can serve as a multiplex approach for introducing novel or improved properties to organisms in the context of evolution [63]. For example, MAGE has been applied to optimize biosynthesis pathways [66].

CAGE

Conjugative assembly genome engineering (CAGE) is a conjugation-based genome editing tool that allows large-scale hierarchical assembly of multiple modified genomic fragments. Conjugation refers to the process in which genes are transferred between bacteria. In conjugation, the donor bacterium that contains fertility factor (F-factor) produces a pilus that draws the donor and recipient cell close, thus

enabling transfer of genetic material (usually a plasmid or a small, circular piece of DNA). In each cycle of CAGE, the strains are divided into conjugation pairs, with each pair containing a donor strain and a recipient strain.

In CAGE, genome transfer is controlled by the precise placement of an origin of transfer (oriT), where DNA transfer is initiated, as well as positive and positive-negative selectable markers that function as anchor points in the hierarchical assembly process [67, 68]. Note that the insertion of these markers is based on λ-Red recombineering and determines the composition of the yielded strain [69].

In the donor strain, the desired genomic region of transfer is flanked by an upstream oriT and a downstream positive selective marker [67, 68]. The donor strain also contains an F-factor to allow conjugation. In the recipient strain, the desired region is flanked with a different upstream positive selective marker and a downstream positive–negative selective marker [67, 68]. With the F-plasmid in the donor strain, conjugation occurs, and the desired sequence is transferred. Subsequently, a specific set of three simultaneous selections matching the selective marker is conducted to yield a recombinant strain, which should contain both desired regions from its conjugation parents [67]. This recombinant strain could participate in another CAGE cycle to produce a new, higher-order recombinant strain. By implementing this conjugation-based and selective strategy in multiple stages, genomes of multiple strains could be merged into a single strain containing all the desired genomic regions [67].

Integrated Application of MAGE and CAGE in Evolutionary

MAGE enables the introduction of small mutations by oligo-mediated allelic replacement targeting multiple sites, while repeating the MAGE cycle contributes to combinatorial diversity. CAGE enables successive assembly of multiple genomic sequences into a single genome. These two genomic-editing tools offer an approach to study central evolutionary issues in which natural genetic variation is limited or biased [70].

A significant example is codon reassignment. By replacing the target codon synonymously, the target codon is yielded blank and could be allocated to novel functions. This large-scale replacement procedure relies on the integrated application of MAGE and CAGE to a considerable extent [71, 72]. In such fields, MAGE and CAGE provide easy and efficient access to expand the genetic code and construct genomically recoded organisms.

4.4 Directed Evolution and Artificial Design

4.4.1 Directed Evolution: PACE and RAGE

Having been proven to be a highly effective and broadly applicable framework in basic and applied biology, directed evolution mimics natural evolution in more controlled settings to evolve biological systems toward user-defined phenotypes through iterative rounds of genetic diversification and library screening or selection [73–75].

Generally, there is no essential difference between directed evolution and natural selection. Mutagenesis and recombination of wild-type DNA sequences are carried out to establish a variety of mutant libraries. In the process of evolution, the mutant library is screened to achieve the desired purpose, either the most suitable protein becomes the parent of the next generation, or several more suitable mutants are recombined [76, 77].

Directed evolution often relies on genetic diversity and library screening to produce increasingly complex and particular phenotypes over tolerable timescales. Therefore, the capacity to produce functional diversity and appropriate screening, or selection techniques to detect variations, plays a significant role in directed evolution.

Phage-assisted continuous evolution (PACE) uses a modified M13 phage cycle to boost evolution efficiency and link the expression of a phage protein to the trait of interest [78]. PACE can accomplish hundreds of rounds of directed evolution in less than a week with little assistance from researchers because a full phage replication cycle can be completed in as little as 10 min [79]. More recently, PACE has been improved by adjusting selection stringency and the ability to perform negative selection against undesired variations [80, 81].

RNAi-assisted genome evolution (RAGE) is a genome-wide engineering approach that can continuously enhance desired traits by accumulating beneficial downregulations of genes [82]. RNAi is a gene-silencing mechanism widely found in eukaryotes in which messenger RNA (mRNA) is targeted for degradation by homologous double-stranded RNA (dsRNA) [83, 84]. RNAi screening achieves genome-wide knockdown interference in the absence of allele modification and is widely used in eukaryotic systems [85, 86]. RNA-based whole-genome engineering can generate a high-coverage repression pool library without causing permanent changes to the genome.

To date, directed evolution has evolved into a powerful, convenient tool for the engineering of protein and whole-cell biocatalysts, making contributions to diverse fields such as human health, energy, and the environment [87, 88].

4.4.2 *Artificial Design Tools*

Synthetic biology can synthesize genes, pathways, genomes, and chromosomes. Some individual synthetic elements, such as promoters, coding regions, and transcription terminators, can be cloned and characterized individually. However, when the various components are combined together, how to form a whole system and work together to minimize interference from each other remains a challenge. Therefore, researchers have designed numerous artificial design tools to guide the assembly and use of components.

The Lawrence Berkeley National Laboratory reported website-based biological computer-aided design software DeviceEditor [89] and J5 [90] in 2012. They integrate various related resources and design automated DNA parts assembly schemes through software, which reduces the error and the number of off-target products and makes it possible for computer-aided design of multiple parts combinations. Researchers from the Johns Hopkins University developed BiopartsBuilder [91] in 2015 to design synthetic biological components that comply with related DNA assembly methods such as golden gates. It retrieves biological sequences from different databases, standardizes assembly design, and provides design solutions for parts, providing useful integrated tools for synthetic biology. SynBIS, an information system based on the synthetic biology website developed by the lab from Imperial College London [92], integrates a variety of biological computer-aided design programs, DNA assembly element modules, and characteristic parameters of various biological parts.

In addition to the assembly of parts, the artificially assisted evolutionary design of biologically active components has received increasing attention in recent years. Researchers in UCSF developed a modular protein sensing system using protein computational design tools [93], and its computational design strategy opened up a broad channel for the connection of biological outputs and new signals. In recent years, the use of artificial intelligence technologies such as machine learning to promote the complex design of synthetic biology has also received increasing attention. Machine learning systems for biological elements such as promoters [94], synthetic pathways [95], and their assembly schemes can achieve automated design of the "design–build–test–learn" (DBTL) process.

Researchers can also use directed evolution guided by machine learning to design proteins and optimize protein functions without the need for basic physical or biological pathway models [96]. By learning from the characteristics of the variants and selecting improved characteristic sequences, directed evolution can be effectively accelerated to discover unknown protein functions and reveal the relationship between protein sequences and functions, providing a brand-new way for the on-demand development of high-quality protein parts. In the CASP competition where protein structure prediction is the main goal, machine learning methods represented by Alphafold/Alphafold2 [97] have been getting closer to this possibility.

4.5 Chassis

4.5.1 From Microorganisms to Cells

Natural products from plants and animals are widely used in the fields of industry, medical treatment, environment, food, and nutrition. Traditionally, they are extracted from native organisms, which is usually inefficient and complex. These natural producers may not be ideal hosts for bioproduction either due to their slow growth rate or lack of efficient genetic manipulation techniques. Therefore, the heterologous expression of the synthetic pathway for natural products in microbial cells has attracted increasing attention.

Microorganisms are small single-celled organisms consisting of bacteria, fungi, and viruses that can be found either on land or in the sea. Archaeologists have solid evidence that microorganisms are the ancestors of multicellular organisms. These small creatures might seem inconspicuous but are complete machines. Even relatively ancient prokaryotes have ribosomes to accomplish basic life activities and have circular genomes to pass on their genetic material to the next generation. The sums of microorganisms are enormous and macroscopic. The discovery of plasmids changed their fate, as well as the fate of mankind. Plasmids are small circular DNA molecules independent of the genome. Microorganisms can absorb plasmids from outside by nature or by manual methods. Genes carried on the plasmids give microorganisms unique abilities that they do not have before. Plasmids give microorganisms great characteristics for heterologous expression and are used to transfer specific genes in microorganisms to realize the functions we need. The microorganisms used to be engineered in synthetic biology are called chassis. The most widely used chassis for microorganisms is *Escherichia coli*. As chassis, they have advantages due to their fast growth rates, well-studied genomes, and metabolic networks. Massive strains of *E. coli* (e.g., DH5α) were engineered to meet different requirements and realize heterologous expression by transferring plasmids into the engineered strain. This process is easy to manipulate and can produce large-scale and high-value natural products. Therefore, *E. coli* has become the most classic chassis.

Chasses expand from microorganisms to cells with the development of synthetic biology. In addition to prokaryotes, scientists have further developed methods to transfer plasmids into eukaryotic microorganisms, making them chassis.

Engineered yeast has the same advantages as prokaryotes but higher biosafety since it was used in fermentation thousands of years ago. The earliest yeast used for protein expression was *Saccharomyces cerevisiae*. Alcohol oxidase gene-1 (AOX1) is contained in most of the expression vectors of methanol yeast, which can induce foreign gene expression under the condition of a strong promoter and methanol as the sole carbon source. Protein expressed by methanol yeast usually takes a long time to reach the peak level.

Baculovirus–insect cell expression system is based on baculovirus vectors to direct the expression of foreign genes, most of which contain the polyhedrin

promoter of *Autographa californica* nuclear polyhedrosis virus (AcNPV). The protein expression level of the baculovirus system is high, and most of the proteins can remain soluble. The baculovirus genome is large (130 kb) and can accommodate large foreign DNA fragments.

Mammalian cells can also be used as chassis. The eukaryotic proteins it expresses can usually be modified correctly and are closest to natural proteins in terms of molecular structure, physical and chemical properties, and biological functions. It can be accurately positioned in cells and is widely used in medical research. Although the expression in mammalian cells is more difficult, time-consuming, and cost-intensive than the expression in *E. coli,* it is still very practical for researchers to express protein with cell culture.

4.5.2 Animals and Plants

With the deepening of research, the objective of synthetic biology has gradually stepped into the complex multicellular eukaryotic system from the initial single-celled prokaryotic microorganism system, among which synthetic biology research based on plants and animals has gradually attracted attention [98].

Compared with single-celled microorganisms, plants have various endomembrane systems and organelles. Their complex spatial characteristics provide the optimal environment for the synthesis of different metabolites, as well as an excellent model system for the study of synthetic biology. Plant chassis have natural advantages, requiring only CO_2 and water as raw material. Through photosynthesis, they can synthesize all kinds of complex natural products. The metabolic networks can be further redesigned by synthetic biology, and the carbon flow can be further redirected to more valuable specific metabolites. Microbial chassis, on the other hand, requires a cheaper source of energy and more demanding fermentation conditions. Moreover, as multicellular organisms with complex differentiation, the fine division of labor and collaboration among different organs of plants also makes it possible to realize artificial design of complex functions. In some cases, natural products are synthesized in different organelles or tissues in plant cells, and the localization of enzymes in different organelles can reduce the toxic effects of intermediates on cells, thus effectively increasing production. For example, Malhotra et al. [99] designed the partition of the synthesis path by using tobacco, the commonly used plant chassis, to synthesize artemisinin. ADS was localized in mitochondria by signal peptide Cox4 fusion. CYP71AV1, CPR, and DBR2 were localized in chloroplasts by plastid-conducting peptides. Meanwhile, the entire MVA pathway of yeast was introduced into chloroplasts to increase the supply of terpenoid precursors, thus improving the yield of artemisinin. In addition, plant chassis also has obvious advantages in membrane protein expression, precursor supply, product tolerance, etc. Aside from tobacco, *Arabidopsis thaliana* [100], rice [101], algae [102], and tomato [103] are also the commonly used plant chassis.

Plant synthetic biology research is still in its infancy. To date, a few natural products of plant origin have been fully analyzed because plant genomes are larger than those of microorganisms, and only a few have been accurately analyzed. The accuracy of the synthetic pathway of natural products and the sequence of enzymatic reactions in the metabolic pathway are also important factors that hinder the further development of plant chassis [104]. In the process of analyzing new metabolic pathways and discovering new enzymatic reactions, still many challenges remain to confirm the function of candidate genes. An increasing number of studies have shown that some proteins without catalytic activity themselves play key regulatory roles in catalytic reactions [105]. Moreover, the functional characterization of plant regulatory elements remains at the qualitative and quantitative level of single-regulatory elements and single-regulatory circuits, and the internal functions and formation mechanisms of most plant DNA and protein elements remain unclear and are still in the stage of further exploration.

Compared with plant chassis, animal chassis is less studied, but its importance cannot be denied. Over the past years, synthetic biology applied to higher eukaryotes, such as mammalian cells or transgenic animals, has evolved from simple gene switches and gene networks to more complex and treatment-oriented circuits [106]. Mammalian synthetic biology provides strategies for gene and cellular therapies, such as personalized medicine, cancer treatment, and metabolic and immune disorders.

4.5.3 Cell-Free System

A cell-free system is a kind of extract that retains the ability of protein biosynthesis and is used for protein expression in vitro. Cell-free systems usually include ribosomes, various tRNAs, various aminoacyl-tRNA synthetases, and other necessary components for protein synthesis, which can synthesize proteins in vitro using exogenous mRNA and DNA as templates. The process of protein synthesis in the cell-free system is roughly consistent with the following. First, mRNA was synthesized in vitro by plasmid DNA or PCR products in the presence of RNA polymerase. Then, mRNA is translated into protein by using translation factors in the system. Enzymes are required to synthesize proteins, energy substances, tRNA, and amino acids.

Compared with traditional protein expression systems in vivo, cell-free systems have the following advantages [107, 108]. The cell-free system is not affected by the complex environment inside the cell. It can directly control protein synthesis and processes by regulating the reaction conditions and avoid the formation of inclusion bodies at the same time. Cell-free systems avoid the toxic effect of proteins and can be used to express proteins that are harmful to cells. In the process of product separation, a cell-free system does not need cell fragmentation or separation, which reduces the cost of the product.

Early cell-free systems used batch methods, such as the *E. coli* S30 system established by Zubay [109] in 1973. However, this method has some problems, such as a short reaction time and low yield. In the mid-1980s, Spirin et al. invented the continuous-flow cell-free (CFCF) method, in which energy and substrate are continuously fed into the acellular system, and then an ultrafiltration membrane is used to filter out the product at the same rate [110]. This makes the cell-free system remain always in a state of dynamic equilibrium, which greatly prolongs the reaction time and increases the yield of protein synthesis. Afterward, a more efficient CECF was developed [111]. In this CECF, the reaction system and the supplementary system are separated by a semipermeable membrane with selective permeability, and the substrate and energy are supplemented by diffusion. At present, there are several mature systems, such as *E. coli* extract, wheat germ extract, and rabbit reticulocyte extract [112, 113].

At present, the technology of protein synthesis using cell-free systems has been widely used in proteomics research, vaccine development, high-throughput drug screening, and other biomedical fields. On the other hand, cell-free expression technology can also be used for the synthesis of small molecular natural products such as artemisinin and lycopene.

References

1. Benner, S.A., Sismour, A.M.: Synthetic biology. Nat. Rev. Genet. **6**(7), 533–543 (2005)
2. Michelson, A.M., Todd, A.R.: Nucleotides part XXXII. Synthesis of a dithymidine dinucleotide containing a 3′: 5′-internucleotidic linkage. J. Chem. Soc., 2632–2638 (1955)
3. Beaucage, S.L.C.M.H.: Deoxynucleoside phosphoramidites - a new class of key intermediates for deoxypolynucleotide synthesis. Tetrahedron Lett. **22**(20), 1859–1862 (1981)
4. Kosuri, S., Church, G.M.: Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods. **11**(5), 499–507 (2014)
5. LeProust, E.M., et al.: Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Res. **38**(8), 2522–2540 (2010)
6. Tian, J., Ma, K., Saaem, I.: Advancing high-throughput gene synthesis technology. Mol. BioSyst. **5**(7), 714–722 (2009)
7. Ma, S., Tang, N., Tian, J.: DNA synthesis, assembly and applications in synthetic biology. Curr. Opin. Chem. Biol. **16**(3–4), 260–267 (2012)
8. Kosuri, S., et al.: Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nat. Biotechnol. **28**(12), 1295–1299 (2010)
9. Matzas, M., et al.: High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. Nat. Biotechnol. **28**(12), 1291–1294 (2010)
10. Bollum, F.J.: Thermal conversion of nonpriming deoxyribonucleic acid to primer. J. Biol. Chem. **234**(10), 2733–2734 (1959)
11. Bollum, F.J.: Oligodeoxyribonucleotide-primed reactions catalyzed by calf thymus polymerase. J. Biol. Chem. **237**(6), 1945–1949 (1962)
12. Jensen, M.A., Davis, R.W.: Template-independent enzymatic oligonucleotide synthesis (TiEOS): its history, prospects, and challenges. Biochemistry. **57**(12), 1821–1832 (2018)
13. Motea, E.A., Berdis, A.J.: Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. Biochim. Biophys. Acta. **1804**(5), 1151–1166 (2010)

14. Tjong, V., et al.: Amplified on-chip fluorescence detection of DNA hybridization by surface-initiated enzymatic polymerization. Anal. Chem. **83**(13), 5153–5159 (2011)
15. Galdzicki, M., et al.: Standard biological parts knowledgebase. PLoS One. **6**(2), e17005 (2011)
16. Densmore, D., et al.: Algorithms for automated DNA assembly. Nucleic Acids Res. **38**(8), 2607–2616 (2010)
17. Vick, J.E., et al.: Optimized compatible set of BioBrick vectors for metabolic pathway engineering. Appl. Microbiol. Biotechnol. **92**(6), 1275–1286 (2011)
18. Stahl, F.W.: The Holliday junction on its thirtieth anniversary. Genetics. **138**(2), 241–246 (1994)
19. Meselson, M.S., Radding, C.M.: A general model for genetic recombination. Proc. Natl. Acad. Sci. U. S. A. **72**(1), 358–361 (1975)
20. Szostak, J.W., et al.: The double-strand-break repair model for recombination. Cell. **33**(1), 25–35 (1983)
21. Valancius, V., Smithies, O.: Testing an "in-out" targeting procedure for making subtle genomic modifications in mouse embryonic stem cells. Mol. Cell. Biol. **11**(3), 1402–1408 (1991)
22. Hasty, P., et al.: Introduction of a subtle mutation into the Hox-2.6 locus in embryonic stem cells. Nature. **350**(6315), 243–246 (1991)
23. Askew, G.R., Doetschman, T., Lingrel, J.B.: Site-directed point mutations in embryonic stem cells: a gene-targeting tag-and-exchange strategy. Mol. Cell. Biol. **13**(7), 4115–4124 (1993)
24. Engler, C., Kandzia, R., Marillonnet, S.: A one pot, one step, precision cloning method with high throughput capability. PLoS One. **3**(11), e3647 (2008)
25. Engler, C., et al.: Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. PLoS One. **4**(5), e5553 (2009)
26. Gibson, D.G., et al.: Complete chemical synthesis, assembly, and cloning of a mycoplasma genitalium genome. Science. **319**(5867), 1215–1220 (2008)
27. Gibson, D.G., et al.: Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods. **6**(5), 343–345 (2009)
28. Larionov, V., et al.: Specific cloning of human DNA as yeast artificial chromosomes by transformation-associated recombination. Proc. Natl. Acad. Sci. U. S. A. **93**(1), 491–496 (1996)
29. Larionov, V., et al.: Direct isolation of human BRCA2 gene by transformation-associated recombination in yeast. Proc. Natl. Acad. Sci. U. S. A. **94**(14), 7384–7387 (1997)
30. Kouprina, N., Larionov, V.: TAR cloning: perspectives for functional genomics, biomedicine, and biotechnology. Mol Ther Methods Clin Dev. **14**, 16–26 (2019)
31. Kouprina, N., Larionov, V.: Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology. Chromosoma. **125**(4), 621–632 (2016)
32. Lee, N.C., Larionov, V., Kouprina, N.: Highly efficient CRISPR/Cas9-mediated TAR cloning of genes and chromosomal loci from complex genomes in yeast. Nucleic Acids Res. **43**(8), e55 (2015)
33. Theis, J.F., Newlon, C.S.: The ARS309 chromosomal replicator of Saccharomyces cerevisiae depends on an exceptional ARS consensus sequence. Proc. Natl. Acad. Sci. U. S. A. **94**(20), 10786–10791 (1997)
34. Baudin, A., et al.: A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. Nucleic Acids Res. **21**(14), 3329–3330 (1993)
35. Kouprina, N., et al.: Cloning of human centromeres by transformation-associated recombination in yeast and generation of functional human artificial chromosomes. Nucleic Acids Res. **31**(3), 922–934 (2003)
36. Noskov, V.N., et al.: A general cloning system to selectively isolate any eukaryotic or prokaryotic genomic region in yeast. BMC Genomics. **4**(1), 16 (2003)
37. Noskov, V., et al.: A genetic system for direct selection of gene-positive clones during recombinational cloning in yeast. Nucleic Acids Res. **30**(2), E8 (2002)

38. Sternberg, N., Hamilton, D.: Bacteriophage P1 site-specific recombination. I. Recombination between loxP sites. J. Mol. Biol. **150**(4), 467–486 (1981)
39. Sauer, B.: Functional expression of the cre-lox site-specific recombination system in the yeast Saccharomyces cerevisiae. Mol. Cell. Biol. **7**(6), 2087–2096 (1987)
40. Sternberg, N., et al.: Bacteriophage P1 cre gene and its regulatory region. Evidence for multiple promoters and for regulation by DNA methylation. J. Mol. Biol. **187**(2), 197–212 (1986)
41. Hoess, R.H., Ziese, M., Sternberg, N.: P1 site-specific recombination: nucleotide sequence of the recombining sites. Proc. Natl. Acad. Sci. U. S. A. **79**(11), 3398–3402 (1982)
42. Sauer, B.: Inducible gene targeting in mice using the Cre/lox system. Methods. **14**(4), 381–392 (1998)
43. Garrick, D., et al.: Repeat-induced gene silencing in mammals. Nat. Genet. **18**(1), 56–59 (1998)
44. Carroll, D.: Genome engineering with zinc-finger nucleases. Genetics. **188**(4), 773–782 (2011)
45. Palpant, N.J., Dudzinski, D.: Zinc finger nucleases: looking toward translation. Gene Ther. **20**(2), 121–127 (2013)
46. Li, H.L., Nakano, T., Hotta, A.: Genetic correction using engineered nucleases for gene therapy applications. Develop. Growth Differ. **56**(1), 63–77 (2014)
47. Gaj, T., Gersbach, C.A., Barbas 3rd, C.F.: ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. **31**(7), 397–405 (2013)
48. Kim, Y., et al.: A library of TAL effector nucleases spanning the human genome. Nat. Biotechnol. **31**(3), 251–258 (2013)
49. Makarova, K.S., Wolf, Y.I., Koonin, E.V.: Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. **41**(8), 4360–4377 (2013)
50. Wiedenheft, B., Sternberg, S.H., Doudna, J.A.: RNA-guided genetic silencing systems in bacteria and archaea. Nature. **482**(7385), 331–338 (2012)
51. van der Oost, J., et al.: CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem. Sci. **34**(8), 401–407 (2009)
52. Heler, R., Marraffini, L.A., Bikard, D.: Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. Mol. Microbiol. **93**(1), 1–9 (2014)
53. Makarova, K.S., et al.: An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. **13**(11), 722–736 (2015)
54. Perez Rojo, F., et al.: CRISPR-Cas systems: ushering in the new genome editing era. Bioengineered. **9**(1), 214–221 (2018)
55. Burstein, D., et al.: New CRISPR-Cas systems from uncultivated microbes. Nature. **542**(7640), 237–241 (2017)
56. Harrington, L.B., et al.: Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. Science. **362**(6416), 839–842 (2018)
57. Liu, J.J., et al.: CasX enzymes comprise a distinct family of RNA-guided genome editors. Nature. **566**(7743), 218–223 (2019)
58. Yan, W.X., et al.: Functionally diverse type V CRISPR-Cas systems. Science. **363**(6422), 88–91 (2019)
59. Garneau, J.E., et al.: The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. **468**(7320), 67–71 (2010)
60. Jinek, M., et al.: A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. **337**(6096), 816–821 (2012)
61. Zhang, J.H., et al.: Optimization of genome editing through CRISPR-Cas9 engineering. Bioengineered. **7**(3), 166–174 (2016)
62. Kleinstiver, B.P., et al.: High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature. **529**(7587), 490–495 (2016)
63. Wang, H.H., et al.: Programming cells by multiplex genome engineering and accelerated evolution. Nature. **460**(7257), 894–898 (2009)

64. Ellis, H.M., Yu, D., DiTizio, T.: High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. Proc. Natl. Acad. Sci. **98**(12), 6742–6746 (2001)
65. Wang, H.H., et al.: Genome-scale promoter engineering by coselection MAGE. Nat. Methods. **9**(6), 591–593 (2012)
66. Raman, S., et al.: Evolution-guided optimization of biosynthetic pathways. Proc. Natl. Acad. Sci. **111**(50), 17803–17808 (2014)
67. Isaacs, F.J., et al.: Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. Science. **333**(6040), 348–353 (2011)
68. Güell, M.: Conjugative assembly genome engineering (CAGE). In: de la Cruz, F. (ed.) Horizontal Gene Transfer: Methods and Protocols, pp. 399–409. Springer US, New York, NY (2020)
69. Datsenko, K.A., Wanner, B.L.: One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. **97**(12), 6640–6645 (2000)
70. Pál, C., Papp, B., Pósfai, G.: The dawn of evolutionary genome engineering. Nat. Rev. Genet. **15**(7), 504–512 (2014)
71. Lajoie, M.J., et al.: Genomically recoded organisms expand biological functions. Science. **342**(6156), 357–360 (2013)
72. Lajoie, M., et al.: Probing the limits of genetic recoding in essential genes. Science. **342**(6156), 361–363 (2013)
73. Bassalo, M.C., Liu, R., Gill, R.T.: Directed evolution and synthetic biology applications to microbial systems. Curr. Opin. Biotechnol. **39**, 126–133 (2016)
74. Chen, K., Arnold, F.H.: Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. Proc. Natl. Acad. Sci. U. S. A. **90**(12), 5618–5622 (1993)
75. Stemmer, W.P.: Rapid evolution of a protein in vitro by DNA shuffling. Nature. **370**(6488), 389–391 (1994)
76. Moore, J.C., Arnold, F.H.: Directed evolution of a Para-nitrobenzyl esterase for aqueous-organic solvents. Nat. Biotechnol. **14**(4), 458–467 (1996)
77. Moore, J.C., et al.: Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences. J. Mol. Biol. **272**(3), 336–347 (1997)
78. Esvelt, K.M., Carlson, J.C., Liu, D.R.: A system for the continuous directed evolution of biomolecules. Nature. **472**(7344), 499–503 (2011)
79. Nelson, F.K., Friedman, S.M., Smith, G.P.: Filamentous phage DNA cloning vectors: a noninfective mutant with a nonpolar deletion in gene III. Virology. **108**(2), 338–350 (1981)
80. Sergeeva, A., et al.: Display technologies: application for the discovery of drug and gene delivery agents. Adv. Drug Deliv. Rev. **58**(15), 1622–1654 (2006)
81. Yuan, L., et al.: Laboratory-directed protein evolution. Microbiol. Mol. Biol. Rev. **69**(3), 373–392 (2005)
82. Si, T., HamediRad, M., Zhao, H.: Regulatory RNA-assisted genome engineering in microorganisms. Curr. Opin. Biotechnol. **36**, 85–90 (2015)
83. Fire, A., et al.: Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. **391**(6669), 806–811 (1998)
84. Hannon, G.J.: RNA interference. Nature. **418**(6894), 244–251 (2002)
85. Echeverri, C.J., Perrimon, N.: High-throughput RNAi screening in cultured cells: a user's guide. Nat. Rev. Genet. **7**(5), 373–384 (2006)
86. Boutros, M., Ahringer, J.: The art and design of genetic screens: RNA interference. Nat. Rev. Genet. **9**(7), 554–566 (2008)
87. Si, T., et al.: Automated multiplex genome-scale engineering in yeast. Nat. Commun. **8**, 15187 (2017)
88. Zhao, H., Chockalingam, K., Chen, Z.: Directed evolution of enzymes and pathways for industrial biocatalysis. Curr. Opin. Biotechnol. **13**(2), 104–110 (2002)
89. Chen, J., et al.: DeviceEditor visual biological CAD canvas. J. Biol. Eng. **6**(1), 1 (2012)

90. Hillson, N.J., Rosengarten, R.D., Keasling, J.D.: j5 DNA assembly design automation software. ACS Synth. Biol. **1**(1), 14–21 (2012)
91. Yang, K., et al.: BioPartsBuilder: a synthetic biology tool for combinatorial assembly of biological parts. Bioinformatics. **32**(6), 937–939 (2016)
92. Clarke, L.J., Kitney, R.I.: Synthetic biology in the UK—an outline of plans and progress. Synth Syst Biotechnol. **1**(4), 243–257 (2016)
93. Glasgow, A.A., et al.: Computational design of a modular protein sense-response system. Science. **366**(6468), 1024–1028 (2019)
94. de Boer, C.G., et al.: Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat. Biotechnol. **38**(1), 56–65 (2020)
95. HamediRad, M., et al.: Towards a fully automated algorithm driven platform for biosystems design. Nat. Commun. **10**(1), 5150 (2019)
96. Yang, K.K., Wu, Z., Arnold, F.H.: Machine-learning-guided directed evolution for protein engineering. Nat. Methods. **16**(8), 687–694 (2019)
97. Callaway, E.: It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature. **588**(7837), 203–204 (2020)
98. Liu, W., Stewart Jr., C.N.: Plant synthetic biology. Trends Plant Sci. **20**(5), 309–317 (2015)
99. Malhotra, K., et al.: Compartmentalized metabolic engineering for artemisinin biosynthesis and effective malaria treatment by oral delivery of plant cells. Mol. Plant. **9**(11), 1464–1477 (2016)
100. Besumbes, O., et al.: Metabolic engineering of isoprenoid biosynthesis in Arabidopsis for the production of taxadiene, the first committed precursor of Taxol. Biotechnol. Bioeng. **88**(2), 168–175 (2004)
101. Ye, X., et al.: Engineering the provitamin a (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. Science. **287**(5451), 303–305 (2000)
102. Anterola, A., et al.: Production of taxa-4(5),11(12)-diene by transgenic Physcomitrella patens. Transgenic Res. **18**(4), 655–660 (2009)
103. Kovacs, K., et al.: Redirection of carotenoid metabolism for the efficient production of taxadiene [taxa-4(5),11(12)-diene] in transgenic tomato fruit. Transgenic Res. **16**(1), 121–126 (2007)
104. Nutzmann, H.W., Huang, A., Osbourn, A.: Plant metabolic clusters—from genetics to genomics. New Phytol. **211**(3), 771–789 (2016)
105. Ban, Z., et al.: Noncatalytic chalcone isomerase-fold proteins in Humulus lupulus are auxiliary components in prenylated flavonoid biosynthesis. Proc. Natl. Acad. Sci. U. S. A. **115**(22), E5223–E5232 (2018)
106. Kis, Z., et al.: Mammalian synthetic biology: emerging medical applications. J. R. Soc. Interface. **12**(106) (2015)
107. Carlson, E.D., et al.: Cell-free protein synthesis: applications come of age. Biotechnol. Adv. **30**(5), 1185–1194 (2012)
108. Hodgman, C.E., Jewett, M.C.: Cell-free synthetic biology: thinking outside the cell. Metab. Eng. **14**(3), 261–269 (2012)
109. Zubay, G.: In vitro synthesis of protein in microbial systems. Annu. Rev. Genet. **7**, 267–287 (1973)
110. Spirin, A.S., et al.: A continuous cell-free translation system capable of producing polypeptides in high yield. Science. **242**(4882), 1162–1164 (1988)
111. Kim, D.M., Choi, C.Y.: A semicontinuous prokaryotic coupled transcription/translation system using a dialysis membrane. Biotechnol. Prog. **12**(5), 645–649 (1996)
112. Endo, Y., Sawasaki, T.: Cell-free expression systems for eukaryotic protein production. Curr. Opin. Biotechnol. **17**(4), 373–380 (2006)
113. Endo, Y., Sawasaki, T.: High-throughput, genome-scale protein production method based on the wheat germ cell-free expression system. Biotechnol. Adv. **21**(8), 695–713 (2003)

Chapter 5
Synthetic Biology: Development Issues

Kang Ning, Yi Zhan, and Dan Zhao

Abstracts Many development issues have arisen in synthetic biology, including a lack of connection among omics data, the availability of various hardware resources such as parts and modules, and well-curated software resources a scarcity of standard platforms to evaluate the synthetic system's effectiveness, etc. The vision of synthetic biology includes three aspects, including a predictable engineering discipline, a competitive supply chain, and accessible and desirable products and services; thus, the three main challenges faced by synthetic biology are predictability, supply chain, and accessibility. In this chapter, we will look at these challenges and issues of synthetic biology.

Keywords Challenges · Standard platform · Development issues

Many development issues have arisen in synthetic biology, including a lack of connectivity with omics data, the availability of various hardware resources such as parts and modules, and well-curated software resources a scarcity of standard platforms to evaluate the synthetic system's effectiveness, etc. (Fig. 5.1).

5.1 Challenges Faced by Synthetic Biology

5.1.1 Predictability

Since biology is incredibly unpredictable, it is currently impossible to ensure that organisms can output according to the requirements and predicted results of the initial design [1].

K. Ning (✉) · Y. Zhan · D. Zhao
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: ningkang@hust.edu.cn

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8_5

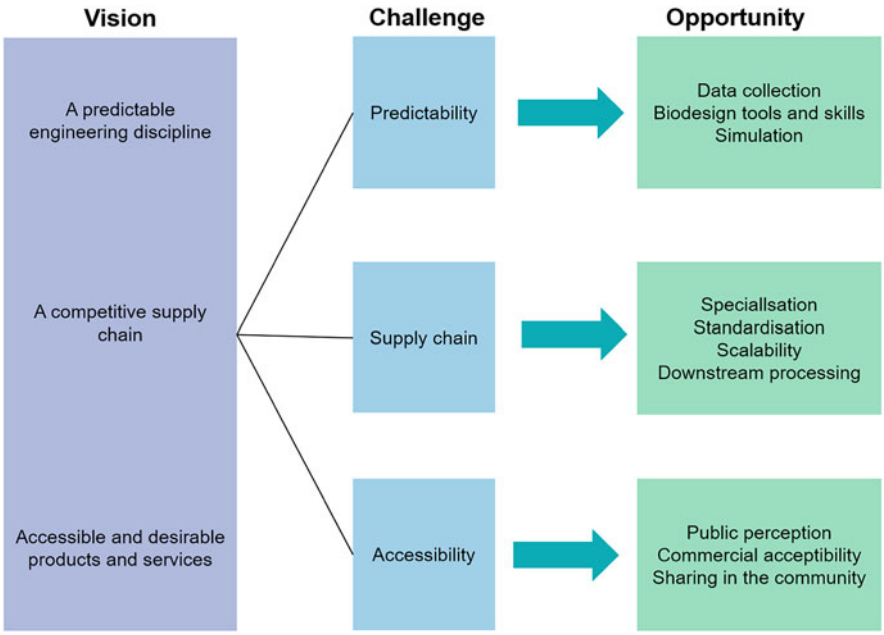

Fig. 5.1 Three main challenges faced by synthetic biology. The vision of synthetic biology includes three aspects, including a predictable engineering discipline, a competitive supply chain, and accessible and desirable products and services; thus, the three main challenges faced by synthetic biology are predictability, supply chain, and accessibility

The synthetic biology industry includes a series of design tools and various procedures in which raw materials are transformed into advanced equipment and systems. However, these tools and procedures used in synthetic biology have not been fully defined and standardized, so we cannot predict output based on input.

By comparing with other industries, we can determine what is needed to achieve better predictability in the synthetic biology industry. First, in the design and manufacturing process of some products, the desired characteristics of products can be defined in the early stages of the design process. However, for synthetic biology, this could only be achieved in one circumstance: if the representative cycle "design–build–test" runs enough times so that the input will produce the expected output.

Second, there are no comprehensive, universal tools to design for synthetic biology, and we need to gather a series of complete, interoperable toolboxes so that any biological designer is capable of using these tools. In addition, biological designers themselves must have the necessary skills to use these tools [2].

Finally, only when the design process is simulated by modeling tools can the process of production be entered. In biological design, computer modeling and simulation also play a critical role in transforming design inputs into predictable outputs [3].

To achieve a predictable biological design that meets our expectations in the next 5–10 years, we need to make a breakthrough in the fields of data collection, tools and technology of biological design, and computer-based simulation.

Data Collection

A key factor that drives the transition from synthetic biology to biological design is the ability to collect and analyze high-quality data from biological and engineering processes. Nevertheless, there are problems both in collecting and comparing data from multiple systems at the hardware and software levels. Currently, in synthetic biology, we rely on simple measurement methods, such as fluorescence, temperature, dissolved oxygen, and pH, to evaluate biological processes in microbes or cell culture. There is almost no way to directly measure parameters such as the cell physiological state, the expression level of products, or the carbon input concentration in real time. Fluorescence might be a reliable method, but there are many problems. In the process of fluorescence detection, proteins need to be highly expressed, and the degradation rate of fluorescent proteins will limit the ability to measure kinetic characteristics [4]. Consequently, the lack of data hinders the ability to build and test predictive models. Therefore, we need to develop reliable, low-cost systems to analyze biological processes. One possibility is to develop software that converts information from a series of sensors and instruments into standard output.

At the same time, two more problems are difficult to solve. The first is the lack of measurement tools, limiting the collected data. The solution is to collect more data. The second is the opposite. With the promotion of omics measurement tools [5], we have obtained an increasing amount of data at various omics levels, but we are not capable of analyzing them for valuable information. The solution is to develop more software for data analysis. Machine learning/AI methods may help us to transform large datasets into usable information. In conclusion, only if improvements in all the key areas mentioned above are made can we collect the wide-range, multilevel, and parameter-rich data that are needed for predictable and overall modeling.

Tools and Technology of Biological Design

A well-known problem in synthetic biology is that combining elements into genetic circuits often leads to results that are different from expectations. To solve the problem in genetic circuit design, synthetic biologists often choose high-throughput methods, conduct many trial-and-error experiments, and produce many design prototypes. This actually impairs the ability of synthetic biologists to apply design principles to biology and increases the cost of biological design. The design space covered by other more traditional methods is actually smaller and more dependent on the so-called “manual” ability of the scientist. What would happen if biological design tools were widely adopted to avoid the need for high-throughput methods and combined with the “manual” capabilities of biological design? As some participants

pointed out, there is indeed a need for standardization throughout the supply chain. Biological design tools have seamless interoperability [6], scalability, and the ability to support the development of biological products that meet specifications. These tools need to be adopted throughout the supply chain in a manner comparable to existing engineering design tools while allowing biological designers to apply design thinking and take advantage of the unique innovation capabilities of biology.

5.1.2 Competitiveness of Supply Chains

Supply chains of synthetic biology lack competitiveness and are not suitable for applications [7].

5.1.3 Accessibility

It is difficult to make nonprofessionals believe in the potential of biological design to solve problems in an understandable way [8].

5.2 The Lack of a Standard Platform to Evaluate the Effectiveness of the Synthetic System

5.2.1 Systems Modeling, Standards, and Metrology

Regardless of the biological system under investigation, a reliable model of that system is required if predictable changes to genetic circuits are to be suggested and put into practice with any degree of assurance. To design these models, enormous amounts of data from measurements of various cell behavior characteristics in a variety of contexts are needed [9].

As a result of technological developments in high-throughput data monitoring and collection equipment, large datasets are projected to be created correctly. Artificial intelligence or machine learning approaches can be used to study them to enhance the design of synthetic biological products and prevent unsuccessful trial-and-error approaches [10].

By coming to an agreement on standards of design, assembly, data transfer, data measurement, and regulatory requirements, as well as on the terminology that is used, the interdisciplinary and international cooperation required to progress the subject will be improved. This is challenging for a group with such a diversity of interests and points of view particularly when data exchange, curation, and quality control are not common procedures.

However, many synthetic biology techniques and products do not work well in industrial settings that depend on repeatable procedures and are subject to stringent regulatory requirements without some sort of agreed-upon standards.

Most academic researchers are motivated by a desire to comprehend how intricate nature is (and publish their work in high-impact journals). Although vital, standardization and increased manufacturing have less scholarly appeal than the discovery of a novel product. The possibility of financial support, business collaborations, and academic reputation are a few potential incentives for conducting this kind of study [11].

5.3 iGEM's Issues

5.3.1 iGEM: Is It for Elites or Commons?

As synthetic biology is currently at the frontier of biology research, it is undoubtedly costly, requiring a high-level wet-lab platform and advanced computation platform as well as its current form. Therefore, it would be natural to ask the question: Is iGEM for elites or commons? This is a question hard to ask in front of the students, but it is unfortunately a question that needs to be asked. To answer this question, I think some analogy should be conducted first as below.

We can learn from other international student competitions, such as RoboCup (for robotics) [12] and CASP (for protein folding prediction) [13]. RoboCup started in 1997 and has become popular for robotics, artificial intelligence, and other research areas. It has been in partnership with industrialized research giants such as MathWorks and has gathered more than 5000 attendants every year, mostly students. It has become very international as well, with its 19th competition held in China this year with more than 2000 competitors from 47 countries. With its RoboCup@Home portal and other portals, it has constantly attracted students and researchers interested in robotics and artificial intelligence, and hundreds of software packages and research papers have been produced along its history. CASP is more about solving real scientific hardcore problems. Since it started in 1994, its participants have proposed many novel methods that actually boost the advancement of protein folding prediction. The common properties of these two successful competitions include computing-centric competition being easy for students to participate and the results being more accessible both for publication and for the public. iGEM does not possess these two advantages. However, iGEM has unique advantages, including its "real" results as in the registry, as well as its participants' ideas that directly or indirectly influence serious synthetic biology research. Therefore, similar to competition such as RoboCup, which aims for commons, iGEM could improve its broader impact by making a more accessible platform (or even virtual platform) for competitors and, more broadly, students who are interested in synthetic biology.

What can we learn from other synthetic biology initiatives, including BIOFAB [14] and Synberc [15]? BIOFAB refers to International Open Facility Advancing

Biotechnology. It is a set of open facilities that are "designed to produce broadly useful collections of standard biological parts that can be made freely available to both academic and commercial users." Synberc has provided "engineering principles to biology to develop tools that improve how fast—and how well—we can go through the design-test-build cycle." Both of these initiatives aim for uniform platforms that have standardized parts as well as tools and testbeds for the efficient development of synthetic biology systems. A uniform platform could benefit a lot for iGEM as well because all competitors could have a fair-play platform for competition and because it could contain parts in the iGEM registry and greatly improve the "from idea to implementation" speed of many iGEM teams. Such a uniform platform would be costly but would for sure benefit both iGEM and students who joined the competition in the long run. Of course, the development of a common platform would need some elite designs, but a successful platform could be more accessible to common students who only have a basic environment for synthetic biology research.

Synthetic biologists have also been greatly helped by the advancement of machine learning techniques, which have made mining functional genes part of resources [16], as well as computational modeling and simulation of synthetic systems possible [17].

Thus, there is plenty of evidence to believe that powered by both techniques (such as machine learning and deep learning) [16] and new collaboration models such as citizen science [18], iGEM could become the ideal competition and leaning opportunities for commons. However, the expensive computational hardware and accessibility of up-to-date knowledge might be new bottlenecks for commons to keep pace with the latest development of synthetic biology [19].

All in all, based on common students unified in the iGEM competition, who can provide numerous novel ideas, together with iGEM's elite-designed well-organized platform including parts and to-be-developed easily accessible system, I believe iGEM would have a bright future in years to come.

References

1. Fritz, B.R., et al.: Biology by design: from top to bottom and back. J. Biomed. Biotechnol. **2010**, 232016 (2010)
2. Clarke, L., Kitney, R.: Developing synthetic biology for industrial biotechnology applications. Biochem. Soc. Trans. **48**(1), 113–122 (2020)
3. Patra, P., et al.: Recent advances in systems and synthetic biology approaches for developing novel cell-factories in non-conventional yeasts. Biotechnol. Adv. **47**, 107695 (2021)
4. Wang, S., et al.: Fluorescence imaging of pathophysiological microenvironments. Chem. Soc. Rev. **50**(16), 8887–8902 (2021)
5. Ashkarran, A.A., Mahmoudi, M.: Magnetic levitation Systems for Disease Diagnostics. Trends Biotechnol. **39**(3), 311–321 (2021)
6. Hylock, R.H., Zeng, X.: A blockchain framework for patient-centered health records and exchange (HealthChain): evaluation and proof-of-concept study. J. Med. Internet Res. **21**(8), e13592 (2019)

7. Frazar, S.L., et al.: Defining the synthetic biology supply chain. Health Secur. **15**(4), 392–400 (2017)
8. Murch, R.S., et al.: Cyberbiosecurity: an emerging new discipline to help safeguard the bioeconomy. Front. Bioeng. Biotechnol. **6**, 39 (2018)
9. Fletcher, A., et al.: Realist complex intervention science: applying realist principles across all phases of the Medical Research Council framework for developing and evaluating complex interventions. Evaluation. **22**(3), 286–303 (2016)
10. Castillo-Hair, S.M., Seelig, G.: Machine learning for designing next-generation mRNA therapeutics. Acc. Chem. Res. **55**(1), 24–34 (2022)
11. Minor, P.: International reference preparations for standardization of biological medicinal products. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. **57**(10), 1145–1151 (2014)
12. Nguyen, Q.D., Prokopenko, M.: Structure-preserving imitation learning with delayed reward: an evaluation within the RoboCup soccer 2D simulation environment. Front Robot AI. **7**, 123 (2020)
13. Herminghaus, A., Picker, O.: Colon ascendens stent peritonitis (CASP). Methods Mol. Biol. **2321**, 9–15 (2021)
14. Saviranta, P., et al.: In vitro enzymatic biotinylation of recombinant fab fragments through a peptide acceptor tail. Bioconjug. Chem. **9**(6), 725–735 (1998)
15. Kunjapur, A.M., Tarasova, Y., Prather, K.L.: Synthesis and accumulation of aromatic aldehydes in an engineered strain of Escherichia coli. J. Am. Chem. Soc. **136**(33), 11644–11654 (2014)
16. Carbonell, P., Radivojevic, T., Garcia Martin, H.: Opportunities at the intersection of synthetic biology, machine learning, and automation. ACS Synth. Biol. **8**(7), 1474–1477 (2019)
17. Mienda, B.S., Drager, A.: Genome-scale metabolic modeling of Escherichia coli and its chassis design for synthetic biology applications. Methods Mol. Biol. **2189**, 217–229 (2021)
18. Koepnick, B., et al.: De novo protein design by citizen scientists. Nature. **570**(7761), 390–394 (2019)
19. Goni-Moreno, A., Nikel, P.I.: High-performance biocomputing in synthetic biology-integrated transcriptional and metabolic circuits. Front. Bioeng. Biotechnol. **7**, 40 (2019)

Chapter 6
Synthetic Biology: Safety Issues

Xue Zhu, Dan Zhao, and Kang Ning

Abstract Safety issues are the key issues for the healthy development of synthetic biology. The generation of novel life forms, biosafety, and biosecurity are the three main ethical issues of synthetic biology. Regulation of novel inventions, management of novel invention patents, benefit sharing, and research integrity are a few additional ethical concerns that were raised. Because biosynthetic technologies have the potential to be used maliciously against humans or the environment, there are certain moral dilemmas concerning biosecurity. Because synthetic biology poses ethical and biosecurity issues, humanity must consider and create preparations for how to handle potentially hazardous products and what ethical measures might be employed to deter hostile biosynthetic technologies. In this chapter, we will look at some of the safety issues facing synthetic biology and some of the existing safety control methods.

Keywords Bio-safety · Safety control

Safety issues are the key issues for the healthy development of synthetic biology.

6.1 Biosafety Issues

The creation of new life and the modification of existing life has raised ethical concerns in synthetic biology, which is actively discussed [1]. The generation of novel life forms, biosafety, and biosecurity are the three main ethical issues of synthetic biology [2]. Regulation of novel inventions, management of novel invention patents, benefit sharing, and research integrity are a few additional ethical concerns that were raised. Recombinant DNA and genetically modified organism

X. Zhu (✉) · D. Zhao · K. Ning
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: zhuxue@hust.edu.cn

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8_6

(GMO) technologies have ethical concerns, and many governments have strict laws governing pathogen research and genetic engineering. Amy Gutmann, the former chair of the Presidential Bioethics Commission, emphasized that we should resist the urge to overregulate genetic engineering, in particular, and synthetic biology, in general [3]. Gutmann asserts this as: "It is especially important for emerging technologies, where there is a strong desire to stifle innovation due to uncertainty and a fear of the unknown. Legislative and regulatory restrictions may work against security and safety by preventing the development of efficient safeguards and preventing the sharing of novel benefits" [4].

6.1.1 The "Creation" of Life

The morality of creating new life forms frequently referred to as "playing God", which is one ethical issue. The development of new life forms that are not found in nature is currently done on a modest scale, the potential advantages and risks are still unclear, and the majority of experiments are conducted with careful consideration and oversight [2]. Many proponents emphasize the enormous potential utility of generating artificial life forms for various industries, including agriculture, medicine, and academic research. Science could learn much more through the creation of new entities than it does now through the study of natural happenings. However, there is concern that the existence of artificial life forms could diminish the "purity" of nature (i.e., nature might be compromised by human intervention and manipulation) and possibly lead to the adoption of more engineering-like principles rather than biodiversity- and nature-focused ideals. Some people are also worried that releasing artificial life into the wild could reduce biodiversity by outcompeting natural species for resources (e.g., algal blooms can wipe out marine species). Another issue is how freshly produced beings should be treated morally if they happen to have self-awareness, sentience, and pain sensitivity. Do such lives need moral or legal protections? How, if so?

6.1.2 Biosafety

In regard to biosafety measures, what is the most ethically acceptable option? Can synthetic life be prevented from being accidentally introduced into the natural environment? These questions have received much ethical and critical attention. In addition to biological containment, biosafety also refers to measures taken to ensure that the public is protected against potentially harmful biological agents. While such concerns remain unaddressed, not all synthetic biology products pose a risk to biological safety or environmental harm. Most synthetic technologies have been argued to be "harmless" and incapable of thriving in the real world because of their "unnatural" features, which have yet to be shown in the wild.

Standard risk assessment methods and regulations for traditional genetically modified organisms (GMOs) are generally considered adequate for synthesized species [5]. Biosafety cabinets and gloveboxes, as well as personal protective equipment, are examples of "extrinsic" biocontainment methods used in laboratories. Pollen barriers and isolation distances are common in agricultural biocontainment, and they are related to GMO biocontainment techniques [6]. By using "intrinsic" biocontainment technologies, synthetic organisms can be made to limit their proliferation in uncontrolled environments or inhibit horizontal gene transfer to natural species, which can help reduce hazards [7]. These include auxotrophy, biological death switches, the inability to reproduce or pass modified or synthetic genes to offspring, and the usage of xenobiological creatures employing alternative biochemistry, for example, using artificial Xeno nucleic acids (XNA) instead of DNA [8]. A key amino acid for all life, histidine, can be manufactured out of bacteria and yeast using auxotrophy. Because these organisms can only be cultivated in a laboratory environment, there are no worries that they could spread to undesired locations.

6.1.3 Biosecurity

Because biosynthetic technologies have the potential to be used maliciously against humans or the environment [9], there are certain moral dilemmas concerning biosecurity. Because synthetic biology poses ethical and biosecurity issues, humanity must consider and create preparations for how to handle potentially hazardous products and what ethical measures might be employed to deter hostile biosynthetic technologies [10]. Extensive regulations of genetic engineering and pathogen research are already in place in many jurisdictions, with the exception of those governing synthetic biology and biotechnology companies. The issues were raised during the earlier discussions of recombinant DNA and genetically modified organisms (GMOs) [11].

6.2 Design Principles for Safety Control

6.2.1 Potential Risks of Synthetic Biology

Synthetic biology, as previously mentioned, opens up new avenues for the modification or creation of living entities. Despite this, synthetic biology faces the "dual-use issue" of technologies, which asserts that technology can be used for both good and ill purposes. While it is impossible to prevent synthetic biology from being abused, the risks can be reduced via education about the hazards and the implementation of proper ethical and regulatory processes. In the sections that follow, we

examine the dangers from the viewpoints of biosafety, biosecurity, and ethical concerns [12].

6.2.2 Biosafety Concerns

Biosafety issues are associated with dual-use biotechnology [13], and not enough effort has been made to detect or analyze biosafety hazards in the synthetic biology sector. The intricacy of synthetic biology makes it difficult to perform a risk assessment based on a comparison of alternatives. When comparing genetically modified organisms, it is straightforward to discover a donor organism that shares the same genes as the intended recipient [12]. A gene with an unknown function or a new pathway involving numerous genes is common in synthetic biology designs, which are typically more complex and require more time and resources to implement. Additionally, xenobiology is a branch of synthetic biology that deals with the creation of life through the use of noncanonical base pairs or amino acids. These components are not found in nature; hence, there is no natural comparison for these situations [14].

Although it has been asserted that biological ecosystems are self-sustaining and that synthetic organisms are vulnerable to extinction from encroachment by native species, biosafety is a major concern in synthetic biology due to accidental or intentional releases of synthetic organisms into the environment during research and application [15]. A purposeful release of genetically altered microbes for plant growth improvement or bioremediation has been financed by the European Union in recent years. Genetically created bacteria and naturally occurring microbes had approximately the same environmental impact, according to these research authors. Because they can be quickly destroyed by competitors or predators and are subject to strong ecological constraints, synthetic microbes may have a short-term advantage in a population, but long-term survival will be difficult. This idea is consistent with the fact that most attempts to genetically engineer microbes for environmental applications have failed [16].

Synthetic biology's ability to go from the laboratory to the real world is hindered by horizontal gene transfer, which is a widespread occurrence in the natural world. Due to the natural lysis of microbes, up to 1 g of nucleic acids are found in a gram of soil and 80 g per liter in seawater [17]. Extracellular DNA can remain in the environment for months before it is ingested by either prokaryotic or eukaryotic cells in the natural environment. Contrary to popular belief, which holds that the natural rate of mutation in bacteria is approximately 1108, synthetic DNA circuits based on mobilized genes or sequences during conjugation or transduction can achieve a significantly higher rate of horizontal gene transfer, with high risks to environmental genetic structure.

Third, bioterrorism is concerned with the rise in antibiotic-resistant superbugs. Antibiotic resistance genes are commonly used as selection markers in plasmids used for DNA or pathway synthesis. For example, superbugs can escape host cells

and make their way outside because of the lack of selection pressure. Antibiotic-resistant "superbugs" can be created in nature when these self-replicating organisms infect and multiply with other bacteria [18].

6.2.3 Ethical Concerns

Synthia, a human-made cell, sparked a worldwide discussion on the ethics of synthetic biology in 2010. As a result, some critics said that the work had the potential to undermine people's fundamental ideas about life and to produce both environmental and health disasters [18]. Ethics in Synthetic Biology and Emerging Technologies, a report published in 2010, examined this issue. Previous researchers have reported that rather than attempting to create life solely from inorganic chemicals, current research relies on an already-existing natural host. Even in the near future, fully artificial life will still be a long way off [19]. As mentioned in the papers, there are five ethical criteria to follow when developing synthetic biology: public benefit, responsible stewardship, freedom, accountability for one's ideas, democratic discussion, and justice and equity. The reports cover all of these ethical considerations in detail. It was suggested that these foundations be followed to ensure that advances in synthetic biology promote human health and the general public good while also identifying and reducing any risks as synthetic biology progressed [20].

Scientists in the field announced in June 2016 that they would form the Human Genome Project-Write consortium to create the synthetic biology technologies needed to synthesize a human genome chemically [21]. Transplantation of human organs and the creation of cancer-resistant therapeutic cell lines are only a few possible uses for this new technology if it is fully developed [16]. In the wake of the revelation, there were considerable arguments over the ethics of cutting-edge biological research in the public sphere once again.

Several moral objections were raised: Is DNA insertion in human embryos a part of the research? Regulators face a dilemma in regard to ensuring fairness, given the hefty expense of the equipment. Is technology going to be a privilege for the wealthy? An increase in public apprehension about technology is possible if the HGP-Write project's results are abused. Unanticipated ramifications of the HGP-Write initiative, such as prenatal genetic testing and selective abortion, have been raised in numerous nations. Last but not least, although scientists declare that the initiative is nonprofit, corporate investment may be engaged; will the successes of this research be monopolized by a few large businesses and exclusively in wealthy countries? Many people were confused when the scientists later clarified that the project's goals were not to create human beings or to usher in an age of eugenics but rather to improve the capability for large-scale DNA synthesis that can be applied to industrial biotechnology or agriculture. For this reason, a working group on ethical, social, and legal aspects was established during the federation's second annual

meeting in May 2017 to facilitate an open dialogue and ensure that the project's ethical boundaries are respected.

6.3 Code of Action for Safety Control

6.3.1 Code of Conduct for Scientists

Scientists need to serve as the first line of defense against the exploitation and abuse of this cutting-edge weapon of science and medicine at the forefront of synthetic biology. Self-discipline and scientists' obligations are two areas where the synthetic biology community and governments agree [22]. Protective measures, such as preventing the dissemination of results that could be exploited for illicit purposes, should be in place at all times during the study process. According to Kuhlau et al., life scientists working on dual-use research should be concerned not only with preventing misuse but also with preventing harm that can be predicted in the future [23].

Researchers must also adhere to regulations in various nations. Some examples are the Responsible Conduct of Research Code of Australia, the revised Ethics Code of the Japan Science Council, and the Self-discipline of Moral Behavior for Scientific and Technical Employees of China, to name just a few [24]. The Chinese and Pakistani delegations jointly submitted a "Model Code of Conduct for Biological Scientists" in 2016 to the BWC's Eighth Review Conference. All researchers can benefit from the "Model Code of Conduct for Biological Scientists," which includes several rules and principles. As a group, synthetic biologists have an obligation, and these recommendations fall within this category. To ensure that research benefits the maximum number of people while limits any harm that may be done, life scientists must carefully analyze the ethical and moral concerns of biotechnology. Researchers in the field of life sciences are also required to conduct a thorough risk assessment and feasibility certification of the potential health and societal risks that biological research procedures and achievements may create. Finally, specialists in the field of life sciences must undertake a thorough risk assessment and feasibility certification of prevention measures [25].

6.3.2 Nationwide Governance

In addition to ethical considerations, consideration should be given to the formation of national governance or regulations. Concerns on synthetic biology biosafety and biosecurity date back to 1999. According to the reports from Cho et al., national and international governments should seriously consider monitoring and regulating information related to the development of biological weapons. Scientists and the scientific community should have more say in how synthetic biology research with

dual-use applications is conducted. Miller and Selgelid proposed a middle ground between the two extremes: there should be a combination of government and institutional controls or an independent authority-based governance system [26].

Restrictions on synthetic biology are nearly identical in both the EU and the United States. To examine the potential challenges associated with risk assessment in synthetic biotechnology, the French High Council for Biotechnology, Germany's Central Committee on Biological Safety, the Netherlands Commission on Genetic Modification, and Belgium's Scientific Institute of Public Health convened an international scientific workshop in 2012. According to this study, synthetically manufactured bacteria or entities are unlikely to provide new environmental risks in the medium term because they are difficult to imagine from existing species. These processes fall under Directives 2009/41/EC and 2001/18/EC, which prohibit the use or release of genetically modified organisms into our environment, respectively, according to the researchers' findings. Synthetic biology, according to the US National Research Council, is not a specific genetic engineering method and does not represent any distinct dangers in comparison to other genetic alteration technologies [25].

According to the International Scientific Workshop held by the European Union, synthetic biology can result in organisms that are fundamentally different from those found in nature [27]. However, the chemically modified products of xenobiology may fall under a new regulatory framework. The term "synthetic nucleic acid molecules" was added to the NIH's "NIH Guideline for Research Involving Recombinant or Synthetic Nucleic Acid Molecules" in 2013. Regardless of whether the DNA molecules are synthesized or recombined conventionally through genetic manipulation, the new guideline emphasizes that the requisite biocontainment is needed. A notable aspect of this decision-making is that it is based on synthetic biology progress that has yet to be fully understood. The J. Craig Venter Institute recently completed research entitled "Synthetic biology and the U.S. Biotechnology Regulatory System: Challenges and Options," which identified two main challenges for the existing US regulatory system in the long run: (I) the number of genetically altered species that are not subject to evaluation by the Animal and Plant Health Inspection Service will rise as a result of synthetic biology. It is currently up to the Animal and Plant Health Inspection Service (APHIS) to determine whether plant pests are used in the design of the plant, but synthetic biology provides new solutions for genetically modified organisms. EPA biotechnology initiatives may be overwhelmed by these entities. (II) As these genetically altered organisms become more complex, risk assessment would be more difficult. When dealing with modified organisms, these agencies have little or no jurisdiction.

6.3.3 Efforts by International Societies

Concerns regarding the use of synthetic biology for bioterrorism or as a weapon have captivated the entire world. According to a report from the 8th BWC Congress in

2016, synthetic biotechnology has increased the scope and destructive capacity of biological weapons. Synthetic biology materials, including rubber and metal parts that corrode more quickly, fuel or food degrades more rapidly, or are destroyed by synthetic life, are examples of "material damage factors." For both civilian and military purposes, this could have serious implications. The National Academies of Science, Engineering, and Medicine in the United States published a paper in 2018 titled "Biodefense in the age of synthetic biology." Synthetic biology has created a growing number of defense issues, necessitating new strategies for chemical and biological defense, as well as new methodologies for combining synthetic biology's new capabilities. An evaluation mechanism for synthetic biology capabilities was also devised. Resurrecting known lethal viruses, making existing bacteria more dangerous, and generating harmful biochemicals in situ are three of the most important biosecurity concerns.

References

1. Persson, E.: Synthetic life and the value of life. Front. Bioeng. Biotechnol. **9**, 701942 (2021)
2. Sture, J., Whitby, S., Perkins, D.: Biosafety, biosecurity and internationally mandated regulatory regimes: compliance mechanisms for education and global health security. Med. Confl. Surviv. **29**(4), 289–321 (2013)
3. Holm, S., Powell, R.: Organism, machine, artifact: the conceptual and normative challenges of synthetic biology. Stud. Hist. Phil. Biol. Biomed. Sci. **44**(4 Pt B), 627–631 (2013)
4. Gutmann, A.: The ethics of synthetic biology: guiding principles for emerging technologies. Hast. Cent. Rep. **41**(4), 17–22 (2011)
5. Devos, Y., Reheul, D., De Schrijver, A.: The co-existence between transgenic and non-transgenic maize in the European Union: a focus on pollen flow and cross-fertilization. Environ. Biosaf. Res. **4**(2), 71–87 (2005)
6. Manoel, R.O., et al.: Landscape barriers to pollen and seed flow in the dioecious tropical tree Astronium fraxinifolium in Brazilian savannah. PLoS One. **16**(8), e0255275 (2021)
7. Cai, Y., et al.: Intrinsic biocontainment: multiplex genome safeguards combine transcriptional and recombinational control of essential yeast genes. Proc. Natl. Acad. Sci. U. S. A. **112**(6), 1803–1808 (2015)
8. Chen, Z., et al.: Artificial intelligence in aptamer-target binding prediction. Int. J. Mol. Sci. **22**, 7 (2021)
9. Bakanidze, L., Imnadze, P., Perkins, D.: Biosafety and biosecurity as essential pillars of international health security and cross-cutting elements of biological nonproliferation. BMC Public Health. **10**(Suppl 1), S12 (2010)
10. in Soliciting Stakeholder Input for a Revision of Biosafety in Microbiological and Biomedical Laboratories (BMBL): Proceedings of a Workshop. 2016: Washington (DC)
11. Jose, J., Pai, S.: Comparison of regulatory framework of clinical trial with genetically modified organism-containing vaccines in the Europe, Australia, and Switzerland. Clin. Exp. Vaccine Res. **10**(2), 93–105 (2021)
12. Li, J., et al.: Advances in synthetic biology and biosafety governance. Front. Bioeng. Biotechnol. **9**, 598087 (2021)
13. Cook-Deegan, R.M., et al.: Issues in biosecurity and biosafety. Science. **308**(5730), 1867–1868 (2005)
14. Peintner, L., et al.: Eight years of collaboration on biosafety and biosecurity issues between Kazakhstan and Germany as part of the German Biosecurity Programme and the G7 Global

Partnership against the spread of weapons and materials of mass destruction. Front. Public Health. **9**, 649393 (2021)
15. Kelle, A.: Synthetic biology and biosecurity. From low levels of awareness to a comprehensive strategy. EMBO Rep. **10**(Suppl 1), S23–S27 (2009)
16. Kang, M., et al.: Synthetic biology approaches in the development of engineered therapeutic microbes. Int. J. Mol. Sci. **21**, 22 (2020)
17. Chen, G.Q., Jiang, X.R., Guo, Y.: Synthetic biology of microbes synthesizing polyhydroxyalkanoates (PHA). Synth. Syst. Biotechnol. **1**(4), 236–242 (2016)
18. Claesen, J., Fischbach, M.A.: Synthetic microbes as drug delivery systems. ACS Synth. Biol. **4**(4), 358–364 (2015)
19. Jochems, C.E., et al.: The use of fetal bovine serum: ethical or scientific problem? Altern. Lab. Anim. **30**(2), 219–227 (2002)
20. Rager-Zisman, B.: Ethical and regulatory challenges posed by synthetic biology. Perspect. Biol. Med. **55**(4), 590–607 (2012)
21. Boeke, J.D., et al.: GENOME ENGINEERING. The Genome Project-Write. Science. **353**(6295), 126–127 (2016)
22. Ehni, H.J.: Dual use and the ethical responsibility of scientists. Arch. Immunol. Ther. Exp. **56**(3), 147–152 (2008)
23. Kuhlau, F., et al.: Taking due care: moral obligations in dual use research. Bioethics. **22**(9), 477–487 (2008)
24. Cho, M.K., et al.: Policy forum: genetics. Ethical considerations in synthesizing a minimal genome. Science. **286**(5447), 2087 (1999) 2089-90
25. Miller, S., Selgelid, M.J.: Ethical and philosophical consideration of the dual-use dilemma in the biological sciences. Sci. Eng. Ethics. **13**(4), 523–580 (2007)
26. Buhk, H.J.: Synthetic biology and its regulation in the European Union. New Biotechnol. **31**(6), 528–531 (2014)
27. in Safety of Genetically Engineered Foods: Approaches to Assessing Unintended Health Effects. 2004: Washington (DC)

Chapter 7
Synthetic Biology: Data Resources, Web Services, and Visualizations

Yuzhu Zhang and Yi Zhan

Abstract Synthetic biology is a study field concerning new biological components, devices, and systems that replicate existing natural biological systems or create a new "artificial life." It is of great significance to study the origin of life and biological evolution and has shown broad application prospects in medicine, energy, environmental protection, and agriculture. Data resources are very important in the study of synthetic biology, and we will provide an overview of the data sources, web services, and visualization tools used recently in the field of synthetic biology in this chapter.

Keywords Synthetic biology · Data resources · Web services · Visualizations · iGEM registry

Synthetic biology is a study field concerning new biological components, devices, and systems that replicate existing natural biological systems or create a new "artificial life." It is of great significance to study the origin of life and biological evolution and has shown broad application prospects in medicine, energy, environmental protection, and agriculture [1]. Synthetic biology aims to design and build engineering systems such as gene circuits, signal cascades, and metabolic networks, among other things. It allows the system to control compound synthesis, manufacture materials, generate energy, improve human health and living conditions, and obtain new cell behavior in a predictable and reliable manner in response to changes in external information [2]. Producing a variety of bulk and fine chemicals from previously unimaginable renewable resources can also be achieved thanks to synthetic biology. However, creating the necessary microbial cell factories is still a protracted, labor-intensive project with unknown results. The absence of dependable, distinctive, and standardized biological components needed for predictable strain engineering presents a significant challenge in this regard. Functional DNA fragments have been standardized so that they can easily connect to other fragments.

Y. Zhang · Y. Zhan (✉)
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: zhanyi@hust.edu.cn

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8_7

Table 7.1 Representative databases

Database	Type	Description	License
parts.igem.org	BioBricks	It was developed at MIT and is now maintained by the biobricks.org organization as the primary parts library.	Open source, noncommercial
biofab.synberc.orgregistry.synberc.org	BioBricks	The International Open Facility Advancing Biotechnology was established in December 2009 as the world's first biological design-build facility (BIOFAB). The purpose of BioFab is to develop broadly applicable collections of common biological components that may be freely distributed to academic and industrial users.	Open source
biobricks.org	BioBricks	The BioBrick™ Public Agreement (BPA) is a freely used legal instrument that allows individuals, corporations, and organizations to share their standardized biological components with others.	Open source, special license
acs-registry.jbei.org	BioBricks	BioBricks has a focus on biofuels and has revealed around 200 open-source sequences.	Open source, noncommercial
NCBI/GenBank	DNA/Bio Security	GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.	
EBI/EMBL	Everything	The EBI is a scalable text search engine that provides easy and uniform access to the biological resources hosted at the EMBL-EBI.	
DDBJ	Bio Security	DDBJ Center gathers nucleotide sequence data and makes it publicly available. It also runs a supercomputer system, to promote life science research.	
Biocatalogue	Web Services	Providing a curated catalog of life science Web services.	
Swiss-Prot/UniProt	Annotated Protein database	Swiss-Prot/UniProt is a freely accessible database of protein sequences and functional information. Many entries were derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from research literature.	

The biological factors used to forge new systems are known as "BioBrick" to build biological systems within living cells, allowing for better combinations of components (Table 7.1) [3].

Biologists will focus on standardized experimental techniques and modular construction of all different kinds of biological components. To understand how a biological system is put together, they will coordinate the control of various components, figuring out how to manage the digital standardization of biology,

quantitative evaluation, and better simulation calculations. This is where synthetic biology's database application really shines. Along with the rapid advancement of contemporary molecular biology and biological engineering, "BioBrick" is in the form of mass. Modern database technology is unavoidably required to organize, store, and manipulate these fundamental biological components of information. The quick evolving of synthetic biology causes increases of the number of internal structures and its greater complexity. Functions are also becoming more sophisticated, including gathering data from commonly used plasmid expression vectors about "biological components," deconstructing these vectors into their component parts, gathering, arranging, and storing the data before using those parts to construct new vectors or plasmids [4].

Another reason for the existence of databases is to ensure data sharing, which is essential to speed up progress, reduce duplication of effort, and allow the reuse of past research. However, efforts to create openness by expanding access to the characteristic parts of synthetic biology are patchy and hard to succeed. Standardization of data sharing is the final step in ensuring the enormous amounts of data generated on synthetic biology components and devices do not end up disconnected, remain inadequate, or even go unutilized. In this regard, scientists need more systematic preservation of complete information about these parts and data. In order to achieve this, data repositories and registries of various components and equipment have been created and routinely maintained [5].

This chapter provides an overview of the data sources, web services, and visualization tools used recently in the field of synthetic biology. EBML-EBI, a nucleotide database maintained by the European Bioinformatics Research Institute (EBI), and DDBJ, a nucleic acid sequence database maintained by the Genetic Information Center of the National Institute of Genetics, Japan, are among the data platforms mentioned above. GenBank, which contains all known nucleic acid and protein sequence and their related literature and biological annotations, is another. These three databases are the leading collections of nucleic acid sequences; they accept user-submitted sequences, store sequences from all living things, and regularly exchange updated data to maintain optimal synchronization between them. International Genetically Engineered Machine(iGEM) Registry and BioBricks Foundation are two expanding communities of biological components committed to assisting in the composition and matching of parts to build synthetic biological devices and systems. The European Institute for Bioinformatics maintains the reviewed and annotated protein sequence database known as Swiss-Prot. In order to create standardized DNA sequence elements, OpenWetWaich encourages the exchange of knowledge, expertise, and wisdom among researchers and communities engaged in biology and bioengineering. Biocatalogue, a general directory of life science Web services, and acs-Registry, an open-source software and platform for managing biological information, are two components of BioFab, which supports the advancement of synthetic biology by better controlling gene expression: to aid researchers in their study of synthetic biology component design and to find fresh design concepts [6].

7.1 iGEM Registry

7.1.1 iGEM Registry Overview

iGEM Registry has over 20,000 documented parts. Several of these components are arranged in its catalog by factors like type, function, chassis, standard, contributor, etc.

7.1.2 Catalog

In the Catalog interface, iGEM Registry provides users with a variety of ways to query components from the iGEM Part library, both through specific loci represented by each part (e.g., Promoter, ribosome-binding site, coding sequence, terminator). They can also be classified according to the particular function of a part. Additionally, iGEM collects parts with unique functions, such as those related to CRISPR, biological repair, drug transport, genes encoding reporter proteins, etc. The parts in the collection are constantly updated. In order to achieve the continual enhancement of the collection library, iGEM annually collects components with defined functions from participating teams and rewards the "Best Collection" award. The Catalog also categorizes elements based on suitable biological chassis, allowing researchers to combine components according to their needs.

7.1.3 Repository

Because the same functional component may perform entirely different biological functions in various cell lines or plasmids, it is necessary to strictly define the functional annotation to preserve the database of plasmid DNA fragments. As a result, if a new part is created in lab, iGEM headquarter requires each year's teams to complete a registration form. Participants must fill out the parts' name, type, short description (typically a biological or technical abbreviation for a part's function), detailed description (detailed specifications for the part's function and use), origin, design considerations (such as removing restriction site mutations, codon optimization, etc.), and add the part's sequence and properties to the registry. Verifying every part complies with the assembly requirements before setting parameters is important in determining which category partition the part will be assigned in the future. Following the completion of the registration form, DNA samples will be sent following shipping specifications, and the "Best Part" will be chosen from samples sent by various teams each year. Due to the global spread of COVID-19, iGEM will no longer accept new samples from 2019. However, once a new component is created, its information must still be added to the iGEM library and adhere to the

competition's rules and specifications. Only parts that have been submitted and meet the criteria can be approved and added to the iGEM database.

7.1.4 Assembly

During experiments, it is often necessary to interconnect multiple part samples to form plasmids capable of performing more complex tasks, so strict standardization of assembly between parts is required to allow the creation of longer, more complex parts. In iGEM assembly requirements, specific prefixes and suffixes need to be used at the beginning and end of a part, respectively. The prefix and suffix contain specific restriction enzyme-cutting sites, which can produce target sample when treated with corresponding restriction enzyme. Meanwhile, DNA ligase can reassemble the sample fragment with other samples to form a new part. Moreover, the new part is formed with the same prefix and suffix as the original ones that can be used for future cut and join. Currently, iGEM uses two sets of standard assembly sets: BioBrick RFC [7] and iGEM Type IIS.

7.2 Literature Resources

7.2.1 PubMed Database (an Important Source of Literature Download)

The National Center for Biotechnology Information (NCBI), a division of the American Library of Medicine, creates the free PubMed database to retrieve biomedical literature. The retrieval function also makes literature searches simple and quick. Its information is derived from MEDLINE, life science journals and books, and online biomedical literature references. These sources primarily offer biology science paper search, resources, and biomedical consulting. However, some literature in the PubMed database does not directly provide the original documents but rather links to the full text [8].

7.2.2 NCBI Bookshelf

Although NCBI Bookshelf is a free online article database, information can be obtained through text search or links in the Entrez database. In NCBI Bookshelf, articles are also sorted by content. Users browsing the book can read other passages or directly search the book's specific content to find the information they need. Every month, thousands of people access NCBI Bookshelf, which offers free full-text data

on life medicine. It offers free, high-quality textbooks to researchers and students, and is now a leading example of open education [9].

7.2.3 BLAST

NCBI BLAST is a DNA or protein sequence similarity analysis tool developed by NCBI. Homology among sequences was compared by similarity with sequences in GenBank nonredundant database, human Genome database and Human RefSeq database in NCBI, Swiss-Prot, PDB, PIR, and PRF, as well as some other patent database and environmental sample database. At the same time, statistical significance can infer the evolutionary relationships of different sequences and help identify gene family members. Finally, BLAST will sort the results according to the score and sequence similarity and present the comparison results [9].

7.2.4 Blast-Type Selection

Nucleotide BLAST The program is used to compare nucleic acid sequences and nucleic acid sequences. In the program's interface, users can choose the subroutine they need again according to the purpose and source of the sequence [10].

Megablast The program is mainly used to compare very similar sequences, such as intraspecific comparisons, and can be used to compare long sequences thanks to the use of "fuzzy algorithms" to speed up comparisons. Megablast is also the default option in NCBI Genome BLAST [11].

Discontiguous Megablast The program is mainly used to analyze sequences with large differences and low similarity, such as cross-species comparison [12].

Blastn The sequence is compared against each known sequence in the database on a one-to-one basis, so it takes more time, mainly used for searching short queries [11].

Protein BLAST This program is used to compare protein sequences, which is the same as the comparison of nucleic acid sequences. After entering the interface, users can choose the subroutine they need again according to the search time and search method [13].

Blastp Matching protein sequences to sequences in a protein database.

Quick BLASTP As NCBI's nonredundant database becomes larger, the speed of comparison becomes slower. Therefore, NCBI launches Quick BLASTP to realize the acceleration of BLASTP. It adopts the K-mer matching method to speed up the search for similar proteins, but the accuracy is not as good as BLASTP.

Delta-blast The program is mainly used for more sensitive and refined protein similarity search.

Psi-blast Firstly, the sequence with the highest score was obtained through BLASTP comparison to establish the position-specific score matrix (PSSM). Then alignment was conducted again through this PSSM, and the results were continuously iterated until convergence, which was used to identify protein families or similar proteins of distant species.

Phi-blast It limits the pattern of proteins searched, and the final result only contains protein alignments of this pattern.

Blastx Blastx can search protein databases using a translated nucleotide query. The main process includes two steps: firstly, translate the input of the nucleic acid sequence (according to six reading frames, the translation may have six protein); secondly, compare the results with protein sequences in database. Its benefit is to identify potential protein encoded by nucleotide query result.

Tblastn Tblastn is a compare program that puts protein sequences into translated nucleotide databases for comparison. The main process is to translate all nucleic acids in the nucleic acid database into proteins according to different reading frames, and then compare them with the protein sequences queried. It is mainly used to find new coding areas in the database where the sequence has yet to be labeled.

Tblastx Tblastx translates a given nucleic acid sequence into a protein using six reading frameworks and compares it to sequences in the protein database. Tblastx is used to query similar nucleotide sequences according to the coding protein, which can only be used under special circumstances.

7.2.5 Result Analysis

Score It indicates the homology between two sequences. The higher the Score is, the more similar the two sequences are. General results are sorted according to the Score.

E Value The smaller the E value is, the more reliable the result is. The E value is an evaluation of the reliability of Score, indicating the possibility that other sequences are more similar to the sequence we need than Score under random circumstances. When the E value is less than 10^{-5}, it indicates that the two sequences have high homology, not because of calculation error. When the E value is less than 10^{-6}, the homology of the two sequences is very high.

7.3 Tools

7.3.1 ENTREZ Search System

The NCBI's Entrez database is an integrated retrieval system that gives users access to sequence, structure, and reference data all in one place, which incorporates PubMed. The PubMed contains GenBank, RefSeq, and PDB, as well as SwissProt, PIR, PRF, and PDB protein sequence databases, genome databases, human genetics databases, species taxonomy databases, gene expression and microarray data, markers, genetic mapping databases, macromolecule 3D structure databases, etc. Interlinking gives users access to information from all NCBI database systems, including molecular types and molecular biology, biochemistry, and genetics.

EMBL-EBI Database

The main molecular biological database services developed and offered by EBML-EBI (European Molecular Biology Laboratory's European Bioinformatics Institute) are the following: EMBL nucleic acid sequence database, TrEMBL and SWiss-Prot protein sequence databases, storage of biological macromolecules of the 3-d coordinate data structure of macromolecules (EBI-MSD), and radiation hybrid mapping database (RHdb database) [14]. Over 70 molecular biology databases in the field of special databases are prepared and distributed by EBI in addition to these main databases. Three major databases—EMBL, GenBank, and DDBJ—each compiles a portion of the world's reported sequence data and daily updates its sequence data in real time every day [15].

EBI maintains FASTA, BLAST, CLUSTALW, and Smith & Waterman as public sequence database search and analysis tools. GeneQuiz is a highly automated analysis system of protein sequences and predictable protein biomedical functions. DALI is a tool for demonstrating the similarity of 3D protein structures. The most useful system is the SRS system, an index and database information cross-references program that offers a single entry into the molecular biological database, integrated analysis tools, and advanced analytical tools for fragmenting and restructuring information. The data is stored in ASCII text format.

The EMBL nucleotide sequence database is the primary nucleotide data source in Europe. EBML nucleic acid sequence database. It is retrieved through the Sequence Extraction System and is kept up to date by the European Institute of Bioinformatics (EBI) (SRS). The information in this database was gathered by lab technicians, genome sequencing facilities, the European Patent Office, and daily data exchanges with GenBank in the United States and DDBJ in Japan to keep the information current. The nucleic acid annotation formats used by all three are the same (including identifiers, descriptions, citations, characteristic information, sequence length, and other information). Most bioinformatics software can be compatible with this

format, which minimizes data duplication while ensuring database cross-referencing and enables quick and easy exchange of nucleic acid data between databases [16].

TrEBML and Swiss-Prot protein sequence database: The Swiss-Prot database is a collection of annotated protein sequences by biologists and computers [17]. It is maintained by the European Institute of Bioinformatics (EBI) and is now incorporated into the UniProt database. It contains a large amount of information about proteins, and provides up-to-date information on protein sequences for researchers working on genomes and proteomics. Swiss-Prot establishes cross-reference links with 25 databases. Users can directly access related entries in other databases. This extensive and practical database network connection gives Swiss-Prot a central position in the database and data focus function. To reduce data redundancy, UniProt lumps together all the proteins with the same sequence, regardless of whether the proteins are from the same source. At the same time, the UniProt database and more than 30 databases, including nucleic acid database, protein sequence database and protein structure database, were cross-referenced, and all the protein information was annotated in detail. UniProt standardized the protein format to make it similar to the format in EBML nucleic acid sequence database, making it easier to cross-reference [17].

7.3.2 Service

Data Retrieval

Tool ArrayExpress/ChEBI Web Services/ChEmbl Web Services/EBI Search/ENA Browser/Gene Expression Atlas API/MartService/PDBe/PSICQUIC/Rhea/ UniProt.org/WSDbfetch [17–20].

Protein Function Analysis

Tool InterProScan (protein structure and function site database to predict protein function)/HMMER (biological sequence analysis tool based on hidden Markov model)/PFAMScan (search tool for protein family database Pfam)/Phobius (combined transmembrane protein topology) Structure and Signal Peptide Predictor/Pratt (Prediction of conserved structure of protein sequences) /RADAR (Detection and Comparison of repeated fragments in protein sequences) [21].

Sequence Similarity Analysis

Tool FASTA (provided by Smith-Waterman/FASTM (short peptides are placed into protein sequence database for comparison)/BLAST (local similarity comparison

of nucleic acids or proteins)/HMMER (protein homology analysis using hidden Markov model) [22].

Multiple Sequence Alignment

Tool Clustal Omega (multisequence alignment via Seeded Guide Trees and HMM Profile-profile Techniques)/Kalign (fast and accurate multi-sequence alignment), MAFFT (high-speed multisequence alignments)/MUSCLE (compare ClustalW2) or T-coffee alignment program with higher average accuracy and faster speed)/ MView (results of reformatting sequence similarity analysis or multisequence alignment)/T-Coffee (multisequence alignment can be used with multiple alignment) [23].

Double Sequence Alignment

Tool EMBOSS Matcher (local dual sequence alignment using the strict algorithm in Bill Pearson's Lalign Application)/EMBOSS Needle (dual sequence global alignment using Needleman–Wunsch algorithm) EMBOSS Stretcher (optimizes results using classic dynamic programming algorithms for linear space based on EMBOSS Needle)/EMBOSS Water (dual sequence local alignment using Smith–Waterman algorithm) /GeneWise (alignment of protein sequences with DNA sequences that allow intron and frame-shift errors) /LALIGN (non-cross local alignment of protein or nucleotide sequences to detect internal duplications, a tool for suboptimal pairing) [24].

7.3.3 Phylogenetic Analysis

Tool Simple Phylogeny (phylogenetic tree generation based on multiple sequence alignments).

7.3.4 Sequence Format Conversion

Tool EMBOSS seqret (for sequence conversion, supports input in any sequence format) [25].

7.3.5 Sequence Statistics

Tool EMBOSS CPgplot (identify CpG islands in the input nucleotide sequence)/EMBOSS Isochore (identify and plot alleles in the input DNA sequence)/EMBOSS Pepinfo (can mark the characteristics of amino acids in the input protein sequence)/EMBOSS Pepstats (can calculate protein molecular weight, isoelectric point, etc.)/SAPS (use statistical methods to evaluate more protein sequence properties) [25].

7.3.6 Sequence of Translation

Tool EMBOSS Transeq (translate the input nucleic acid sequence into the corresponding peptide chain using six reading boxes)/EMBOSS Sixpack (translate the input nucleic acid sequence into the corresponding peptide chain and output it in the form of visualization)/EMBOSS Backtranseq (predicts the most likely nucleic acid sequence from the incoming protein sequence)/EMBOSS Backtranambig (reads the protein sequence and outputs the nucleic acid sequence it may have come from) [25].

7.4 Databases

7.4.1 GenBank

GenBank Overview

GenBank is a comprehensive database that contains publicly available nucleotide sequences for over 340,000 formally described species. Each record is annotated with coding region (CDS) characteristics and also includes amino acid translations. GenBank belongs to an international collaboration of sequence databases, including EMBL and DDBJ in addition to NCBI [26]. Global coverage is maintained by daily data sharing with the European Nucleotide Archive (ENA) [27] and the DNA Data Bank of Japan (DDBJ). Records from GenBank may be retrieved using the Entrez search engine, which combines this information with additional sources.

7.4.2 DDBJ

DDBJ Overview

The DNA Data Bank of Japan provides a nucleotide sequence archive database and accompanying database tools for sequence submission, entry retrieval, and

annotation analysis [15]. The National Institute of Genetics in Japan manages and updates the DDBJ, a database of Japanese DNA. The DDBJ regularly exchanges data with GenBank and EMBL as a member of INSDC (International Collaboration on Nucleotide Sequence Databases) and updates data in real time. It also offers a range of data analysis services, manages and conducts routine maintenance on Japanese supercomputers, and collects and provides substantial nucleotide sequence data. A significant amount of new information about biomolecules has been produced as genomics research has progressed on the basis of experiments like genome sequencing, protein sequencing, and structure analysis. Because of this, the DDBJ center adheres to the unified INSD rules and has a wealth of data annotation information. It utilizes cutting-edge computer technology to gather, manage, and store raw data for later retrieval and use. The database will then be made available online to everyone in the world to support life science research activities.

Database

The database in DDBJ includes AGD (Japan's genome data sharing database, which can only be accessed after the application is approved) [28] and BioProject (a database for collecting and sharing information of biological research projects, mainly including project materials, protocol, funding sources, time information, and so on) [29]. BioSample (a database for collecting and sharing information on biological samples, providing standardized annotated information on biological samples, provides a wealth of sample attributes) [29], DDBJ, DRA (DDBJ's public archive high-throughput sequencing database), GEA (Functional Genomics Database) [30], JGA (de-identification phenotype database), JVar (Japanese Variation Database) [31], short genetic variation JvAR-SNP and structural variation JvAR-SV, and NHA (a database for storing various data generated from human specimens established in the context of the rise of next-generation sequencing and other analytical data).

Data Retrieval

Data retrieval tools in DDBJ include AOE (accessing the public gene expression database by indexing), ARSA (accessing DDBJ by fast retrieval or advanced retrieval), BLAST (retrieval system based on sequence similarity), CRISPR direct (designing CRISPR for specific nucleotide sequences/Cas target), DBCLS SRA (either by keyword or Accession Number searched SRA, BioProject, BioSample databases), GGGenome (high-speed nucleotide sequence search engine), GGRNA (keyword search for genes and transcripts), RefEx (gene expression database search tool), TXSearch (NCBI classification search) system retrieval tool, TogoAnnotator (gene annotation database retrieval tool, which can provide more annotation content of target genes), TogoVar (Japanese genetic variation database retrieval tool), and Getentry (flat file search system) [15, 29, 32].

Data Analysis

Tool BLAST (retrieval system based on sequence similarity alignment), ClustalW (multisequence alignment program for DNA or proteins), Maser (large reading management and analysis system for next-generation sequencing data NGS), VecScreen (tools that can be input into sequences to identify vectors, connectors, primers, and other foreign DNA), and WABI (Web API for DDBJ) [33].

7.4.3 UniProt

UniProt Overview

UniProt (Universal Protein) is the most comprehensive and well-researched free protein database currently accessible because of the integration of Swiss-Prot, TrEMBL, and PIR-PSD data. As a publicly available database, UniProt has the advantage of easily available, exhaustive functional annotation of protein sequences [17].

Database

UniProt contains UniProtKB, UniParc, Proteomes, and UniRef databases, in which UniParc will accept original data from GenBank, EMBL, DDBJ, and other public databases, and obtain nonredundant protein information resources after processing. This information is then used to provide reliable data sets based on the information types of the different subdatabases UniProtKB, Proteomes, and UniRef. The Swiss-Prot in UniProtKB is a high-quality nonredundant database manually annotated by TrEMBL. It is also one of the protein databases commonly used by researchers [17].

UniParc

UniParc is a comprehensive database of protein sequences. It stores a large number of protein sequence resources and provides a comprehensive history of all proteins. UniParc includes UniProtKB Database, EBML, GenBank, DDBJ, Ensembl, EPO, Model Biology Database (FlyBase), H-Invitational Database (H-INV), IPI, Japan Patent Office (JPO), Protein Information Resource (PIR-PSD), Protein Data Bank (PDB), Protein Research Foundation (PRF), Saccharomyces Genome Database (SGD), The Arabidopsis Information Resource (TAIR), TROME, US Patent Office (USPTO), and Genome and Genome Annotation DatabaseWormBase (VEGA) [34].

All the latest and revised protein sequences from the source are updated daily, cross-referenced, and revised in time, thus ensuring the completeness of the data included. To avoid data redundancy, UniParc combined protein sequences from

different but identical sources into one record and added annotated information such as basic identifiers, source database retrieval numbers, and version numbers to each record.

UniProtKB

Interactive lookups with other databases provide users comprehensive and detailed protein synthesis information about target proteins. The database consists of two parts. One is UniProtKB/Swiss-Prot, which is mainly responsible for sequences annotated by professional biologists and their related literature information and sequences analyzed by the computer. Meanwhile, it also summarizes and integrates all relevant literature on the same protein (including its variants and related diseases). It also lists protein data for cross-reference with other nucleic acid, species-specific, domain, and disease databases. UniProtKB/Swiss-Prot also included proteins translated from specific sequences, such as sequences encoding small fragments, synthetic sequences, T-cell receptor sequences, patented sequences, and highly overexpressed sequences. These opals need to be annotated by a large number of people before being incorporated into the UniProtKB/Swiss-Prot database. The other, UniProtKB/TrEMBL, contains sequences automatically annotated and classified after computer analysis, and proteins translated from sequences from GenBank/EBML/DDBJ/Arabidopsis Information Repository (TAIR)/Yeast Genome Database (SGD)/Human Ensembl database are included [17].

The search area of this page is mainly for researchers to find proteins of interest. In the search box, they can be searched by protein name, sequence, ID number, etc. Users can use the filter in the lower-left corner of this page to restrict the data sources and perform a species classification search to reduce the number of possible results. The sequence is UniProt ID, the protein database name, the protein name, the gene name, the species, and the sequence length are all listed in the search result column. The detailed protein annotation will be shown on the page after clicking UniProt ID, including annotations on protein structure, interactions, splice isomers, related disease information, post-translational modifications, subcellular localization, tissue specificity, developmental stage specificity, domains or key sites of biological significance, and more.

UniRef

In order to include all data completely, without omission or redundancy, UniRef classifies and summarizes all kinds of data from UniProt and selects some data from UniParc, and the clustering information in UniRef will be updated synchronously with the update of UniParc and UniProtKB. The identity of the database is divided into three levels: 100%, 90%, and 50% correspond to UniRef100, UniRef90, and UniRef50 respectively, among which UniRef100 is the most comprehensive database of nonredundant protein sequences. On the basis of UniRef100, UniRef90, and

UniRef50 reduce the amount of data so that sequence similarity search can be carried out at a faster speed and with less error. At the same level, each sequence can only belong to the same cluster, and these sequences can only be contained in the same parent set sequence and subset sequence. The UniRef100 database integrates the same sequence data and subfragment data, and users can search it once through a search portal. The UniRef90 database is obtained by clustering on the basis of UniRef100, while the UniRef50 database is based on UniRef90. Each cluster contains the following information: library origin, protein name, taxonomic information, number of items under the cluster, and so on [35].

Proteomes

A proteome is a collection of proteins expressed by living things. For species whose entire genomes have been sequenced, UniProt offers a proteome. They cover well-researched model organisms and other organisms relevant to phylogeny and biomedical research. The proteome's ID, the source species' ID, and annotated details about the proteins it contains are all available to us.

7.4.4 OpenWetWare

OpenWetWare Overview

Openness in OpenWetWare manifests as a "wiki," which enables anyone with web browser software to edit or create linked web pages as a part of a website that offers LABS, groups, resources, reference materials, and blogs in addition to substantial web links to other online resources. Most of them have something to do with molecular biology or bioengineering.

OpenWetWare is a sizable collaborative resource library that includes links to blogs and other online resources in addition to resources and references from biology LABS and groups all over the world. For biochemistry, molecular biology, and biomedicine laboratories, OpenWetWare offers thorough and accurate records of laboratory activities. Provide an experiment "recipe" or "method" that is standardized, reproducible in the lab, and includes instructions, a list of the raw materials needed for the experiment, and notes and reminders to remind the experimenters what to pay attention to during the experiment and how to solve problems. Researchers can use this website to look up references and perhaps find some improvement inspiration, despite the precise procedures and laboratory equipment varying by country and region.

OpenWetWare is officially open, and all of its content can be edited (to varying degrees), commented on, or added to by any registered user, and every time a change is made to the database, a detailed record of the changes is left for the entire database to manage.

Reference Materials

Computing Both information technology and biotechnology are high and new technologies, which complement each other and jointly promote the rapid development of each other. Programming provides a powerful computing tool for the development of biology. Today, it is hard to separate biotechnology advances from high-performance computing developments. In the future, more and more powerful computers and software will be used to collect, store, analyze, simulate, and publish information. In addition, information technology has enhanced all kinds of database management, information transfer, retrieval, and resource sharing in biotechnology. Therefore, understanding and proficiency in programming technology will be one of the necessary conditions for every biological researcher in the future. OpenWetWare lists the programming languages, web pages, systems, software, protocols, programs, and biological tools that may be used in research in order to enable biologists to learn and master the necessary programming techniques more quickly and effectively. Each page contains tutorials, examples, references, and an application framework for biological research. Researchers can learn how to use these tools by following the links provided on the web page.

Equipment This page lists equipment descriptions from different laboratories. Each device page should contain usage protocols, calibration curves, etc. Publish the problem and the solution simultaneously so that it is easier to troubleshoot the problem if it arises again.

Materials This page provides information on various reagents discussed and published by laboratories. Displays descriptions of medium, antibiotics, enzymes, Reporters, acids and bases, and other chemicals, ingredient descriptions, and experimental dosing with links. This page also provides reagents for SDS gel electrophoresis, agarose gel electrophoresis, and other protein purification methods.

Microscopy As the main means to observe the microscopic world, biological microscopic imaging has emerged in many studies in recent years. This page presents common methods of fixing, staining, sectioning, and cleaning in the use of microscopes in various laboratories, and also gives the constitution, use procedures, and related issues of special microscopes.

Strains It mainly introduced the laboratory commonly used bacteria (*Escherichia coli*, *Bacillus subtilis*, mycobacteria, *Pseudomonas aeruginosa*, *Salmonella typhimurium*, coli), microalgae, fungi, and yeast (*Saccharomyces cerevisiae* and *Candida albicans*), describes some strains (*Escherichia coli*, *Saccharomyces cerevisiae*) genotype, standard strains named and abbreviations, and lists the common strains. In addition, researchers will rely on the Belgian coordinated collections of microorganisms and the collection of strains provided by the laboratory for the experimenter's reference. In the meantime, OpenWetWare is working on building standardized *E. coli* strains, and researchers can see progress and participate in the project on this page.

Statistics The Statistical portal is the hub for everything related to statistical analysis in biological research. This page provides beginners with a primer on important concepts in statistics. It also provides links to pages that discuss individual statistical topics in detail and lists useful software and recommended external pages. Users can help the page grow by contributing content.

References

1. Holm, S.: Organism and artifact: proper functions in Paley organisms. Stud. Hist. Phil. Biol. Biomed. Sci. **44**(4 Pt B), 706–713 (2013)
2. Clarke, L., Kitney, R.: Developing synthetic biology for industrial biotechnology applications. Biochem. Soc. Trans. **48**(1), 113–122 (2020)
3. Matsumura, I.: Methylase-assisted subcloning for high throughput BioBrick assembly. PeerJ. **8**, e9841 (2020)
4. Ho-Shing, O., et al.: Assembly of standardized DNA parts using BioBrick ends in E. coli. Methods Mol. Biol. **852**, 61–76 (2012)
5. Nora, L.C., et al.: The art of vector engineering: towards the construction of next-generation genetic tools. Microb. Biotechnol. **12**(1), 125–147 (2019)
6. Cavalcoli, J.D.: Genomic and proteomic databases: large-scale analysis and integration of data. Trends Cardiovasc. Med. **11**(2), 76–81 (2001)
7. Alnahhas, R.N., et al.: The case for decoupling assembly and submission standards to maintain a more flexible registry of biological parts. J. Biol. Eng. **8**(1), 28 (2014)
8. Motschall, E., Falck-Ytter, Y.: Searching the MEDLINE literature database through PubMed: a short guide. Onkologie. **28**(10), 517–522 (2005)
9. Hoeppner, M.A.: NCBI bookshelf: books and documents in life sciences and health care. Nucleic Acids Res. **41**(Database issue), D1251–D1260 (2013)
10. Cameron, M., Williams, H.E.: Comparing compressed sequences for faster nucleotide BLAST searches. IEEE/ACM Trans. Comput. Biol. Bioinform. **4**(3), 349–364 (2007)
11. Chen, Y., et al.: High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. **43**(16), 7762–7768 (2015)
12. Ibironke, O., et al.: Species-level evaluation of the human respiratory microbiome. Gigascience. **9**, 4 (2020)
13. Singh, H., Raghava, G.P.: BLAST-based structural annotation of protein residues using Protein Data Bank. Biol. Direct. **11**(1), 4 (2016)
14. McEntyre, J., Birney, E.: The EMBL-EBI channel. F1000Res. **5**, 52 (2016)
15. Fukuda, A., et al.: DDBJ update: streamlining submission and access of human data. Nucleic Acids Res. **49**(D1), D71–D75 (2021)
16. Kinney, N., et al.: Ethnically biased microsatellites contribute to differential gene expression and glutathione metabolism in Africans and Europeans. PLoS One. **16**(3), e0249148 (2021)
17. Boutet, E., et al.: UniProtKB/Swiss-Prot. Methods Mol. Biol. **406**, 89–112 (2007)
18. Athar, A., et al.: ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res. **47**(D1), D711–D715 (2019)
19. Capecchi, A., et al.: PubChem and ChEMBL beyond Lipinski. Mol. Inform. **38**(5), e1900016 (2019)
20. in *NTP Research Report on National Toxicology Program Approach to Genomic Dose-Response Modeling: Research Report 5*. 2018: Durham (NC)
21. Jones, P., et al.: InterProScan 5: genome-scale protein function classification. Bioinformatics. **30**(9), 1236–1240 (2014)
22. Ferraro Petrillo, U., et al.: FASTA/Q data compressors for MapReduce-Hadoop genomics: space and time savings made easy. BMC Bioinformatics. **22**(1), 144 (2021)

23. Sievers, F., Higgins, D.G.: Clustal omega. Curr. Protoc. Bioinformatics. **48**, 3 13 1-16 (2014)
24. Greene, E.C.: DNA sequence alignment during homologous recombination. J. Biol. Chem. **291**(22), 11572–11580 (2016)
25. Rice, P., Longden, I., Bleasby, A.: EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. **16**(6), 276–277 (2000)
26. Benson, D.A., et al.: GenBank. Nucleic Acids Res. **46**(D1), D41–D47 (2018)
27. Amid, C., et al.: The European Nucleotide Archive in 2019. Nucleic Acids Res. **48**(D1), D70–D76 (2020)
28. Sun, R., et al.: *AGD: Aneurysm Gene Database.* Database (Oxford), 2018. **2018**
29. Barrett, T., et al.: BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. **40**(Database issue), D57–D63 (2012)
30. Smyth, E.C., et al.: Checkpoint inhibitors for gastroesophageal cancers: dissecting heterogeneity to better understand their role in first-line and adjuvant therapy. Ann. Oncol. **32**(5), 590–599 (2021)
31. Higasa, K., et al.: Human genetic variation database, a reference database of genetic variations in the Japanese population. J. Hum. Genet. **61**(6), 547–553 (2016)
32. Ma, Y., Zhang, L., Huang, X.: Genome modification by CRISPR/Cas9. FEBS J. **281**(23), 5186–5193 (2014)
33. Johnson, M., et al.: NCBI BLAST: a better web interface. Nucleic Acids Res. **36**(Web Server issue), W5-9 (2008)
34. Leinonen, R., et al.: UniProt archive. Bioinformatics. **20**(17), 3236–3237 (2004)
35. Suzek, B.E., et al.: UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. **23**(10), 1282–1288 (2007)

Chapter 8
Synthetic Biology: Case Studies

Pei Liu, Yi Zhan, and Kang Ning

Abstract The international Genetically Engineered Machine (iGEM) is mainly for undergraduate synthetic biology competitions. iGEM represents one of the symbols for the synthetic biology research area, especially in the minds of many students around the world, with an increasing number of teams every year. In China, there are already national projects that support synthetic biology research, and the internal newsletters, which began in 2013, have already covered a wide area of synthetic biology frontiers. Chinese synthetic biologists have already developed several genetic circuits for algae, plant, and even mammalian cells. Teams from China have joined iGEM since 2007, and in 2014, more than 1/4 of iGEM teams came from China. In this chapter, we will introduce several representative projects of student teams participating in iGEM competition.

Keywords The international Genetically Engineered Machine (iGEM) · Representative projects

The international Genetically Engineered Machine (iGEM) was born at the peak of the last frenzy of synthetic biology; thus, it also bears the hope of many synthetic biologists. During its more than 10 years of running, it has produced some novel ideas for synthetic biology research, as well as many potential biological parts that have been stored in the registry.

In China, there are already national projects that support synthetic biology research, and the internal newsletters, which began in 2013, have already covered a wide area of synthetic biology frontiers. Chinese synthetic biologists have already developed several genetic circuits for algae, plant, and even mammalian cells. China's teams have joined iGEM since 2007, and in 2014, more than 1/4 of iGEM teams came from China. During its years of journey in iGEM, China won more than

P. Liu · Y. Zhan · K. Ning (✉)
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: ningkang@hust.edu.cn

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8_8

10 golden and more than 20 silver and copper medals, representing national university students' interests in iGEM.

8.1 iGEM Project by HUST-China 2021

In daily life, many people are keen to perm or dye their hair, but traditional chemical reagents will cause adverse effects on human health. Therefore, the HUST-China team decided to use natural pigments and short peptides, which are safe and harmless materials, to perm and dye their hair. In addition, they designed a fermentation bottle and a dyeing comb, providing users with a convenient way to produce the pigments and dye their hair entirely on their own.

This project Mr. Tony's goal is to utilize synthetic biology methods to generate natural pigments and short peptides and use them in the hairdressing industry, reducing the harm caused by chemical reagents. In our project, they chose Pichia pastoris as our chassis organism and produced three natural pigments: indigo, curcumin, and lycopene. In addition, they produce several peptides to perm the hair. In addition to the health benefits of this method of perming and dyeing hair, there is another advantage, that is, it can quickly recover. In the future, they hope that this product can be used by more people and has broader application prospects [1].

8.2 iGEM Project by HUST-China 2019

According to reliable sources, 410 million people worldwide consume bananas as a basic meal, and 8.8 billion tons of stalks are produced annually. However, many banana stalks are removed each year and are only used as fertilizer in the field the following year. In light of this, we chose to turn banana stalks into fine fibers by biological fermentation. The purpose of this project is to develop a biological technology that is environmentally friendly and green for processing banana stems and producing high-quality banana fiber. The project is divided into three sections: fermentation, a pH-responsive system, and a degradation system for breaking down the lignin and pectin in banana stalks.

Banamax is an engineered *Pichia pastoris* created by HUST-China that responds to environmental pH and adaptively controls the quantity of degrading enzymes. By integrating 3 pH-responsive promoters, 6 various signal peptides, and 3 biological degumming enzymes, high-enzyme activity kits were created. At pH 7 and pH 5, we successfully decomposed pectin and lignin, respectively. To sustain the enzyme activity and act as a pH buffer for the environment, alkali was expressed at pH 2. Trial results demonstrating the successful extraction of crude fiber from banana stalk samples demonstrate the viability of the overall biological intelligent manufacturing system [2].

8.3 iGEM Project by HUST-China 2018

One of the main sources of new energy, photovoltaic power generation, still has several problems, such as high costs, high pollution, and high energy consumption. For a long time, microbial fuel cells (MFCs) have been the subject of intensive investigation. The high energy conversion efficiency, accessibility to temperature and ambient conditions, lack of need for waste gas treatment, and other benefits of MFCs over other modern technologies that use other organic molecules to produce energy are just a few.

Additionally, our project is founded on the background of globalization and proposes a novel notion for the conventional solar industry: to research a device that can convert light into electricity while remaining commercially viable. Our gadgets can perfectly replace batteries compared to solar cells, and they also reduce environmental pollution and the issue of large power plant footprints. By increasing the lactic acid output of *Synechocystis* and *Rhodopseudomonas* pallidum, we boosted the production of *Shewanella*. *Shewanella* production can currently reach 20%, which is sufficient to replace solar panels with 17–18% of the electricity generated on the market. Optopia is intended to function as a photovoltaic system made up of two subsystems: an electrogenic microbe system and a photosynthetic microbe system (*Synechocystis* sp. or *Rhodopseudomonas palustris*) (*Shewanella oneidensis*).

To find an improved version of Optopia, HUST-China 2018 built a *Synechocystis–Shewanella* MFC and a *Rhodopseudomonas–Shewanella* MFC. This maximizes the conversion of optical energy to electric energy [3].

8.4 iGEM Project by HUST-China 2017

Due to the significance that rare-earth elements (REEs) play in several fields—permanent magnets, catalysts, rechargeable batteries, and other high-tech products—they have drawn an increasing amount of public interest. However, if not effectively managed, operations such as mining, refining, and recycling of rare earth elements seriously impact the ecosystem. For instance, the refining process results in the production of hazardous acids, which are specifically designed to enter the common water supply. To detect and effectively capture rare earth elements from industrial sewage, they designed REEBOT, which is an engineered *Escherichia coli*. Rare earth ions can be recycled, and sewage water can be made clean by creating and building genetic circuits. Our ultimate objective is to create a highly clever and effective engineering circuit that uses particular peptides.

HUST-China 2017 developed an engineering *E. coli* named REEBOT, whose system has been divided into a sensing part and a capturing part, to limit the pollution of rare earth elements and recycle those rare earth elements effectively. To recycle rare earth ions, they employ surface display systems and lanthanide-binding

peptides. The PmrA-B-C system, which was found in *Salmonella*, serves as the foundation for the sensing component [4].

8.5 iGEM Project by XMU-China 2020

Tea is deeply rooted in Chinese culture. For a long period, a large amount of glyphosate has been used as an herbicide, which raises a severe problem of pesticide residues in tea food. XMU-China 2020 decided to engineer strains that can rapidly detect and effectively degrade glyphosate.

XMU-China 2020 aims to develop an efficient glyphosate detection and degradation system using synthetic biology technology. It is hoped that this project could provide new ideas for the detection and degradation of pesticide residues [5].

8.6 iGEM Project by TUDelft 2017

According to the World Health Organization (WHO), "The resistance of antibiotics is one of the biggest threats to global health, food security and development currently." The abuse of antibiotics often leads to the emergence of bacterial drug resistance, and drug-resistant bacteria enter the human body through food or direct contact, posing a serious threat to human health. In view of this, the team project of 2017 TUDelft focused on using synthetic biology to solve the problem of antibiotic resistance. The goal of this project is to develop a simple and reliable tool that enables farmers to detect whether the pathogens that cows are infected with are drug-resistant bacteria. Based on the test results, farmers could adjust the use of antibiotics to realize targeted treatment.

TUDelft utilized the characteristics of Cas13a in CRISPR–Cas technology and targeted the target RNA to generate a detectable signal. This system could detect whether cows are infected with resistant bacteria under nonlaboratory conditions and actually solve the problem of antibiotic abuse in agriculture. In addition, TUDelft also created various biological components, including different types of Tardigrade protein (TDP), Cas13a, and a composite part containing a nonsense spacer. Apart from solving the problem of antibiotic resistance, TUDelft created a bond between synthetic biology and problems remaining in agriculture and motivated the public to learn more about synthetic biology [6].

8.7 iGEM Project by TJUSLS China 2016

Discarded plastics are garbage that is difficult to degrade in nature and is currently threatening the earth's environment. Traditional methods to deal with plastic waste require centralized treatment at high temperature, which requires high energy but has low efficiency and could cause secondary pollution. TJUSLS China, a team from the School of Chemical Engineering of Tianjin University, created an efficient plastic biodegradation system based on a mixed bacteria system using bacteria to degrade plastics. In November 2016, this project was nominated for the Gold Award and the Best Environmental Project Individual Award in the 2016 iGEM Competition (iGEM) hosted by the Massachusetts Institute of Technology.

The mixed bacteria system is a bacterial organization based on artificial design, allowing different bacteria to perform their duties. In this system, some of the bacteria degrade the large molecules in the plastic into small molecules, and others absorb or convert the small molecules into other beneficial substances. Through hundreds of mixed bacteria experiments, TJUSLS continued to explore bacterial culture conditions and finally developed a mixed bacteria system that allows different bacteria to coexist. They designed a metabolic pathway in the mixed bacteria system, which reduced the nutrient competition between bacteria and bacteria and consequently realized stability of the system. This mixed bacteria system could completely degrade polyethylene terephthalate (PET), which is a common plastic used in daily life. In the future, people could use this technique to realize the on-site decomposition of plastic waste.

To degrade plastics faster, TJUSLS also applied gene-editing technology to modify the PETase enzyme, the key to degrading plastics, and successfully improved the efficiency of this enzyme. In addition, TJUSLS also used cell-free culture (CELL-FREE) technology to build an enzyme production factory identical to the cells by putting cell substances into test tubes. In this way, the operation steps for screening PETase enzymes were greatly simplified, and the production rate of PETase enzymes was greatly increased.

TJUSLS also used cell-free culture (CELL-FREE) technology to build an enzyme production factory identical to the cells by putting cell substances into test tubes [7].

8.8 iGEM Project by SCAU-China 2016

Traditional iGEM projects mostly use microbes as templates. However, since 2010, there have been many outstanding projects using plants as templates. One of them is the rice endosperm multigene system (TGS II) established by the South China Agricultural University team (SCAU-China), in which astaxanthin (ASTA) was successfully expressed in the endosperm, thereby obtaining rice that was rich in astaxanthin (aSTARice).

Plants are autotrophic eukaryotic and food-related species that have a large number of descendants, and they are considered one of the most promising templates of synthetic biology. However, since genetic manipulation is difficult and the growth cycle is long, the establishment of plant templates is challenging. SCAU-China has set an example in considering plants as templates. In the iGEM competition, many other teams were also committed to overcoming this problem [8].

8.9 iGEM Project by SJTU-Software 2019

Apart from all kinds of iGEM projects mentioned above, there are plenty of projects focusing on software, web science, and databases. For instance, in 2019, SJTU software created an online synthetic biology website "Phosyme" that integrated information on plant biobricks and plant metabolism, containing several computational tools related to plant synthetic biology.

After investigation, they noticed that existing websites of plant synthetic biology are incapable of meeting research needs. In addition, the official website of iGEM has several problems, such as incomplete information and defective search engines, resulting in many restrictions for teams that focus on plant template synthesis.

Based on these reasons, SJTU software has established an online synthetic biology website "Phosyme", which integrates information on plant biobricks and plant metabolism and contains three databases: PartData, PlasmidData, and MetaData. Moreover, three tools are provided:

1. Prediction, which is referred to as training data based on the convolutional neural network CNN model. This function was designed for predicting the reaction probability between the enzyme sequence and substrate during photosynthetic carbon fixation.
2. Molecular structure visualization.
3. Online modification, export and format conversion via SBML (Systems Biology Markup Language).

In summary, "Physyme" provided users with a relatively complete experience by integrating databases and computational tools. It has also been noted that deep learning has a promising future in synthetic biology applications [9].

References

1. *iGEM Project by HUST-China 2021*. Available from: https://2021.igem.org/Team:HUST-China/Team
2. *iGEM project by HUST-China 2019*. Available from: https://2019.igem.org/Team:HUST-China
3. *iGEM project by HUST-China 2018*. Available from: https://2018.igem.org/Team:HUST-China
4. *iGEM project by HUST-China 2017*. Available from: https://2017.igem.org/Team:HUST-China
5. *iGEM project by XMU-China 2020*. Available from: https://2020.igem.org/Team:XMU-China

6. *iGEM project by TUDelft 2017*. Available from: https://2017.igem.org/Team:TUDelft
7. *iGEM project by TJUSLS China 2016*. Available from: https://2016.igem.org/Team:TJUSLS_China
8. *iGEM project by SCAU-China 2016*. Available from: https://2016.igem.org/Team:SCAU-China
9. *iGEM project by SJTU-software 2019*. Available from: https://2019.igem.org/Team:SJTU-software

Chapter 9
Concluding Remarks

Dan Zhao and Kang Ning

Abstract As a basic science, synthetic biology has grown from simply biomedical engineering to an area that stands on both engineering and omics big data, and thus it has become an area with strong problem-driven and data-driven features. As an application science, it has already enabled complex modification of genomes, yielding important metabolites useful in a broad spectrum of applications, based on rational editing of the genomes. In summary, it has undergone the scientific renaissance that could prove it to be a multidiscipline research area, from where we can answer questions on why, what, and how to build a synthetic life. The major aim of this book is that the readers could not only get access to the latest resource for synthetic biology, but also gain knowledge about how to conduct and optimize their own synthetic biology project.

Keywords Basic science · Application science · Multidiscipline · Synthetic biology project

Synthetic biology has gone through a hard time in proving its effectiveness in healthcare and industry in the past 10–20 years [1]. As a basic science, it has grown from simply biomedical engineering to an area that stands on both engineering and omics big data, and thus it has become an area with strong problem-driven and data-driven features [2]. As an application science, it has already enabled complex modification of genomes, yielding important metabolites useful in a broad spectrum of applications, based on rational editing of the genomes [3]. In summary, it has undergone the scientific renaissance that could prove it to be a multidiscipline research area, from where we can answer questions on why, what, and how to build a synthetic life (Fig. 9.1).

Synthetic biology is problem-driven, meaning that a synthetic biology study has always been dependent on specific problems. The major problems include but are

D. Zhao · K. Ning (✉)
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: ningkang@hust.edu.cn

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8_9

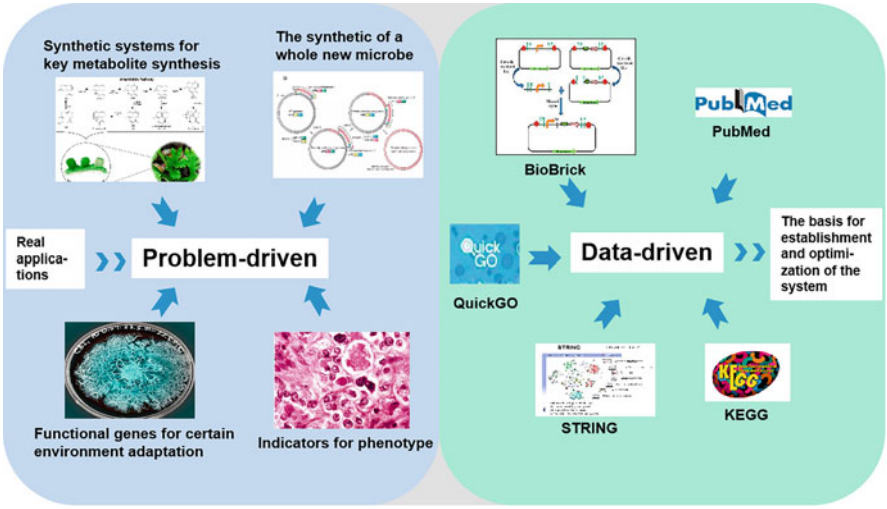

Fig. 9.1 The synthetic biology has become a problem-driven and data-driven multidiscipline engineering research area. The problem comes from real applications and promotes the synthetic biology; data resources is the basis for establishment and optimization of the system

not limited to, firstly, indicators for phenotype, such as indicators of soil or water environment pollution [4], indicators of cancer or other diseases [5], as well as indicators of special metabolites in engineering facilities [6]. Secondly, functional genes for certain environment adaptation, such as the bacterial adaptation for extreme conditions [7], the bacterial adaptation for animal stomach or gut [8], and the bacterial adaptation in resource-constrained environment [9]. Thirdly, synthetic systems for key metabolite synthesis, exemplified by the synthesis of artemisinin [10] and cannabinoids [11]. Fourthly, the synthetic of a whole new microbe such as *Escherichia coli* [12] and yeast [13]. Finally, synthetic systems for the promotion or suppression of phenotypes. Collectively, synthetic biology is driven by problems arising from real applications, which would affect multiple aspects of our health, the better production of materials needed by us, as well as the environment around us.

Synthetic biology is also data-driven since current synthetic biology is tightly linked with multi-omics studies [14]. For example, in the collection of parts and modules, there are already more than 50,000 parts registered in the BioBrick parts database [15], while more than millions of possible databases are available for mining of functional genes from representative databases such as iGEM Registry, UniProt [16], QuickGO [17], KEGG [18], BioGRID [19], ExplorEnz [20], STRING [21], and PubMed [22]. In the optimization of chassis, a common problem in its application is the host-interference problem, for which the most proficient way to overcome the host-interference problem is through host-interference problem [23], and genome simplification is largely dependent on the multi-omics analysis [24]. In the optimization of the synthetic system, a key part is about the design and

optimization of a system that could work best under the specific condition, which means that a large quantity of omics data at different levels, as well as millions of attempts, would need to be conducted. In summary, data-driven approach has already been realized in most of the synthetic biology studies, serving as the basis for the establishment and optimization of the system.

Synthetic biology still has a strong flavor of engineering, and it would continue to has flavor of engineering. The engineering idea, both top-down and bottom-up, still dominated the synthetic biology studies [25]. Top-down and bottom-up synthetic biology has developed methods for building and manipulating living systems. Both fields face challenges, but both also present opportunities. For example, top-down synthetic biology has equipped existing cells with new functionalities, for example, enabling *E. coli* to produce artemisinin by means of introducing artemisinin synthesis pathways [26]. While the bottom-up approach starts with cellular parts to study their function in isolation, which involves creating new biological systems in vitro by bringing together "nonliving" biomolecular components, for example, the synthetic of a whole new microbe such as *E. coli* [12] and yeast [13]. Collectively, only by the combination of multi-omics data and concrete problems the engineering of biological system could achieve optimal status in a synthetic biology study.

However, much more questions are still in front of synthetic biology, for both young and established researchers, to solve. For example, what is the upper limit of efficiency for producing key metabolite by a synthetic life? How far are we from the "digital life system"? Can we design a synthetic life from scratch for a target metabolite? All of these issues remain for us to explore.

We wrote this book to summarize the current developments in synthetic biology, with a special focus on iGEM. The hope is that the readers could not only get access to the latest resource for synthetic biology but also gain knowledge about how to conduct and optimize their own synthetic biology project.

Good luck!

References

1. Cuccato, G., Della Gatta, G., di Bernardo, D.: Systems and synthetic biology: tackling genetic networks and complex diseases. Heredity. **102**(6), 527–532 (2009)
2. Hagemann, M., Hess, W.R.: Systems and synthetic biology for the biotechnological application of cyanobacteria. Curr. Opin. Biotechnol. **49**, 94–99 (2018)
3. Palazzotto, E., et al.: Synthetic biology and metabolic engineering of actinomycetes for natural product discovery. Biotechnol. Adv. **37**(6), 107366 (2019)
4. Capeness, M.J., Horsfall, L.E.: Synthetic biology approaches towards the recycling of metals from the environment. Biochem. Soc. Trans. **48**(4), 1367–1378 (2020)
5. Weber, E.W., Maus, M.V., Mackall, C.L.: The emerging landscape of immune cell therapies. Cell. **181**(1), 46–62 (2020)
6. Bills, G.F., Gloer, J.B.: Biologically active secondary metabolites from the fungi. Microbiol. Spectr. **4**, 6 (2016)
7. Kitada, T., et al.: Programming gene and engineered-cell therapies with synthetic biology. Science. **359**, 6376 (2018)

8. Cho, S., Shin, J., Cho, B.K.: Applications of CRISPR/Cas system to bacterial metabolic engineering. Int. J. Mol. Sci. **19**, 4 (2018)
9. Banner, A., Toogood, H.S., Scrutton, N.S.: Consolidated bioprocessing: synthetic biology routes to fuels and fine chemicals. Microorganisms. **9**, 5 (2021)
10. Wani, K.I., et al.: Enhancing artemisinin content in and delivery from Artemisia annua: a review of alternative, classical, and transgenic approaches. Planta. **254**(2), 29 (2021)
11. Gulck, T., et al.: Synthetic biology of cannabinoids and cannabinoid glucosides in Nicotiana benthamiana and Saccharomyces cerevisiae. J. Nat. Prod. **83**(10), 2877–2893 (2020)
12. Aslan, S., et al.: Design and engineering of E. coli metabolic sensor strains with a wide sensitivity range for glycerate. Metab. Eng. **57**, 96–109 (2020)
13. Patra, P., et al.: Recent advances in systems and synthetic biology approaches for developing novel cell-factories in non-conventional yeasts. Biotechnol. Adv. **47**, 107695 (2021)
14. Kim, Y.M., et al.: Editorial: multi-omics technologies for optimizing synthetic biomanufacturing. Front. Bioeng. Biotechnol. **9**, 818010 (2021)
15. Boyle, P.M., et al.: A BioBrick compatible strategy for genetic modification of plants. J. Biol. Eng. **6**(1), 8 (2012)
16. Boutet, E., et al.: UniProtKB/Swiss-Prot. Methods Mol. Biol. **406**, 89–112 (2007)
17. Binns, D., et al.: QuickGO: a web-based tool for gene ontology searching. Bioinformatics. **25**(22), 3045–3046 (2009)
18. Kanehisa, M., et al.: KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. **45**(D1), D353–D361 (2017)
19. Oughtred, R., et al.: The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. **30**(1), 187–200 (2021)
20. McDonald, A.G., et al.: ExplorEnz: a MySQL database of the IUBMB enzyme nomenclature. BMC Biochem. **8**, 14 (2007)
21. Szklarczyk, D., et al.: The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. **49**(D1), D605–D612 (2021)
22. White, J.: PubMed 2.0. Med. Ref. Serv. Q. **39**(4), 382–387 (2020)
23. Beites, T., Mendes, M.V.: Chassis optimization as a cornerstone for the application of synthetic biology based strategies in microbial secondary metabolism. Front. Microbiol. **6**, 906 (2015)
24. Cotten, C., Reed, J.L.: Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. BMC Bioinformatics. **14**, 32 (2013)
25. Coudreuse, D.: Insights from synthetic yeasts. Yeast. **33**(9), 483–492 (2016)
26. Farhi, M., et al.: Metabolic engineering of plants for artemisinin synthesis. Biotechnol. Genet. Eng. Rev. **29**, 135–148 (2013)

Appendix

Databases for Synthetic Biology

Database	Type	Description
parts.igem.org	BioBricks	Is the central parts library funded by the biobricks.org foundation and originated from the MIT.
openwetware.org	Lab notes	OpenWetWare is an effort to promote the sharing of information, know-how, and wisdom among researchers and groups who are working in biology and biological engineering. It contains lab notes, material descriptions, etc.
biofab.synberc.org registry.synberc.org	BioBricks	International Open Facility Advancing Biotechnology (BIOFAB) was founded in December 2009 as the world's first biological design-build facility. BioFab projects will be designed to produce broadly useful collections of standard biological parts that can be made freely available to both academic and commercial users, while also enabling the rapid design and prototyping of genetic constructs needed to support specific needs of partner efforts such as SynBERC Testbeds.
biobricks.org	BioBricks	20 parts issued under the BioBrick™ Public Agreement (BPA), which is a free-to-use legal tool that allows individuals, companies, and institutions to make their standardized biological parts free for others to use.
acs-registry.jbei.org	BioBricks	BioBricks specialized for BioFuels, around 200 open-source sequences are disclosed.
Literature database	Literature	Literature database that relates to DNA2.0 Research

(continued)

K. Ning, Y. Zhan (eds.), *Synthetic Biology and iGEM: Techniques, Development and Safety Concerns*, https://doi.org/10.1007/978-981-99-2460-8

(continued)

Database	Type	Description
NCBI/GenBank	DNA / Bio Security	GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.
EBI/EMBL	Biological database	The EBI Search engine, also known as EB-eye, is a scalable text search engine that provides easy and uniform access to the biological data resources hosted at the EMBL-EBI. The data resources represented in the EBI Search engine include: nucleotide and protein sequences at both the genomic and proteomic levels, structures ranging from chemicals to macro-molecular complexes, gene-expression experiments, binary level molecular interactions as well as reaction maps and pathway models, functional classifications, biological ontologies, and comprehensive literature libraries covering the biomedical sciences and related intellectual property.
DDBJ	Bio Security	
Envision / Encore	DNA Search	Presentation on Encore and Envision
Biocatalogue	Web Services	Providing a curated catalog of life science Web services.
SWISS-Prot	Annotated Protein database	Database description: http://en.wikipedia.org/wiki/UniProt Database Structure: http://arep.med.harvard.edu/labgc/jong/Fetch/SwissProtAll.html
UniProt	Protein database	Database description: https://www.uniprot.org Uniprot (Universal Protein), which combines Swiss-Prot, TrEMBL, and PIR-PSD data, is the most informative and resource-rich free protein database available.
OpenWetWare		
ChEBI	Metabolomics database	Metabolomics database and ontology.
GenBank(database)	Proteomics database	Proteomics database open access annotated collection of all publically available nucleotide sequences and their protein transitions.
Human Metabolome Database (HMDB)	Human metabolite and pathway database	Human metabolite and pathway database.

(continued)

(continued)

Database	Type	Description
KEGG	Collected database	Collection of databases dealing with genomes biological pathways, disease, drugs, and chemical substances.
PubMed Database	An important source of literature download	PubMed database is a free biomedical literature retrieval system developed by the National Center for Biotechnology Information (NCBI), which is affiliated to the American Library of Medicine.

Software for Synthetic Biology

Type	Name	Description
Software List	iGem Software Tools	List of software developed in iGem competitions, maybe there are some interesting things
Genome Designer	Gene Designer 2.0	Genome Designer with Drag and Drop
Gene Search	CRISPR gRNA Design tool	
Gene Analysis	Atlas Plasmid Mapper	Enter a DNA and it shows you what's in it
Gene Analysis	DNA2.0 Bioinformatics Toolbox	Many tools to analyze DNA
Genome Designer	Genome Compiler	Genome Design with Drag and Drop of standard parts and printing options, AutoCAD is in the Advisory Board
Genome Designer	Generous	Genome Designer with Drag and Drop
Genome Editing	DESKGEN	CRISPR genome-editing software and bioinformatics platform
Genome and Pathway Designer	Archetype	Discover, analyze, and build synthetic genes and pathways
PARADIGM	Probabilistic graphical models	Probabilistic graphical models using directed factor graphs
iCluster	Joint latent variable model	Joint latent variable model-based clustering method
iClusterPlus	Generalized linear regression	Generalized linear regression for the formulation of the joint model
LRAcluster	Probabilistic	Probabilistic The model with low-rank approximation