IntechOpen

IntechOpen Series
Biomedical Engineering, Volume 32

Bioinformatics

Recent Advances

Edited by Ibrokhim Y. Abdurakhmonov

Bioinformatics - Recent Advances

Edited by Ibrokhim Y. Abdurakhmonov

Published in London, United Kingdom

Bioinformatics - Recent Advances
http://dx.doi.org/10.5772/intechopen.1005956
Edited by Ibrokhim Y. Abdurakhmonov

Contributors
Francesco Pappalardo, Giulia Russo, Ibrokhim Y. Abdurakhmonov, Jianbo Jian, Jun Cheng, Kinjalben Suthar, Mirzakamol S. Ayubov, Nisha Singh, Qiang Gao, Xiaogeng Wan, Ye Yin

First published in London, United Kingdom, 2025 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 167-169 Great Portland Street, London, W1W 5PF, United Kingdom

For EU product safety concerns: IN TECH d.o.o., Prolaz Marije Krucifikse Kozulić 3, 51000 Rijeka, Croatia, info@intechopen.com or visit our website at intechopen.com.

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

Bioinformatics - Recent Advances
Edited by Ibrokhim Y. Abdurakhmonov
p. cm.

This title is part of the Biomedical Engineering Book Series, Volume 32
Topic: Bioinformatics and Medical Informatics
Series Editor: Robert Koprowski
Topic Editor: Slawomir Wilczynski

Print ISBN 978-1-83634-626-5
Online ISBN 978-1-83634-625-8
eBook (PDF) ISBN 978-1-83634-627-2
ISSN 2631-5343

IntechOpen Book Series

Biomedical Engineering

Volume 32

Aims and Scope of the Series

Biomedical Engineering is one of the fastest-growing interdisciplinary branches of science and industry. The combination of electronics and computer science with biology and medicine has improved patient diagnosis, reduced rehabilitation time, and helped to facilitate a better quality of life. Nowadays, all medical imaging devices, medical instruments, or new laboratory techniques result from the cooperation of specialists in various fields. The series of Biomedical Engineering books covers such areas of knowledge as chemistry, physics, electronics, medicine, and biology. This series is intended for doctors, engineers, and scientists involved in biomedical engineering or those wanting to start working in this field.

Meet the Series Editor

Robert Koprowski, MD (1997), Ph.D. (2003), Habilitation (2015), is an employee of the University of Silesia, Poland, Institute of Computer Science, Department of Biomedical Computer Systems. For 20 years, he has studied the analysis and processing of biomedical images, emphasizing the full automation of measurement for a large inter-individual variability of patients. Dr. Koprowski has authored more than a hundred research papers with dozens in impact factor (IF) journals and has authored or co-authored six books. Additionally, he is the author of several national and international patents in the field of biomedical devices and imaging. Since 2011, he has been a reviewer of grants and projects (including EU projects) in biomedical engineering.

Meet the Volume Editor

Ibrokhim Y. Abdurakhmonov received his BS in Biotechnology (1997) from the National University of Uzbekistan, an MS in Plant Breeding (2001) from Texas A&M University, and a Ph.D. in Molecular Genetics (2002) and a D. Sc. in Genetics (2009) from the Academy of Sciences of Uzbekistan. He founded the Center of Genomics and Bioinformatics of Uzbekistan in 2012. He received the 2010 TWAS Prize and the ICAC Cotton Researcher of the Year Award 2013 for his outstanding contributions to cotton genomics and biotechnology. Dr. Abdurakhmonov was elected as a fellow of The World Academy of Sciences (TWAS) in 2014 and a member of the Academy of Sciences of Uzbekistan in 2017. In 2023, he was awarded the UNESCO-Guinea Equatorial Life Science Award for his contributions to the development of molecular markers and disease resistance in cotton. He received the 8th VCANBIO Award for Biosciences and Medicine in 2023.

Contents

Preface

Over the past few decades, the field of bioinformatics has undergone a significant transformation, emerging as an essential bridge between the life sciences and computational technology. As high-throughput technologies generate increasing amounts of biological data, there is a growing demand for new computational methods to analyze, interpret, and utilize this information. This trend is reflected in the rise of research publications over the past several decades, with nearly 250,000 publications appearing in the last five years.

Bioinformatics – Recent Advances addresses this dynamic and rapidly evolving landscape, offering a concise overview of the latest developments and breakthroughs in various fields of bioinformatics. From advancements in algorithmic innovations for sequence analysis and structural biology to the integration of artificial intelligence in genomics and systems biology, each chapter examines a key area where cutting-edge research is shaping the future of life sciences. The book's content showcases the interdisciplinary nature of modern bioinformatics, bringing together insights from experts in biology, computer science, data science, and math.

This volume aims to inform and inspire researchers, students, and professionals in this evolving field. Whether academics seek new methods, practitioners employ bioinformatics tools, or students need a solid foundation in current trends, this book serves as a valuable and accessible resource. The updated chapters explore recent advancements in extracting protein sequence features, the challenges and opportunities of multi-omics applications, breakthroughs in agent-based modeling, and genome assembly algorithms, along with an overview of the latest trends in bioinformatics.

I acknowledge the authors and reviewers who brought this collection to life, as well as their hard work and invaluable contributions to this volume. Their expertise and commitment to advancing bioinformatics research are evident in the quality and clarity of their work. I also want to recognize the growing community of bioinformatics researchers and life science users who are collaborating to expand our understanding of life at the molecular level.

I hope that *Bioinformatics – Recent Advances* will spark further inquiry, foster innovation, and support the next wave of discoveries in the life sciences. I am also grateful to the IntechOpen book department and our author service manager, Ms. Kristina Kardum Cvitan, for the opportunity to work on this project and assist with my editorial duties.

Ibrokhim Y. Abdurakhmonov
Center of Genomics and Bioinformatics,
Academy of Sciences of Uzbekistan,
Tashkent, Uzbekistan

Section 1

New Trends in Bioinformatics

Chapter 1

Latest Trends in Bioinformatics Development

Ibrokhim Y. Abdurakhmonov and Mirzakamol S. Ayubov

Abstract

Breakthroughs in bioinformatics revolutionize our understanding of biological systems and accelerate advancements in life sciences, including agricultural and biomedical research. Thanks to advances in technology, artificial intelligence (AI) and machine learning now enable precise data analysis, leading to accurate predictions and the discovery of complex patterns in whole-genome datasets. Single-cell sequencing technologies are continually improving, providing scientists with a more detailed view of cellular diversity and development. Meanwhile, multi-omics approaches, which combine genomics, transcriptomics, proteomics, and metabolomics, are deepening our understanding of biological processes through ontogenesis. Recent advancements in bioinformatics tools, particularly those utilizing CRISPR-Cas systems, have refined guide RNA design, enhanced off-target prediction, and facilitated functional validation, thereby accelerating the development of gene therapy. Structural bioinformatics has advanced due to the development of enhanced protein structure prediction models, which aid in drug discovery and functional annotation. Research on metagenomics and microbiomes highlights the significant roles of microbial communities in health, agriculture, and environmental processes. The rise of precision medicine and innovative agri-biotechnologies depends on the integration of genomic and farming/clinical data, leading to precision strategies that provide efficient solutions to targeted questions. Bioinformatics-guided drug design and next-generation vaccine development, as well as modeled personalized medicine, will be the developing areas. Further, cloud and quantum computing, automation, and scalable workflows have made high-throughput analysis more accessible, encouraging collaboration and reproducibility. These trends signify a dynamic period for bioinformatics, characterized by technological innovation and increased collaboration, which ultimately accelerates discoveries in the life sciences.

Keywords: bioinformatics development, bioinformatics publications, deep learning tools, AI tools, bioinformatics-based crop design, personalized medicine

1. Introduction

Bioinformatics is a rapidly expanding field that integrates biology, computer science, and information technology to process, analyze, and evaluate scientific data. We periodically reviewed the developments over the past decades in Intech Open [1, 2]. It significantly impacts the future of life sciences, particularly in design, production,

monitoring, and management, especially in agriculture and healthcare. Recently, bioinformatics has made remarkable strides, thanks to the convergence of artificial intelligence (AI), machine learning (ML), and next-generation sequencing (NGS). NGS facilitates genome-wide screening, which fosters crucial innovations for economic growth. Applications of artificial intelligence and deep learning for automated feature extraction and predictive modeling have also shown significant improvements [3, 4]. It is anticipated that bioinformatics will greatly influence the future of healthcare in various ways, including drug discovery, disease detection and prevention, and the management of specific health concerns [5]. Similarly, crop development capable of withstanding farming in the age of global climate change has greatly benefited from current and emerging bioinformatics tools [6].

A PubMed [7] search for "bioinformatics" across all organisms, as of June 2025, yields 587,623 publications from 1958 to the present, including books, documents, clinical trials, meta-analyses, and review articles. Surprisingly, 244,033 of these were published in just the last 5 years (2020–2025), demonstrating the field's rapid expansion (**Figures 1**–**3**; **Table 1**). This body of work encompasses key developments, beginning with the Pre-Bioinformatics Era (the discovery of DNA structure, early protein sequencing, and foundational computational biology) and concluding with the Emergence of Bioinformatics (the establishment of sequence databases, such as GenBank and the Protein Data Bank, the development of pairwise alignment algorithms, and the introduction of BLAST) [1, 2].

The Human Genome Project, microarray technology, NGS [8], and the ENCODE Project all contributed to propelling the Genomics Revolution forward [9]. Later, Systems Biology and Big Data transformed research by integrating multiple omics (genomics, transcriptomics, proteomics, and metabolomics) [10, 11], CRISPR-Cas9 gene editing (utilizing bioinformatics to guide RNA design) [12], single-cell RNA-Seq for analyzing cellular heterogeneity, and AI/ML [13]. Long-read sequencing, personalized medicine, AlphaFold's innovative protein structure predictions, metagenomics and microbiome analysis, and impending quantum computing applications [1, 2, 8, 14, 15] in biological research all characterize the Modern Era today.

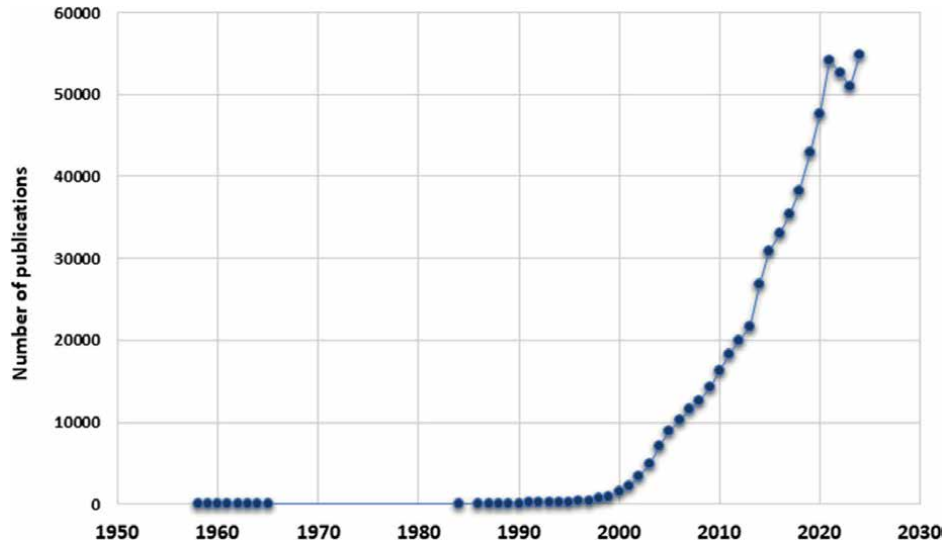

Figure 1.
Timeline of bioinformatics publications based on PubMed [7] data.

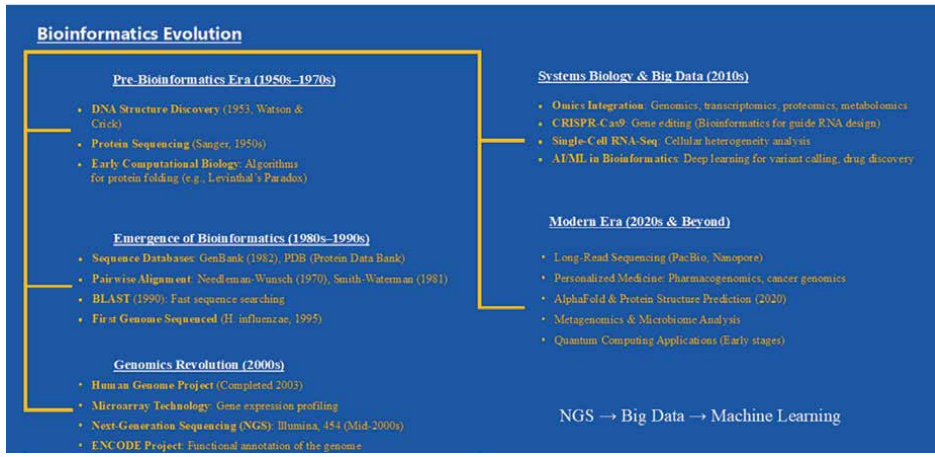

Figure 2.
Bioinformatics evolution: Past and present.

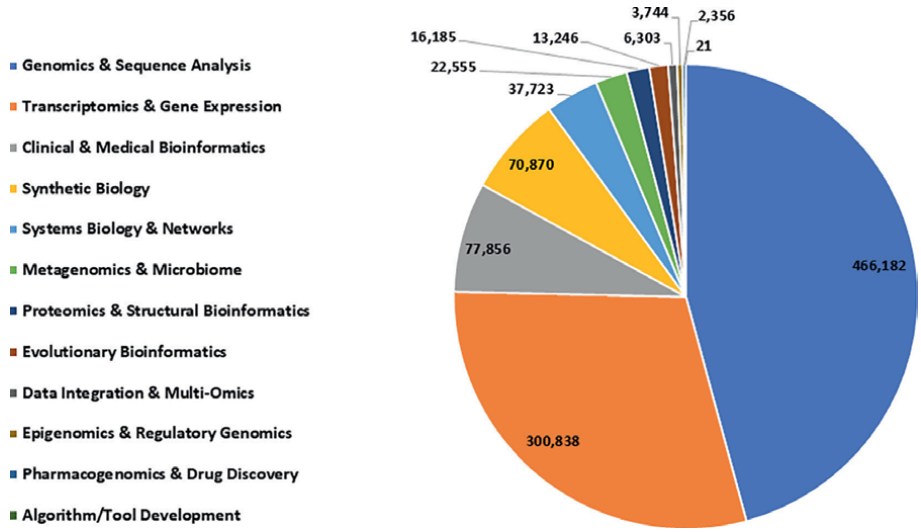

Figure 3.
Publications categorized by direction.

Since 1987, 38,776 articles have been published in agriculture, focusing on the latest bioinformatics-based technologies and their applications. The main efforts have been on building genomic resources, sequencing genomes and germplasm, using sequencing to map important traits, and developing marker-assisted or genome-assisted breeding methods to create crops that can withstand biotic and abiotic stresses and have higher nutrition levels in 12 key crops, including barley, rice, maize, wheat, pearl millet, sorghum, cowpea, soybean, common bean, groundnut, pigeon pea, and chickpea [16]. In recent years, scientists have been working on comparative genomics, transcriptomics, and studies of bacterial microbiomes using next-generation sequencing. Studies have also explored the workflow of transcriptomics and experimental considerations using bioinformatics approaches [17, 18].

The field of structural bioinformatics has undergone significant improvements in recent decades, mainly due to advances in X-ray crystallography, nuclear magnetic

#	Categories	Publications	From to present	PubMed address
1	Genomics and Sequence Analysis	466,182	1963	https://pubmed.ncbi.nlm.nih.gov/?term=Genomics+%26+Sequence+Analysis
2	Transcriptomics and Gene Expression	300,838	1982	https://pubmed.ncbi.nlm.nih.gov/?term=Transcriptomics+%26+Gene+Expression
3	Clinical and Medical Bioinformatics	77,856	1989	https://pubmed.ncbi.nlm.nih.gov/?term=Clinical+%26+Medical+Bioinformatics
4	Synthetic Biology	70,870	1936	https://pubmed.ncbi.nlm.nih.gov/?term=Synthetic+Biology
5	Systems Biology and Networks	37,723	1970	https://pubmed.ncbi.nlm.nih.gov/?term=Systems+Biology+%26+Networks
6	Metagenomics and Microbiome	22,555	2000	https://pubmed.ncbi.nlm.nih.gov/?term=Metagenomics+%26+Microbiome
7	Proteomics and Structural Bioinformatics	16,185	1997	https://pubmed.ncbi.nlm.nih.gov/?term=Proteomics+%26+Structural+Bioinformatics
8	Evolutionary Bioinformatics	13,246	1993	https://pubmed.ncbi.nlm.nih.gov/?term=Evolutionary+Bioinformatics
9	Data Integration and Multi-Omics	6303	2002	https://pubmed.ncbi.nlm.nih.gov/?term=Data+Integration+%26+Multi-Omics
10	Epigenomics and Regulatory Genomics	3744	2002	https://pubmed.ncbi.nlm.nih.gov/?term=Epigenomics+%26+Regulatory+Genomics
11	Pharmacogenomics and Drug Discovery	2356	1989	https://pubmed.ncbi.nlm.nih.gov/?term=Pharmacogenomics+%26+Drug+Discovery
12	Algorithm/Tool Development	21	1999	https://pubmed.ncbi.nlm.nih.gov/?term=Algorithm%2FTool+Development

Table 1.
Bioinformatics directions and their establishment.

resonance (NMR) spectroscopy, and computing resources. The first protein structures were determined in the 1950s and 1960s, and since then, the number of identified structures has increased dramatically [3]. Today, over 170,000 structures have been deposited in the Protein Data Bank (PDB), a public collection of biomolecular structures. Modern advances in healthcare technology have ushered in a new era of personalized medicine. The growth of bioinformatics in personalized medicine has enabled the analysis of vast amounts of data, dramatically transforming the potential for delivering tailored healthcare [4, 5].

2. Bioinformatics in the COVID-19 pandemic

During the COVID-19 pandemic, the importance of bioinformatics in global health crises became clear. Bioinformatics tools played a crucial role in understanding

SARS-CoV-2, developing diagnostic tests, and accelerating vaccine development. Sequencing technologies enabled the quick decoding of the viral genome, while bioinformatics analyzed thousands of viral sequences worldwide [19, 20]. This helped track mutations and new variants, guiding public health responses. Phylogenetic analysis revealed patterns of infection spread among different viral strains [21].

Between 2020 and 2022, whole-genome bioinformatics analyses of the SARS-CoV-2 virus significantly helped in combating this deadly virus. Whole-genome sequencing (WGS) has proven to be a valuable tool for understanding newly emerging infectious diseases, tracking microbial transmission networks, and providing cost-effective insights when integrated into hospital surveillance programs. The COVID-19 pandemic accelerated the use of whole-genome sequencing (WGS), with viral sequencing conducted at unprecedented levels globally. As of this writing, over 21,013,106 SARS-CoV-2 genomes have been shared on the global initiative on sharing all influenza data (GISAID) database [22]. These genomics sequences, powered by bioinformatics, enabled the development and refinement of diagnostic tools, such as PCR primers and rapid antigen tests, ensuring high accuracy and specificity. In this regard, data sharing platforms like GISAID [22] have fostered global collaboration, enabling scientists to exchange viral sequences and insights in real time. All of these have been instrumental in understanding the virus's evolution, monitoring clinically relevant variants, and informing public health policies [23].

Additionally, bioinformatics played a crucial role in vaccine design. By applying structural biology and protein modeling, researchers were able to predict the structures of key viral proteins, including the spike protein, which has become the primary target for vaccine development. Computational modeling also helped optimize vaccine candidates and assess their potential effectiveness before clinical trials began [24]. AI-powered bioinformatics approaches have contributed to drug repurposing efforts and the understanding of viral-host interactions, revealing potential therapeutic targets. Repurposing existing drugs can accelerate the development of new treatments, particularly when their safety profiles are already known from previous disease research. This process can be even quicker when the drugs have already been approved for different diseases and post-marketing safety data are available [25, 26]. As a result, bioinformatics accelerated the scientific response, enhanced surveillance, and facilitated the development of vaccines and treatments, ultimately saving lives and informing policies during the pandemic. This impact underscored the crucial role of data-driven science [27] in combating infectious diseases and preparing for future outbreaks.

3. Artificial intelligence and bioinformatics

Over the past decade, artificial intelligence (AI) and machine learning (ML) have emerged as groundbreaking technologies poised to revolutionize pharmaceutical research and development. This transformation is partly driven by revolutionary advancements in computing power and the removal of traditional barriers to large-scale data collection and analysis [15]. Meanwhile, the costs of bringing new pharmaceuticals to market and to patients have soared. Recognizing these challenges, AI/ML approaches appeal to the pharmaceutical industry due to their automated nature, predictive capabilities, and anticipated increase in efficiency. Over the past 15–20 years, drug discovery has increasingly relied on machine learning technologies. The latest area of drug research where AI/ML is creating positive disruption is in the

design, conduct, and analysis of clinical trials. The COVID-19 pandemic may accelerate the adoption of AI/ML in clinical trials, reflecting a greater reliance on digital technology during trial execution [28].

While progress has been made, current methods fall short without similar structures for comparison. Recent work for CASP14 (the 14th Critical Assessment of Protein Structure Prediction) [29] has led to a new computational approach that enables reliable protein structure predictions without relying on known similar structures. This breakthrough has led to a revamped AlphaFold, a neural network model that often matches experimental accuracy and advances the field of bioinformatics. The latest AlphaFold incorporates physical and biological knowledge, as well as multisequence alignments, to boost its deep learning design [14].

Recent progress in Natural Language Processing (NLP) has been significantly accelerated by the emergence of Large Language Models (LLMs), marking a major advancement in language technology capabilities. These models, built on intricate deep learning architectures known as transformers, comprise billions of parameters and vast amounts of training data, enabling them to achieve impressive accuracy across a wide range of tasks. The transformer architecture of LLMs empowers them to effectively handle context and sequential information, which is essential for understanding and generating human language. Beyond traditional NLP applications, LLMs have demonstrated considerable potential in bioinformatics, transforming the field by addressing challenges associated with vast and complex biological data. LLMs are advantageous in genomics, proteomics, and personalized medicine for identifying patterns, predicting protein structures, and analyzing genetic variants. This ability is crucial, for instance, in drug discovery, where accurate prediction of molecular interactions is required [30]. Cloud computing has made high-throughput analysis more accessible, encouraging collaboration and reproducibility [31]. Automation and scalable workflows enhance data processing, minimize biases, and boost reproducibility. Cloud-based infrastructure, automation, and scalable workflows are fundamental to the advancement of precision healthcare, rendering agriculture a more sustainable and productive sector. These tools facilitate efficient data utilization, stimulate innovation, and allow for targeted, personalized medical interventions within a digital health framework [32]. Furthermore, they equip farmers with actionable insights, enhance resource efficiency, and bolster resilience against the impacts of climate change and challenges related to food security [33].

4. Open-source software and databases

Open-source software and databases are crucial to the development of bioinformatics, offering cost savings, flexibility, and the ability to switch between options. They foster broader collaboration, enable customized solutions, and provide increased security, scalability, improved performance, and learning opportunities. There are many resources available in this area. For instance, the latest version of PyRosetta [34] provides an interactive Python interface to the Rosetta molecular-simulation toolkit. This enables users to develop their own molecular-modeling algorithms utilizing Rosetta's sampling methods and energy functions. Biopython [35], a collection of freely available biological computation tools written in Python, was developed by a multinational team of developers. It is a collaborative project aimed at creating Python modules and applications that meet both current and future bioinformatics research needs. The source code is distributed under the Biopython License,

which is highly permissive and compatible with nearly every license worldwide. Data repositories such as NCBI, UniProt, PDB, and the cancer genome atlas (TCGA) have become increasingly accessible, offering significant opportunities for developers, businesses, and educators. These open-source solutions support open standards, making integration and long-term use more sustainable and reliable.

5. Future perspectives of bioinformatics

Bioinformatics is moving beyond classical statistics to include complex AI models in the processing and interpretation of biological data [36]. Alphafold-like Revolution: DeepMind's AlphaFold2 won a 50-year grand challenge by predicting protein structures with amazing precision [37]. This will ultimately be used to predict protein-protein interactions, protein-DNA/RNA interactions, and mutation effects.

It is conceivable to create AI models for drug discovery that generate data rather than simply analyze it [38]. This includes developing novel proteins, antibodies, and small-molecule medications with specialized therapeutic properties, resulting in significantly shorter development timelines. Bioinformatics can eventually be built utilizing improved pattern recognition. Deep learning will find tiny, sophisticated patterns in multi-omics data (genomics, proteomics, and metabolomics) that humans cannot see, yielding novel disease biomarkers and scientific insights [39].

Multi-omics and data integration may be among the most important directions in bioinformatics [40]. The future will entail merging all layers of biological information rather than focusing simply on one type of data (such as DNA). Bioinformatics will be essential for combining genome, transcriptomics, proteomics, metabolomics, and epigenomics data to create a full, systems-level view of a cell, tissue, or organ. Single-cell multi-omics technologies are being developed to sequence individual cells' DNA, RNA, and epigenetic landscapes [41]. Bioinformatics approaches are necessary to evaluate this data, which reveals significant heterogeneity within tissues.

Personalized and predictive medicine have the most direct impact on human health. The sequencing of a patient's genome is expected to become a widespread and economical diagnostic test [42]. Bioinformatics pipelines will be necessary to instantly evaluate this data in a clinical setting [43]. By combining genomic data with electronic health records, wearable device data, and environmental factors, AI models can forecast an individual's likelihood of developing particular diseases, enabling early, personalized prevention strategies. Bioinformatics is already utilized to identify driver mutations in cancer. Future techniques will drive real time, adaptive cancer therapy by identifying the best drug combination as the tumor grows.

Bioinformatics can also aid in the development of biotech vaccines, such as those based on DNA, RNA, and edible vaccines. It was demonstrated during the COVID-19 pandemic. Furthermore, microbiome analysis, cloud computing and scalable infrastructure, CRISPR and genome editing, and evolutionary and conservation biology all have the potential to be key bioinformatics disciplines.

6. Conclusion

Bioinformatics has become a leading field that combines biology, computer science, and information technology to drive breakthroughs in healthcare, agriculture, and drug development. The rapid rise in scientific research, particularly over the past

5 years, demonstrates the rapid growth of bioinformatics, driven by advancements in next-generation sequencing, artificial intelligence, and machine learning. The COVID-19 pandemic has accelerated the adoption of technologies, especially in areas such as viral-genome sequencing, pharmaceutical development, and health surveillance. Advances in structural bioinformatics and prediction, large-scale language models in genomics, and open-source bioinformatics tools have greatly enhanced the field's capabilities, setting the stage for personalized medicine and agricultural biotechnology. As bioinformatics continues to evolve, its integration with artificial intelligence, Big Data, and quantum computing is likely to bring even more innovations across all areas of the life sciences. The field's interdisciplinary nature means it plays a crucial role in addressing global challenges, from climate-resilient agriculture and personalized healthcare, making it a vital part of modern life sciences.

Author details

Ibrokhim Y. Abdurakhmonov* and Mirzakamol S. Ayubov
Center of Genomics and Bioinformatics of the Academy of Sciences of Uzbekistan, Tashkent, Republic of Uzbekistan

*Address all correspondence to: i.y.abdurakhmonov@gmail.com

IntechOpen

References

[1] Abdurakhmonov IY. Bioinformatics: Updated features and applications. In: Abdurakhmonov I, editor. London, UK: IntechOpen; 2016. DOI: 10.5772/61421. Available from: http://www.intechopen.com/books/bioinformatics-updated-features-and-applications

[2] Abdurakhmonov IY. Bioinformatics in the era of post-genomics and big data. In: Abdurakhmonov I, editor. London, UK: IntechOpen; 2018. pp. 1-177. DOI: 10.5772/intechopen.71349. Available from: https://www.intechopen.com/books/bioinformatics-in-the-era-of-post-genomics-and-big-data

[3] Ziegler SJ, Mallinson SJB, John PC, Bomble YJ. Advances in integrative structural biology: Towards understanding protein complexes in their cellular context. Computational and Structural Biotechnology Journal. 2021;**19**:214-225. DOI: 10.1016/j.csbj.2020.11.052

[4] Chellappa V. The Potential of Bioinformatics in Precision Medicine. 2023. Available from: https://www.linkedin.com/pulse/potential-bioinformatics-precision-medicine-venkatesh-chellappa/

[5] Viramgama N. Bioinformatics: The Key to the Future of Health Science. 2024. Available from: https://mypharmavision.com/pharmacy-exams/pebc/bioinformatics-the-key-to-the-future-of-health-science/

[6] Mishra R, Karada MS, Agnihotri D. Bioinformatics in crop improvement and agricultural genomics. In: Chaudhary A, Sethi SK, Verma A, editors. Unraveling New Frontiers and Advances in Bioinformatics. Singapore: Springer; 2024. DOI: 10.1007/978-981-97-7123-3_13

[7] PubMed Database [Internet]. 2025. Available from: https://pubmed.ncbi.nlm.nih.gov/ [Accessed: September 24, 2025]

[8] Abdurakhmonov IY. DNA sequencing - History, present and future. In: Abdurakhmonov I, editor. London, UK: IntechOpen; 2025. p. 132. DOI: 10.5772/ intechopen.1003355. Available from: https://www.intechopen. com/books/1003888

[9] ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004;**306**(5696):636-640. DOI: 10.1126/science.1105136

[10] Gupta A, Kumar S, Kumar A. Big data in bioinformatics and computational biology: Basic insights. Methods in Molecular Biology. 2024;**2719**:153-166. DOI: 10.1007/978-1-0716-3461-5_9

[11] Veenstra TD. Systems biology and multi-omics. Proteomics. 2021;**21**(3-4):e2000306. DOI: 10.1002/pmic.202000306

[12] Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nature Reviews Genetics. 2015;**16**(5):299-311. DOI: 10.1038/nrg3899

[13] Ji F, Sadreyev RI. Single-cell RNA-seq: Introduction to bioinformatics analysis. Current Protocols in Molecular Biology. 2019;**127**(1):e92. DOI: 10.1002/cpmb.92

[14] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;**596**:583-589. DOI: 10.1038/s41586-021-03819-2

[15] Kolluri S, Lin J, Liu R, Zhang Y, Zhang W. Machine learning and artificial intelligence in pharmaceutical research and development: A review. The AAPS Journal. 2022;**24**(1):19. DOI: 10.1208/s12248-021-00644-3

[16] Thudi M, Palakurthi R, Schnable JC, et al. Genomic resources in plant breeding for sustainable agriculture. Journal of Plant Physiology. 2021;**257**:153351. DOI: 10.1016/j.jplph.2020.153351

[17] Botton A. From the field to the lab: Establishing different abscission potentials in apple fruitlets for transcriptomic and functional studies. Methods in Molecular Biology. 2025;**2916**:121-137. DOI: 10.1007/978-1-0716-4470-6_12

[18] Goh HH, Ng CL, Loke KK. Functional genomics. Advances in Experimental Medicine and Biology. 2018;**1102**:11-30. DOI: 10.1007/978-3-319-98758-3_2

[19] Alfonsi T, Pinoli P, Canakoglu A. High performance integration pipeline for viral and epitope sequences. BioTech (Basel). 2022;**11**(1):7. DOI: 10.3390/biotech11010007

[20] Ma L, Li H, Lan J, et al. Comprehensive analyses of bioinformatics applications in the fight against COVID-19 pandemic. Computational Biology and Chemistry. 2021;**95**:107599. DOI: 10.1016/j.compbiolchem.2021.107599

[21] Tabibzadeh A, Esghaei M, Soltani S, et al. Evolutionary study of COVID-19, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as an emerging coronavirus: Phylogenetic analysis and literature review. Veterinary Medicine and Science. 2021;**7**(2):559-571. DOI: 10.1002/vms3.394

[22] GISAID Database [Internet]. 2025. Available from: https://gisaid.org/ [Accessed: September 24, 2025]

[23] Oude Munnink BB, Nieuwenhuijse DF, Stein M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nature Medicine. 2020;**26**:1405-1410. DOI: 10.1038/s41591-020-0997-y

[24] Abdurakhmonov IY. COVID-19 vaccines - current state and perspectives. In: Abdurakhmonov I, editor. London, UK: IntechOpen; 2023. pp. 1-160. DOI: 10.5772/intechopen.101006. Available from: https://www. intechopen.com/books/11724

[25] Ashburn TT, Thor KB. Drug repositioning: Identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery. 2004;**3**:673-683. DOI: 10.1038/nrd1468

[26] Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: Progress, challenges and recommendations. Nature Reviews Drug Discovery. 2019;**18**:41-58. DOI: 10.1038/nrd.2018.168

[27] Zhang Q. Data science approaches to infectious disease surveillance. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2022;**380**(2214):20210115. DOI: 10.1098/rsta.2021.0115

[28] Ali HH, Ali HM, Ali HM, et al. The role and limitations of artificial intelligence in combating infectious disease outbreaks. Cureus. 2025;**17**(1):e77070. DOI: 10.7759/cureus.77070

[29] Pereira J, Simpkin AJ, Hartmann MD, et al. High-accuracy protein structure prediction in CASP14. Proteins.

2021;**89**(12):1687-1699. DOI: 10.1002/prot.26171

[30] Sarumi OA, Heider D. Large language models and their applications in bioinformatics. Computational and Structural Biotechnology Journal. 2024;**23**:3498-3505. DOI: 10.1016/j.csbj.2024.09.031

[31] Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nature Reviews Genetics. 2018;**19**(4):208-219. DOI: 10.1038/nrg.2017.113

[32] The Lancet Psychiatry. Digital health: The good, the bad, and the abandoned. Lancet Psychiatry. 2019;**6**(4):273. DOI: 10.1016/S2215-0366(19)30102-6

[33] Shafi U, Mumtaz R, García-Nieto J, et al. Precision agriculture techniques and practices: From considerations to applications. Sensors (Basel). 2019;**19**(17):3796. DOI: 10.3390/s19173796

[34] PyRosetta [Internet]. 2025. Available from: https://www.pyrosetta.org/ [Accessed: September 24, 2025]

[35] Biophyton [Internet]. 2025. Available from: https://biopython.org/ [Accessed: September 24, 2025]

[36] Jamialahmadi H, Khalili-Tanha G, Nazari E, Rezaei-Tavirani M. Artificial intelligence and bioinformatics: A journey from traditional techniques to smart approaches. The Gastroenterology and Hepatology From Bed to Bench. 2024;**17**(3):241-252. DOI: 10.22037/ghfbb.v17i3.2977

[37] AlphaFold: A solution to a 50-year-old grand challenge in biology. [Internet]. 2020. Available from: https://deepmind.google/discover/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology/ [Accessed: September 24, 2025]

[38] Fu C, Chen Q. The future of pharmaceuticals: Artificial intelligence in drug discovery and development. Journal of Pharmaceutical Analysis. 2025;**15**(8):101248. DOI: 10.1016/j.jpha.2025.101248

[39] Chen C, Wang J, Pan D, Wang X, Xu Y, Yan J, et al. Applications of multi-omics analysis in human diseases. MedComm (2020). 2023;**4**(4):e315. DOI: 10.1002/mco2.315

[40] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinformatics and Biology Insights. 2020;**14**:1177932219899051. DOI: 10.1177/1177932219899051

[41] Guan A, Quek C. Single-cell multi-omics: Insights into therapeutic innovations to advance treatment in cancer. International Journal of Molecular Sciences. 2025;**26**(6):2447. DOI: 10.3390/ijms26062447

[42] Brittain HK, Scott R, Thomas E. The rise of the genome and personalised medicine. Clinical Medicine (London, England). 2017;**17**(6):545-551. DOI: 10.7861/clinmedicine.17-6-545

[43] Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. The Journal of Molecular Diagnostics. 2018;**20**(1):4-27. DOI: 10.1016/j.jmoldx.2017.11.003

Chapter 2

Genome Assembly Algorithms

Jianbo Jian, Qiang Gao, Jun Cheng and Ye Yin

Abstract

Currently, research has entered the genomic era. The high-throughput sequencing of short reads and long reads has increased, while the cost has decreased. Most of the key genomes have been sequenced, and an increasing number of reference genomes from uncommon species are currently in progress toward completion. What is more, a lot of finished draft genomes have been progressively refined and updated to achieve complete, telomere-to-telomere assemblies. Algorithms primarily focus on de novo assembly, evolving from Overlap-Layout-Consensus (OLC) for Sanger reads, to De Bruijn Graphs (DBG) for short reads, and back to OLC for PacBio or nanopore long reads. Scaffolding facilitates chromosome-level assembly, and graph-based algorithms enable pangenome assembly, which is poised to become a new standard for genomic references. A wide variety of genome assembly software has been extensively adopted, efficiently conserving computational resources while improving genome quality.

Keywords: *de novo* assembly, whole genome shotgun, De Bruijn, scaffolding, telomere-to-telomere

1. Introduction

The significant advancements in genomics, coupled with improvements in genome assembly technologies, have greatly enhanced our understanding of life sciences [1]. Furthermore, these developments have driven innovations across various fields, including medicine [2], agriculture [3], biotechnology [4], and more, offering solutions to global challenges such as disease treatment [5], food security [6], and environmental protection [7]. Genome sequences serve as the foundation for understanding the biology and functionality of an organism, and *De novo* genome assembly is a necessary and crucial initial step in bioinformatics analysis [8]. Genome assembly has made significant advancements over the past four decades, driven by the progression of sequencing technologies, the refinement of algorithms, and enhanced computational capabilities. According to the development of sequencing and assembly strategies, it is possible to define five stages of genome evolution. The first-generation genome, utilizing Sanger sequencing technology, involved the construction of Bacterial Artificial Chromosomes (BACs) for initial genome assembly. Subsequently, Whole Genome Shotgun (WGS) sequencing was applied to further enhance the assembly process. The assembly methodology relied on the Overlap-Layout-Consensus (OLC) approach. During this early stage, the human genome served as a representative example of successful first-generation sequencing efforts (**Figure 1**). The second-generation genome sequencing and assembly rely solely on

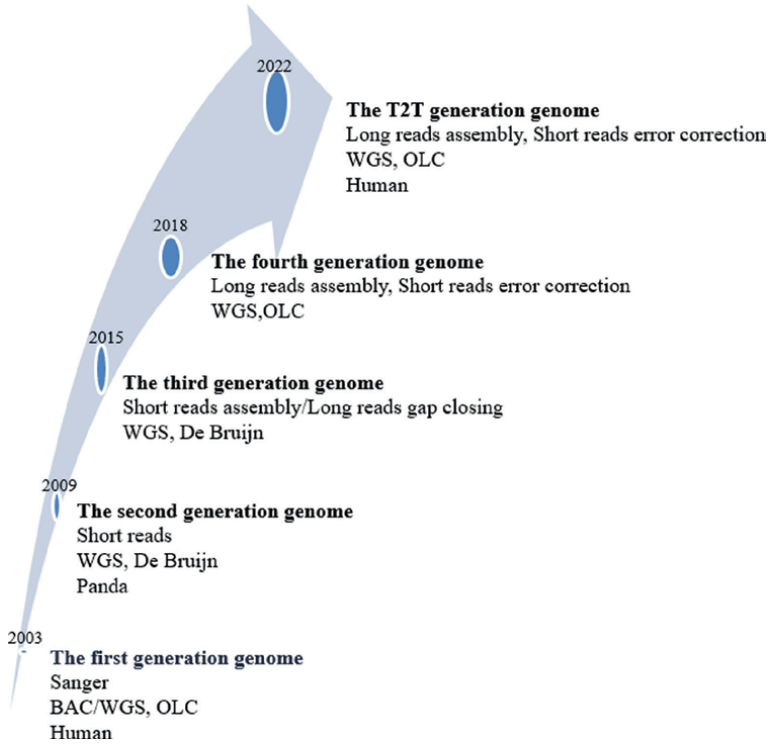

Figure 1.
The different stages of the genome.

short reads using WGS sequencing and De Bruijn Graph assembly methods beginning in 2009. The giant panda genome was the first completed genome assembled exclusively from short reads (**Figure 1**). In 2015, long-read sequencing was still considered expensive despite its potential. The third-generation genome assembly primarily relied on short reads, while long reads were utilized for gap closing. In this stage, the genome assembly still relied on WGS and DBG methods (**Figure 1**). In 2018, the cost of long-read sequencing decreased while the quality and throughput improved.

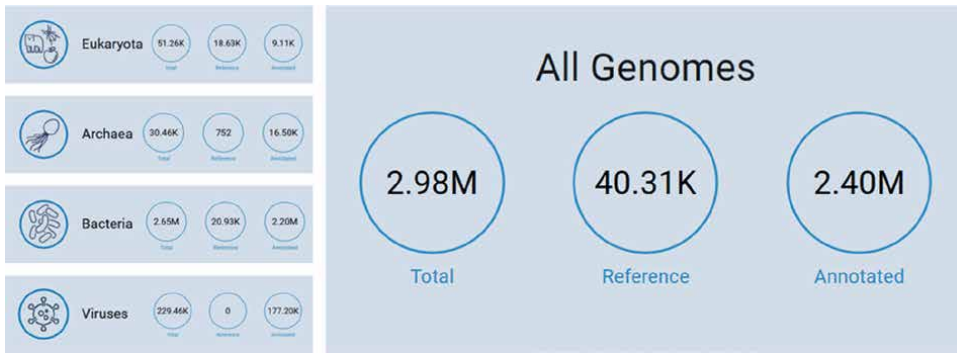

Figure 2.
The statistics of available genomes in NCBI. The data was downloaded from https://www.ncbi.nlm.nih.gov/datasets/genome/ on May 6, 2025.

The fourth-generation genomes were primarily assembled using long reads through WGS and OLC strategies, with short reads employed exclusively for error correction (**Figure 1**). In 2022, the telomere-to-telomere (T2T) generation genome era arrived with the completion of the human T2T genome. The use of ultra-long reads and PacBio HiFi reads has enabled the assembly of the highest-quality genomes to date (**Figure 1**). Until May 6, 2025, the total number of available genomes in NCBI is 2.98 million, including 51.26 K in Eukaryota, 30.46 K in Archaea, 2.65 M in Bacteria, and 229.46 K in Viruses (**Figure 2**). In Eukaryota, a total of 51.26 K genomes have been assembled, including 18.63 K that are referenced to a reference genome and 9.11 K genomes with annotation information (**Figure 2**).

2. The genome project

For de novo genome assembly, the first completed genome sequence was that of bacteriophage phi X174, comprising 5375 nucleotides and achieved in 1977 [9]. In 1995, the first bacterium genome of *Haemophilus influenzae* Rd., consisting of 1,830,137 base pairs, was fully sequenced and assembled [10]. In 1996, the first eukaryotic genome, that of the yeast Saccharomyces cerevisiae (12,068 kb), was fully sequenced and assembled [11]. Then, in 1998, the first genome sequence of a multicellular eukaryotic organism, the nematode C*aenorhabditis elegans* (97 Mb), was released [12]. In 2000, the first plant genome, *Arabidopsis thaliana* (115.4 Mb), was sequenced and decoded [13]. The largest genome project is the Human Genome Project (HGP), which was launched in October 1990 and successfully completed in April 2003, led by an international group [14]. The cost of genome sequencing and assembly decreased as Bacterial Artificial Chromosome (BAC) was replaced by Whole Genome Shotgun (WGS) strategies [15]. In the early stages of the genome project, several key eukaryotic genomes were successfully sequenced through the collaborative efforts of consortia or multiple organizations, requiring substantial time and resources. These included organisms such as model species (*Drosophila melanogaster* and *Takifugu rubripes*) and crops (rice, soybean, maize). Before 2010, all genome assembly efforts relied exclusively on Sanger sequencing technology, which remained costly. Next-generation sequencing (NGS), including sequencing by ligation (SBL) and sequencing by synthesis (SBS), offers an alternative option with greater efficiency and significantly reduced costs [16]. The first draft genome assembly using a short-reads-based next-generation sequencing (NGS) approach was achieved for cucumber, *Cucumis sativus* L., by combining Sanger and next-generation Illumina GA sequencing technologies at the Beijing Genomics Institute (BGI-Shenzhen) [17]. In 2010, the first draft genome of the giant panda (~2.25 Gb) was successfully assembled and completed at BGI-Shenzhen using only NGS technology [18]. Subsequently, a large number of plant and animal genomes were rapidly decoded at significantly reduced costs. The "Thousand Plant and Animal Genomes Project" was launched by BGI-Shenzhen in January 2010. In November 2010, the Genome 10 K Community of Scientists (G10KCOS) launched a project (Genome 10 K) in collaboration with BGI-Shenzhen, aiming to sequence the genomes of all 10,000 extant vertebrate species [19]. The Bird 10,000 Genomes (B10K) Project, for instance, is another project initiative aimed at generating high-quality draft genome sequences representative of all extant bird species [20]. More genome projects have been launched, including 1KMPG (1000 Medicinal Plant Genomes), i5k (Sequencing Five Thousand Arthropod Genomes), 10KP (10,000 Plant Genomes Project), et al. Furthermore, the Earth

BioGenome Project (EBP) aims to sequence life for the future of life. Finally, the Earth BioGenome Project (EBP), which aims to sequence all species on Earth for the future of life, represents a moonshot initiative in genomics [21].

3. The sequencing technology

Many genome projects are progressing rapidly. One of the most important factors is the rapid iteration and upgrade of sequencing technology, which offers higher throughput and lower costs. The traditional Sanger sequencing technology, utilizing the dideoxy chain termination method, played a crucial role in the early stages of genome assembly, including that of the human genome. However, the Sanger method is limited in genome assembly because of its relatively low throughput and high cost per base compared to high-throughput sequencing technologies. Sanger sequencing remains the gold standard sequencing technology due to its high accuracy, long read length capability, and flexibility to support various applications, particularly in PCR product and NGS-related validation.

NGS, also known as second-generation sequencing technology, encompasses platforms such as Roche 454, Illumina (Solexa), and ABI-SOLiD, and has revolutionized the field of omics. In 2005, the first NGS platform was Roche 454 GS-20, whose achievements rapidly captivated the scientific community. In 2006, the Illumina Solexa Genome Analyzer was released, followed by the ABI-SOLiD system in 2007. When the NGS systems became widely adopted, BGI-Shenzhen, which boasts the world's largest sequencing capacity, operated a diverse fleet of NGS systems, including 137 HiSeq 2000 units, 27 SOLiD systems, one Ion Torrent PGM, one MiSeq, and one 454 sequencer [22]. The different NGS systems, including SOLiD/Ion Torrent PGM (Life Sciences), Genome Analyzer/HiSeq 2000/MiSeq (Illumina), GS FLX Titanium/GS Junior (Roche), and Polonator G.007 (Complete genomics), were compared and summarized [22]. The market competition in the NGS field is extremely intense. To date, platforms such as 454 and SOLiD have all been withdrawn from the market. Illumina was once the absolute leader in the sequencing industry, holding more than 70% of the global market share and dominating the high-throughput sequencing market. The advancement and improvement of sequencing platforms are occurring at an extraordinarily rapid pace. Illumina, from acquiring Solexa's Genome Analyzer (1 Gb) in 2006, launched Genome Analyzer II (1.8-2.4 Gb) in 2007, and updated to Genome Analyzer IIx (6.5 Gb) in 2008. In 2010, HiSeq 2000 (200 Gb, PE 100) was introduced. In 2012, the HiSeq 2500 (Rapid model: 120 Gb, standard model: 600 Gb) was launched. In 2015, HiSeq 3000 (750 Gb) and HiSeq 4000 (1.5 Tb) were introduced. In 2014, HiSeq X Ten and HiSeq X Five were launched (one HiSeq X: Two flow cells: 1.6-1.8 Tb, One flow cell (800-900 Gb).

Additionally, in the low- and mid-throughput range with miniaturized sequencers, the MiSeq series, including MiSeq (2011), MiSeq Dx (2013), and MiSeq FGx (2015), has been progressively developed over the years. The other sequencers in the NextSeq series, including the NextSeq 500 (2014), NextSeq 550 (2015), NextSeq 1000 (2020), and NextSeq 2000 (2020), were designed for mid-throughput applications. MiniSeq (2016) was developed for targeted sequencing, small genome sequencing. The NovaSeq series offers ultra-high sequencing throughput. In 2017, NovaSeqTM included NovaSeq 5000 and NovaSeq 6000. The NovaSeq 6000 integrates throughput, flexibility, and ease of use into a single platform, expanding sequencing

capabilities from 80 to 6000 Gb. NovaSeq X Plus (2023) and NovaSeq X (2024) have significantly increased sequencing speed and throughput, thereby reducing sequencing costs.

In summary, Illumina has been highly successful in the field of next-generation sequencing (NGS) and has developed a wide range of sequencer series. However, some other NGS companies, such as Complete Genomics, which was purchased by BGI and with another name MGI Tech (MGI) launched a series of sequencing instruments. In 2015, MGI launched BGISEQ-500, which is a desktop high-throughput gene sequencer, achieving a throughput of 100 Gb. In 2017, the MGISEQ-200 (with the highest capacity of 150Gb) and the MGISEQ-2000 (with the highest capacity of 600Gb) were launched. In 2018, DNBSEQ-T7 was introduced as an ultra-high-throughput sequencer. A single chip can achieve Tb data output. The quadruple flow cell platform allows 1 to 4 chips to operate independently, with the capability of producing up to 6 Tb of data per day. A lot of other different sequencers, including MGISEQ-2000RS FluoXpert (2022), MGISEQ-200, DNBSEQ-G99, DNBSEQ-E25, DNBSEQ-E25 Flash, were facing the market over the years. In 2020, the DNBSEQ-T10 × 4 system, capable of generating 20 Tb of data, was introduced by MGI. In 2023, the DNBSEQ-T20 × 2 was introduced as a fully automated sequencer, integrating all components of a genetic sequencer through robotics. It is capable of processing six slides per run and supports both PE100 (42 Tb) and PE150 (72 Tb) sequencing modes. Recently, on February 24, 2025, the DNBSEQ-T1+ was introduced as one of the fastest T-level benchtop sequencers globally. It can generate up to 1.2 terabases (Tb) of sequencing data within 24 hours. These MGI sequencers achieve even higher throughput, lower costs, and faster turnaround times, thereby demonstrating exceptionally strong competitiveness in the market.

NGS has revolutionized the field of omics; however, its short read lengths have posed significant challenges for genome assembly, leading to fragmented and incomplete draft genomes [23]. Long-read sequencers, also known as third-generation sequencing (TGS) technologies, have been developed and demonstrated superior performance in generating high-quality genome assemblies by addressing the limitations of next-generation sequencing (NGS) [24]. Long-read sequencing technologies are capable of generating reads exceeding 10 kb in length and can even produce sequences spanning several Mb [25]. Three platforms, including Pacific Biosciences (PacBio), Oxford Nanopore Technology (ONT), and CycloneSEQ (a nanopore sequencing technology platform), have contributed to improvements in the integrity and accuracy of genome sequencing and assembly [24]. PacBio, also known as Single-Molecule Real-Time sequencing (SMRT), has two main modes: Continuous Long Reads (CLR) and Circular Consensus Sequencing (CCS), which generates high-fidelity (HiFi) reads. CLR models the typical length of 5–60 kb but the a lower accuracy (typically of 85–92%) [26, 27].

In 2009, PacBio launched the first long-read sequencer, the PacBio RS, which had a throughput of only 90 Mb. Then, in 2013, the throughput of PacBio RS II reached approximately 1 Gb. In 2015, the throughput of the new platform, Sequel, increased significantly to about 10 Gb. In 2019, PacBio introduced the Sequel II and Sequel IIe platforms, which enabled the generation of high-fidelity (HiFi) reads with up to 20 Gb of data per SMRT Cell. In 2023, PacBio Revio was launched, generating 360 GB of HiFi reads using 4 SMRT Cells. In 2024, PacBio launched the Vega, a desktop sequencing platform, which can generate up to 60 GB of data using each SMRT Cell. In 2005, Oxford Nanopore Technologies was founded to develop an electronic single-molecule sensing system based on nanopore science. The first nanopore

sequencing device, MinION, was used for early project planning in 2014 and formally commercialized in 2015 with a capacity of 10-25 Gb. Then, GridION was launched in 2017. In 2018, PromethION beta was launched, followed by PromethION 24 (with a capacity of 3.8 Tb) and PromethION 48 (with a capacity of 5–7.6 Tb) in 2019. The BGI Group officially launches another nanopore sequencing platform CycloneSEQ™. CycloneSEQ-WT02, which was launched on September 9, 2024, supports read lengths ranging from 100 bp to Mbp, with a maximum capacity of 50 Gb per cell. In 2025, the G400-ER was reported on March 1st, 2025, with a maximum capacity of 400 Gb per cell. Some other long-read sequencing platforms may also be laughed at. It is evident that the sequencing platform market is evolving rapidly, and the competition within this field is exceptionally intense.

With all the mentioned sequencing technologies, how should we choose the appropriate sequencing platform for genome assembly? Before the advent of long-read sequencing technologies, short-read sequencing platforms, such as Illumina and DNBSEQ, were used to generate draft genomes. Currently, most genomes are assembled using long reads because these enable the generation of high-quality genome assemblies. The long read includes PacBio Sequel/HiFi/Revio and Nanopore Oxford/Cyclone (**Figure 3**). With HiFi reads, the assembly is not only cost-effective and highly accurate but also eliminates the need for computing resources-intensive error correction based on short reads [28]. The Oxford Nanopore/Cyclone technology can obtain longer reads compared to PacBio, including ultra-long reads (>100 kb), enabling more effective resolution of complex regions in genomes [29]. The short reads can be used for error correction. 10x Genomics Linked Reads and single-tube long fragment read (stLFR) can obtain data from long DNA molecules using cost-effective second-generation sequencing technology. Both technologies can be used for scaffolding and phasing, as well as diploid genome assembly [30, 31]. For chromosome-level genome assembly, a genetic map has traditionally been used for scaffolding based on recombination rates between physical markers (**Figure 3**) [32].

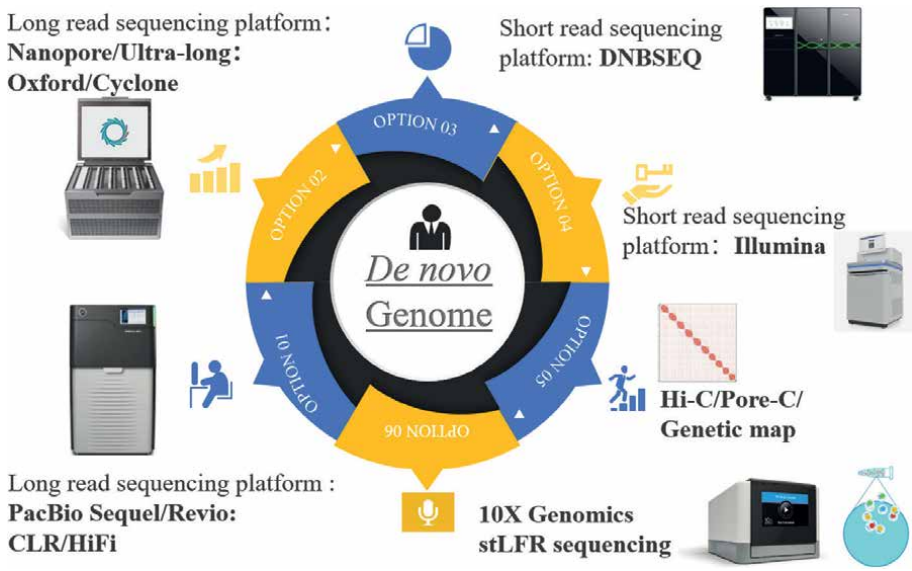

Figure 3.
Different sequencing platforms in de novo *genome assembly.*

However, many species fail to achieve sufficient family population sizes, which incur high costs, require extensive time investment, and result in a low anchoring success rate. Hi-C/Pore-C are economical methods for generating chromosome-scale genome assembly [33].

4. Genome assembly and algorithms

De novo genome assembly is the process of reconstructing the complete genome sequence from the reads generated by DNA sequencing technologies. The accuracy and completeness of genome assembly are primarily influenced by three critical factors: genome complexity (e.g., repetitive sequences, heterozygosity, polyploidy), sequencing technology (including read length, read accuracy, and sequencing coverage), and genome assembly algorithms (which are closely linked to computational resources and processing time). In the mentioned sections, sequencing technologies have revolutionized the genome era. The assembly software and algorithms were developed to process different types of data. OLC and DBG are two common assembly algorithms in genome assembly (**Table 1**).

4.1 Genome survey software

Before genome assembly, it is essential to conduct a preliminary survey of the genome characteristics to determine appropriate strategies. These strategies should address key considerations such as the selection of sequencing platforms, estimation of required data, and balancing cost-effectiveness. For genome size determination, flow cytometry remains the most conventional and widely adopted method because of its straightforward sample preparation process and high efficiency [34]. K-mer analysis based on sequencing has become increasingly popular in genome surveys, as it can be used to estimate genome size, repeat content, heterozygosity levels, and even polyploidy [35]. Jellyfish has been widely used for the fast and memory-efficient counting of k-mers [36]. The frequency of k-mers in a sequence can be efficiently determined using a hash table, which is commonly employed in the analysis of fundamental genomic characteristics, including genomic size estimation and calculation of heterozygosity rates. KMC is a disk-based program designed for efficient counting of k-mers [37].

Description	**OLC (Overlap-Layout-Consensus)**	**DBG (De Bruijn Graph)**
Data type compatibility	Long reads (PacBio/Nanopore/Sanger)	Short reads (Illumina)
Computational complexity	High ($O(N^2)$)	Low ($O(N)$)
Repeat region handling	Stronger (relies on read length to span repeats)	Weaker (short k-mers prone to repeat interference)
Assembly continuity	High (long contigs)	Lower (requires scaffolding)
Error tolerance	Depends on read quality (HiFi > CLR)	Error correction *via* k-mer frequency
Resource consumption	High memory and computational demand	Memory-efficient, scalable

Table 1.
The comparison of two common assembly algorithms.

Subsequent versions, including KMC 2 [38] and KMC 3 [39], introduced significant efficiency improvements. Genomic Character Estimator (GCE), developed by BGI-Shenzhen, is a tool for analyzing k-mer frequency in de novo genome projects. It features a user-friendly interface and detailed instructions, making its operation relatively convenient [40]. The k-mer abundance histograms or unique k-mers can then be utilized for genome survey analysis. KmerGenie is a tool that quickly generates histograms of k-mer abundances, thereby assisting in the determination of genome characteristics [41]. Both ntCard [42] and KmerStream [43] are streaming algorithms designed for cardinality estimation in high-throughput sequencing data. GenomeScope, another open-source web tool for genome survey, has been widely used to estimate the overall characteristics of a genome, such as genome size, heterozygosity rate, and repeat content [44]. The updated version, GenomeScope 2.0, provides a detailed mathematical model for resolving heterozygous and polyploid genomes based on k-mer frequencies [45]. Based on the heterozygous k-mer pairs, the Smudgeplot model in GenomeScope 2.0 is capable of visualizing and estimating the ploidy level as well as the structural composition of a genome. In summary, k-mer analysis is still a robust bioinformatics approach with extensive applications. However, interpreting the results can be complex and requires experience [46].

4.2 Genome assembly based on Sanger sequencing

Sanger sequencing reads have historically been employed for genome assembly. These reads typically range from 500 to 1000 base pairs in length and are characterized by their high accuracy. The assembly algorithm follows the Overlap-Layout-Consensus approach, which involves finding overlaps, laying out reads, and constructing a consensus sequence. PHRAP, which is based on the OLC method, can be utilized for the genome assembly of Sanger reads. It performs pairwise comparisons to identify overlapping regions between reads and constructs contigs accordingly [45]. Another assembly software, the Staden Package, is suitable for small-scale sequence assembly and analysis and is renowned for its reliability and flexibility [47]. Both Consed and Sequencher software are genome assembly editors designed to detect assembly errors, resolve conflicts, and enhance the overall quality of genome assemblies [48]. The FAKII program [49], the GAP4 program [50], and several other programs were developed. The CAP3 sequence assembly program, designed for BAC data, was used for assembly [51]. The Celera Assembler was utilized to generate a consensus sequence through the WGS approach [52]. These software options possess distinct advantages and limitations, and the selection is contingent upon factors such as the size of the genome being assembled, the quality of the data, and the availability of computational resources.

4.3 Genome assembly of short reads

Next-generation sequencing (NGS) is characterized by high throughput and relatively high cost. Another notable feature of NGS is that it generates short reads, typically ranging from 100 to 150 base pairs. With the introduction of novel assembly programs, NGS was rapidly and extensively adopted for many genome assembly projects. The OLC methods have also been implemented in several short-read assemblers, including Newbler [53], Fermi [54], CABOG [55], Edena [54], Shorty [56], Forge [57], SGA [58], and Readjoiner [59]. Due to the short length of reads, overlap strategies are a challenge for genome assembly. The novel assembly algorithm based on De

Bruijn Graphs was utilized. Since WGS assembly using short reads involves processing a large volume of data, the k-mer approach has been incorporated into most NGS assembly software. Different from the OLC method, the DBG methods utilize a k-mer graph or greedy graph algorithms. Several greedy-based software including SSAKE [60], SHARCGS [61], VCAKE [62], QSRA [63] were used for short read assembly. SSAKE is the first tool for assembling short sequence reads into contiguous sequences through a stringent assembly process.

In k-mer graphs, the "bubble" or alternate path must be addressed and resolved appropriately. A lot of new or revised assembly k-mer-based software packages, Velvet [64], Euler-SR [65], ALLPATHS-LG [66], SPAdes [67], ABySS [68], platanus [69], and SOAPdenovo [70], were utilized for the *de novo* assembly of NGS data [71]. The platanus software comprises three primary commands: assemble, scaffold, and gap_close [69]. SOAPdenovo has been widely used to assemble numerous large plant and animal genomes, as it generates assemblies with relatively long contigs and scaffolds. However, this process requires significant computational resources. To achieve a greater length of assembled contigs or scaffolds, libraries with varying insert sizes (e.g., 170, 500, 800 bp, 2, 5, 10, 20, and 40 kb) were utilized to obtain whole genome sequences for genome assembly. Small insert sizes (170, 500, and 800 bp) were utilized to assemble the contigs, while large insert sizes (2, 5, 10, 20, and 40 kb) were employed for scaffolding and subsequent gap filling [72].

4.4 Hybrid assembly using short reads and long reads

The short reads have enabled the cost-effective generation of many genomes. However, the resulting genome assemblies are often of draft quality and fragmented. With long-read sequencing, the throughput, quality, and read length are increasing while the cost is decreasing. In the third or fourth-generation genome, short reads were first used for the initial assembly, and long reads were subsequently employed for gap closing. Alternatively, a hybrid assembly approach combining both short and long reads was utilized. Unicycler, which employs a hybrid assembly approach, excels in assembling bacterial genomes with high accuracy and contiguity [73]. SMARTdenovo demonstrates excellent performance in handling complex genomes but requires higher computational resources [74]. With the advent of long-read sequencing, OLC methods have once again become more widely used. HybridSPAdes can generate accurate assemblies using an algorithm that integrates low-coverage long reads within a hybrid assembly framework [75].

For hybrid assembly, several software tools have been developed to fill gaps using long reads. LR_Gapcloser is a computational tool designed to fill sequence gaps by integrating a tiling path-based approach with long-read alignment and precise positioning algorithms [76]. TGS-GapCloser is a software tool that employs an algorithm to identify reads spanning gap regions between two contigs within a scaffold and assigns the best sequence data to close each gap [77]. DENTIST identifies reliable alignments of long reads, computes a consensus sequence for high base accuracy, and validates gap closure [78]. GMcloser with likelihood-based classifiers to fill gaps in assemblies [79]. WENGAN has been demonstrated to be an efficient tool for hybrid de novo assembly in humans [80].

4.5 *De novo* assembly with long reads and short-read error correction

Hybrid assembly serves as a transitional stage that follows long-read sequencing, which is characterized by higher throughput and cost-effectiveness. Then, most of the

genomes were assembled using long-read sequencing, with short-read error correction employed just as an assisting tool. Even when long-read data are of high quality (e.g., PacBio HiFi reads), the step of short-read error correction may be dispensable. For nanopore and PacBio CLR long reads, their error rates are higher compared to short reads. To improve assembly accuracy, it is recommended to apply short-read correction methods.

Due to the long-read sequencing data, the *de novo* assembly was performed mainly using the OLC algorithm in conjunction with WGS strategies. For nanopore and PacBio CLR long reads, several common assembly software include Canu [81], Falcon/FALCON-Unzip [82], Flye [83], Raven [84], Miniasm [85], MECAT [86], wtdbg2 [87], NECAT [88], and Nextdenovo [89]. Canu's algorithms include an adaptive overlapping strategy based on TF-IDF weighted MinHash and a sparse assembly graph construction that prevents the collapse of diverged repeats and haplotypes [81]. FALCON-based assemblies can identify widespread heterozygous structural variations and resolve the haplotype of a diploid genome [82]. Most of these long-read assemblers follow the Overlap-Layout-Consensus paradigm. The evaluation of different assemblers showed that no single assembler performs the best across all evaluation categories [90].

Error correction is necessary to address the inaccuracies present in the assembled sequencing using nanopore and PacBio CLR long-read sequencing data. LoRDEC, which is a k-mer-based approach, uses a combination of long reads and short reads to perform error correction [86]. RACON generates consensus sequences from overlapping reads [91]. Medaka (https://github.com/nanoporetech/medaka) is a neural network-based tool developed by Oxford Nanopore Technologies for creating consensus sequences and enhancing the accuracy of sequence assembly. The other tool, Minimum Error Correction (MEC), analyzes the k-mer frequency and constructs a graph for error correction [91]. GoldPolish-target, which focuses on specific target regions for polishing, offers a computationally lightweight approach for base error correction [92]. Other software, such as Pilon [93] and NextPolish [94], has been widely used for short-read error correction of assembled sequences. The error correction of long reads or genome assemblies requires significant computational resources.

For PacBio HiFi reads, error correction is unnecessary. With high-quality long reads, sequencing coverage of more than 30-fold is sufficient, unlike noisy long reads, which typically require over 60-fold coverage. HiFi reads are more continuous and complete, and they are also cost-effective [95]. The assembly algorithms are based on the OLC approach. Several software tools, including LJA [96], Flye [83], MBG [97], HiCanu [98], Nextdenovo, and hifiasm [99], have been developed for genome assembly or resolving haplotypes. Hifiasm is one of the top-performing PacBio HiFi assemblers, featuring updated frequency models and having released 53 versions to date. This software is a haplotype-resolved assembler designed for accurate HiFi reads and can use either nanopore or Hi-C data as input to assist in the assembly process.

4.6 T2T *de novo* assembly and pangenome

The telomere-to-telomere (T2T) genome era is emerging, where a T2T genome represents the complete genome assembly that has been successfully achieved for various organisms, including humans, rice, maize, soybean, and some non-model species [8]. The T2T assembly of large eukaryotic genomes was completed by integrating ultra-long reads (≥100 kb), HiFi reads, Hi-C data, and short reads for error

correction. In general, T2T assembly algorithms involve several key steps: error correction of nanopore reads, construction of an assembly graph, simplification of graphs using ultra-long reads, and phasing and scaffolding aided by Hi-C data [99]. A number of cases were used in the T2T genome, as follows: The primary contigs were assembled using HiFi reads, scaffolded with Hi-C data, and further refined with ultra-long nanopore reads for gap closure and resolution of complex regions, including telomeres. Short reads were utilized for error correction.

Currently, only two assemblers, HiFiasm [99] and Verkko [100], can integrate HiFi reads and ultra-long reads to achieve T2T genome assembly. The T2T genome has advanced rapidly, including the assembly of complex polyploid genomes such as cultivated peanut [101], common wheat [102]. The T2T genome assemblies can revolutionize our understanding of repetitive elements and rDNA in gene regulation, pan-genomics, and functional genomics [103]. What is more, a single genome for one species is insufficient. The pangenome refers to the comprehensive set of genetic information encompassed by representative individuals within that species. The algorithms for pangenome construction involve building a variation graph through all-to-all alignments. Several software tools, such as PanGenome Graph Builder [104], Minigraph-Cactus [105], and PanGenome Research Tool Kit (PGR-TK) [106], have been developed for pangenome construction. Some download software pangenome analysis pipeline (PSVCP) was also developed for the next step bioinformatics analysis tools [107].

4.7 Genome assembly evaluation

After *de novo* assembly, the evaluation involves assessing core metrics to determine the quality of the assembled genome. Accuracy, contiguity, and completeness were the three primary criteria for evaluating the assembly. The contiguity evaluation included key metrics such as contig N50 and scaffold N50, the length of the longest contig and scaffold, as well as the total number of contigs and scaffolds. The read mapping rate, coverage, and potential contamination can also serve as assessment criteria.

Benchmarking Universal Single-Copy Orthologs (BUSCO) is one of the most widely used tools for assessing genome completeness based on conserved gene sets [108]. In actuality, for different species with varying genome sizes, relying solely on contig or scaffold length is insufficient. Some pipeline tools, such as Genome Assembly Evaluating Pipeline (GAEP), were developed for assessing the quality, including detecting misassemblies [109]. GenomeQC is another tool that is easy to use and integrates various quantitative measures for genome assemblies and annotations [110]. Merqury can estimate base-level accuracy and completeness without a reference genome using efficient k-mer analysis [111]. Genome Continuity Inspector (GCI) is a tool designed for evaluating genome continuity at both the single-base and T2T levels [112]. Some other tools, including QUAST [113] and REAPR [114], are also used for quality assessment. EvalDNA is a machine learning-based tool designed for evaluating mammalian genome assemblies [115]. For the assessment of completeness, repeats are complex in plant genomes, making them a crucial factor for evaluating genome quality. The LTR Assembly Index (LAI) estimates the percentage of intact LTR retroelements, gauging completeness in repetitive genomic regions [116]. In the future, telomeres, rDNA, and centromeres may also be utilized for evaluation in T2T genomes.

5. Summary

Genome algorithms and software have been updated across different sequencing stages. Five distinct phases can be identified based on differing sequencing strategies, each with associated costs and specific computational requirements. The core genome assembly algorithms are the Overlap-Layout-Consensus (OLC) and De Bruijn Graph (DBG) algorithms. However, various solution strategies and software have been developed based on these algorithms. Some tools are user-friendly, some are resource-efficient, and others produce higher-quality assemblies. It is difficult to determine which one is the best based on real data testing, as different species and sequencing technologies have specific characteristics. In different stages of the genome, the advantages and limitations of assembly algorithms and software were compared (**Table 2**). In the future, machine learning algorithms may be able to solve assembly problems more effectively [117]. Artificial intelligence (AI)-based DNA assemblers, particularly for human or species with a reference genome, could significantly improve accuracy. By training on known sequences, the genomes of other individuals can be assembled more easily and accurately predicted.

Generation	Assembly strategy	Core algorithm	Representative software	Key advantages	Limitations
First-generation	Sanger reads assembly	OLC	PHRAP, CAP3, Celera Assembler	Ultra-high accuracy (>99.9%), simple workflow	Low throughput, limited to small genomes
Second-generation	NGS short-read assembly	DBG	SPAdes, SOAPdenovo, Velvet	Cost-effective, scalable for large genomes	Fragmented contigs, fail in repetitive regions
Third-generation	Hybrid assembly: NGS short reads assembly, PacBio CLR long reads gap filling	DBG-based hybrid	PBJelly, MaSuRCA, Unicycler	Improved continuity, cost-efficient	Dependent on NGS scaffold quality, gaps in complex repeats
Fourth-generation	Hybrid assembly: PacBio CLR/ ONT long reads assembly, polished with NGS short reads	OLC-based hybrid	Falcon, Canu, Flye	Resolves repeats, long contiguous contigs	High computational cost, requires polishing
T2T generation	Integration of HiFi (high accuracy), ONT ultra-long reads (>100 kb), and HiC data, phased assembly + manual curation	Multi-strategy (OLC + graph-based)	Hifiasm, HiCanu, Verkko, 3D-DNA	Telomere-to-telomere completeness resolves centromeres	High cost, complex workflow

Table 2.
The comparison of assembly algorithm and software at different stages.

DOI: http://dx.doi.org/10.5772/intechopen.1011544

Author details

Jianbo Jian[1,2*], Qiang Gao[3], Jun Cheng[1] and Ye Yin[3]

1 BGI Genomics, Shenzhen, China

2 Guangdong Provincial Key Laboratory of Marine Biotechnology, Shantou University, Shantou, China

3 BGI, Shenzhen, China

*Address all correspondence to: jianjianbo@bgi.com

IntechOpen

References

[1] Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal. 2020;**18**:9-19. DOI: 10.1016/j.csbj.2019.11.002

[2] Shendure J, Findlay GM, Snyder MW. Genomic medicine-progress, pitfalls, and promise. Cell. 2019;**177**(1):45-57. DOI: 10.1016/j.cell.2019.02.003

[3] Huang X, Huang S, Han B, Li J. The integrated genomics of crop domestication and breeding. Cell. 2022;**185**(15):2828-2839. DOI: 10.1016/j.cell.2022.04.036

[4] Noll T, Pühler A, Weisshaar B, Wendisch VF. The genomics revolution and its impact on future biotechnology. Journal of Biotechnology. 2014;**190**:1. DOI: 10.1016/j.jbiotec.2014.10.009

[5] Onoyama T, Ishikawa S, Isomoto H. Gastric cancer and genomics: Review of literature. Journal of Gastroenterology. 2022;**57**(8):505-516. DOI: 10.1007/s00535-022-01879-3

[6] Thudi M, Palakurthi R, Schnable JC, Chitikineni A, Dreisigacker S, Mace E, et al. Genomic resources in plant breeding for sustainable agriculture. Journal of Plant Physiology. 2021;**257**:153351. DOI: 10.1016/j.jplph.2020.153351

[7] Theissinger K, Fernandes C, Formenti G, Bista I, Berg PR, Bleidorn C, et al. How genomics can help biodiversity conservation. Trends in Genetics. 2023;**39**(7):545-559. DOI: 10.1016/j.tig.2023.01.005

[8] Li H, Durbin R. Genome assembly in the telomere-to-telomere era. Nature Reviews Genetics. 2024;**25**(9):658-670. DOI: 10.1038/s41576-024-00718-w

[9] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977;**265**(5596):687-695. DOI: 10.1038/265687a0

[10] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;**269**(5223):496-512. DOI: 10.1126/science.7542800

[11] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996;**274**(5287):546, 63-7. DOI: 10.1126/science.274.5287.546

[12] C. elegans Sequencing Consortium. Genome sequence of the nematode *C. Elegans*: A platform for investigating biology. Science. 1998;**282**(5396):2012-2018. DOI: 10.1126/science.282.5396.2012

[13] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;**408**(6814):796-815. DOI: 10.1038/35048692

[14] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;**409**(6822):860-921. DOI: 10.1038/35057062

[15] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;**291**(5507):1304-1351. DOI: 10.1126/science.1058040

[16] Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;**17**(6):333-351. DOI: 10.1038/nrg.2016.49

[17] Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, et al. The genome of the cucumber, *Cucumis sativus* L. Nature Genetics. 2009;**41**(12):1275-1281. DOI: 10.1038/ng.475

[18] Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of the giant panda genome. Nature. 2010;**463**(7279):311-317. DOI: 10.1038/nature08696

[19] Genome 10K Community of Scientists. Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. The Journal of Heredity. 2009;**100**(6):659-674. DOI: 10.1093/jhered/esp086

[20] Zhang G, Rahbek C, Graves GR, Lei F, Jarvis ED, Gilbert MT. Genomics: Bird sequencing project takes off. Nature. 2015;**522**(7554):34. DOI: 10.1038/522034d

[21] Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences of the United States of America. 2018;**115**(17):4325-4333. DOI: 10.1073/pnas.1720115115

[22] Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. Journal of Biomedicine & Biotechnology. 2012;**2012**:251364. DOI: 10.1155/2012/251364

[23] Liao X, Li M, Zou Y, Wu F-X, Yi-Pan and Wang J. Current challenges and solutions of assembly. Quantitative Biology. 2019;**7**(2):90-109. DOI: 10.1007/s40484-019-0166-9

[24] Espinosa E, Bautista R, Larrosa R, Plata O. Advancements in long-read genome sequencing technologies and algorithms. Genomics. 2024;**116**(3):110842. DOI: 10.1016/j.ygeno.2024.110842

[25] Payne A, Holmes N, Rakyan V, Loose M. Bulk V is: A graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. 2019;**35**(13):2193-2198. DOI: 10.1093/bioinformatics/bty841

[26] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;**323**(5910):133-138

[27] Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nature Reviews Genetics. 2020;**21**(10):597-614

[28] Walden KKO, Cao Y, Fields CJ, Hernandez AG, Rendon GA, Robinson GE, et al. High-quality genome assemblies for nine non-model north American insect species representing six orders (Insecta: Coleoptera, Diptera, Hemiptera, hymenoptera, Lepidoptera, Neuroptera). Molecular Ecology Resources. 2024;**24**(8):e14010. DOI: 10.1111/1755-0998.14010

[29] Cuber P, Chooneea D, Geeves C, Salatino S, Creedy TJ, Griffin C, et al. Comparing the accuracy and efficiency of third generation sequencing technologies, Oxford Nanopore technologies, and Pacific biosciences, for DNA barcode sequencing applications. Ecological Genetics and Genomics. 2023;**28**:100181. DOI: 10.1016/j.egg.2023.100181

[30] Wang O, Chin R, Cheng X, Wu MKY, Mao Q, Tang J, et al. Efficient and unique

cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Research. 2019;**29**(5):798-808. DOI: 10.1101/gr.245126.118

[31] Zhang L, Zhou X, Weng Z, Sidow A. Assessment of human diploid genome assembly with 10x linked-reads data. GigaScience. 2019;**8**:11. DOI: 10.1093/gigascience/giz141

[32] Fierst JL. Using linkage maps to correct and scaffold de novo genome assemblies: Methods, challenges, and computational tools. Frontiers in Genetics. 2015;**6**:220. DOI: 10.3389/fgene.2015.00220

[33] Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Computational Biology. 2019;**15**(8):e1007273. DOI: 10.1371/journal.pcbi.1007273

[34] Pellicer J, Powell RF, Leitch IJ. The application of flow cytometry for estimating genome size, ploidy level Endopolyploidy, and reproductive modes in plants. Methods in Molecular Biology (Clifton, NJ). 2021;**2222**:325-361. DOI: 10.1007/978-1-0716-0997-2_17

[35] Pflug JM, Holmes VR, Burrus C, Johnston JS, Maddison DR. Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in beetles (Coleoptera). G3 (Bethesda, Md). 2020;**10**(9):3047-3060. DOI: 10.1534/g3.120.401028

[36] Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;**27**(6):764-770. DOI: 10.1093/bioinformatics/btr011

[37] Deorowicz S, Debudaj-Grabysz A, Grabowski S. Disk-based k-mer counting on a PC. BMC Bioinformatics. 2013;**14**(1):160. DOI: 10.1186/1471-2105-14-160

[38] Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A. KMC 2: Fast and resource-frugal k-mer counting. Bioinformatics. 2015;**31**(10):1569-1576. DOI: 10.1093/bioinformatics/btv022

[39] Kokot M, Dlugosz M, Deorowicz S. KMC 3: Counting and manipulating k-mer statistics. Bioinformatics. 2017;**33**(17):2759-2761. DOI: 10.1093/bioinformatics/btx304

[40] Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv:1308.2012. DOI: 10.48550/arXiv.1308.2012

[41] Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;**30**(1):31-37. DOI: 10.1093/bioinformatics/btt310

[42] Mohamadi H, Khan H, Birol I. ntCard: A streaming algorithm for cardinality estimation in genomics data. Bioinformatics. 2017;**33**(9):1324-1330. DOI: 10.1093/bioinformatics/btw832

[43] Melsted P, Halldórsson BV. KmerStream: Streaming algorithms for k-mer abundance estimation. Bioinformatics. 2014;**30**(24):3541-3547. DOI: 10.1093/bioinformatics/btu713

[44] Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics.

2017;**33**(14):2202-2204. DOI: 10.1093/bioinformatics/btx153

[45] Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature Communications. 2020;**11**(1):1432. DOI: 10.1038/s41467-020-14998-3

[46] Hesse U. K-Mer-based genome size estimation in theory and practice. Methods in Molecular Biology (Clifton, NJ). 2023;**2672**:79-113. DOI: 10.1007/978-1-0716-3226-0_4

[47] Staden R. The Staden sequence analysis package. Molecular Biotechnology. 1996;**5**(3):233-241. DOI: 10.1007/bf02900361

[48] Gordon D, Abajian C, Green P. Consed: A graphical tool for sequence finishing. Genome Research. 1998;**8**(3):195-202. DOI: 10.1101/gr.8.3.195

[49] Larson S, Jain M, Anson EL, Myers EW. An Interface for a Fragment Assembly Kernel. ACM Digital Library; 1996

[50] Bonfield JK, Smith K, Staden R. A new DNA sequence assembly program. Nucleic Acids Research. 1995;**23**(24):4992-4999. DOI: 10.1093/nar/23.24.4992

[51] Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Research. 1999;**9**(9):868-877. DOI: 10.1101/gr.9.9.868

[52] Denisov G, Walenz B, Halpern AL, Miller J, Axelrod N, Levy S, et al. Consensus generation and variant detection by Celera assembler. Bioinformatics. 2008;**24**(8):1035-1040. DOI: 10.1093/bioinformatics/btn074

[53] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;**437**(7057):376-380. DOI: 10.1038/nature03959

[54] Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. Genome Research. 2008;**18**(5):802-809. DOI: 10.1101/gr.072033.107

[55] Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008;**24**(24):2818-2824. DOI: 10.1093/bioinformatics/btn548

[56] Hossain MS, Azimi N, Skiena S. Crystallizing short-read assemblies around seeds. BMC Bioinformatics. 2009;**10**(Suppl. 1):S16. DOI: 10.1186/1471-2105-10-s1-s16

[57] Diguistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK, et al. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biology. 2009;**10**(9):R94. DOI: 10.1186/gb-2009-10-9-r94

[58] Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Research. 2012;**22**(3):549-556. DOI: 10.1101/gr.126953.111

[59] Gonnella G, Kurtz S. Readjoiner: A fast and memory efficient string graph-based sequence assembler. BMC Bioinformatics. 2012;**13**:82. DOI: 10.1186/1471-2105-13-82

[60] Warren RL, Sutton GG, Jones SJ, Holt RA. Assembling millions of short DNA sequences using SSAKE. Bioinformatics. 2007;**23**(4):500-501. DOI: 10.1093/bioinformatics/btl629

[61] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Research. 2007;**17**(11):1697-1706. DOI: 10.1101/gr.6435207

[62] Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, et al. Extending assembly of short DNA sequences to handle error. Bioinformatics. 2007;**23**(21):2942-2944. DOI: 10.1093/bioinformatics/btm451

[63] Bryant DW Jr, Wong WK, Mockler TC. QSRA: A quality-value guided de novo short read assembler. BMC Bioinformatics. 2009;**10**:69. DOI: 10.1186/1471-2105-10-69

[64] Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 2008;**18**(5):821-829. DOI: 10.1101/gr.074492.107

[65] Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. Genome Research. 2008;**18**(2):324-330. DOI: 10.1101/gr.7088808

[66] Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Research. 2008;**18**(5):810-820. DOI: 10.1101/gr.7337908

[67] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology. 2012;**19**(5):455-477. DOI: 10.1089/cmb.2012.0021

[68] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Research. 2009;**19**(6):1117-1123. DOI: 10.1101/gr.089532.108

[69] Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research. 2014;**24**(8):1384-1395. DOI: 10.1101/gr.170720.113

[70] Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Research. 2010;**20**(2):265-272. DOI: 10.1101/gr.097261.109

[71] El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-generation sequence assembly: Four stages of data processing and computational challenges. PLoS Computational Biology. 2013;**9**(12):e1003345. DOI: 10.1371/journal.pcbi.1003345

[72] Vasilinetc I, Prjibelski AD, Gurevich A, Korobeynikov A, Pevzner PA. Assembling short reads from jumping libraries with large insert sizes. Bioinformatics. 2015;**31**(20):3262-3268. DOI: 10.1093/bioinformatics/btv337

[73] Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Computational Biology. 2017;**13**(6):e1005595. DOI: 10.1371/journal.pcbi.1005595

[74] Liu H, Wu S, Li A, Ruan J. SMARTdenovo: A de novo assembler using long noisy reads. GigaByte. 2021;**2021**:gigabyte15. DOI: 10.46471/gigabyte.15

[75] Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes:

An algorithm for hybrid assembly of short and long reads. Bioinformatics. 2016;**32**(7):1009-1015. DOI: 10.1093/bioinformatics/btv688

[76] Xu GC, Xu TJ, Zhu R, Zhang Y, Li SQ, Wang HW, et al. LR_Gapcloser: A tiling path-based gap closer that uses long reads to complete genome assembly. GigaScience. 2019;**8**(1):1-14. DOI: 10.1093/gigascience/giy157

[77] Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. GigaScience. 2020;**9**(9):1-11. DOI: 10.1093/gigascience/giaa094

[78] Ludwig A, Pippel M, Myers G, Hiller M. DENTIST-using long reads for closing assembly gaps at high accuracy. GigaScience. 2022;**11**:1-12. DOI: 10.1093/gigascience/giab100

[79] Kosugi S, Hirakawa H, Tabata S. GMcloser: Closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. Bioinformatics. 2015;**31**(23):3733-3741. DOI: 10.1093/bioinformatics/btv465

[80] Di Genova A, Buena-Atienza E, Ossowski S, Sagot MF. Efficient hybrid de novo assembly of human genomes with WENGAN. Nature Biotechnology. 2021;**39**(4):422-430. DOI: 10.1038/s41587-020-00747-w

[81] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. 2017;**27**(5):722-736. DOI: 10.1101/gr.215087.116

[82] Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 2016;**13**(12):1050-1054. DOI: 10.1038/nmeth.4035

[83] Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology. 2019;**37**(5):540-546. DOI: 10.1038/s41587-019-0072-8

[84] Vaser R, Šikić M. Time- and memory-efficient genome assembly with raven. Nature Computational Science. 2021;**1**(5):332-336. DOI: 10.1038/s43588-021-00073-4

[85] Li H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;**32**(14):2103-2110. DOI: 10.1093/bioinformatics/btw152

[86] Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, et al. MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nature Methods. 2017;**14**(11):1072-1074. DOI: 10.1038/nmeth.4432

[87] Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nature Methods. 2020;**17**(2):155-158. DOI: 10.1038/s41592-019-0669-3

[88] Chen Y, Nie F, Xie SQ, Zheng YF, Dai Q, Bray T, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nature Communications. 2021;**12**(1):60. DOI: 10.1038/s41467-020-20236-7

[89] Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, et al. NextDenovo: An efficient error correction and accurate assembly tool for noisy long reads. Genome Biology. 2024;**25**(1):107. DOI: 10.1186/s13059-024-03252-4

[90] Cosma BM, Shirali Hossein Zade R, Jordan EN, van Lent P, Peng C,

Pillay S, et al. Evaluating long-read de novo assembly tools for eukaryotic genomes: Insights and considerations. GigaScience. 2022;**12**:1-12. DOI: 10.1093/gigascience/giad100

[91] Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research. 2017;**27**(5):737-746. DOI: 10.1101/gr.214270.116

[92] Zhang E, Coombe L, Wong J, Warren RL, Birol I. GoldPolish-target: Targeted long-read genome assembly polishing. BMC Bioinformatics. 2025;**26**(1):78. DOI: 10.1186/s12859-025-06091-7

[93] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;**9**(11):e112963. DOI: 10.1371/journal.pone.0112963

[94] Hu J, Fan J, Sun Z, Liu S. NextPolish: A fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020;**36**(7):2253-2255. DOI: 10.1093/bioinformatics/btz891

[95] Espinosa E, Bautista R, Fernandez I, Larrosa R, Zapata EL, Plata O. Comparing assembly strategies for third-generation sequencing technologies across different genomes. Genomics. 2023;**115**(5):110700. DOI: 10.1016/j.ygeno.2023.110700

[96] Bankevich A, Bzikadze AV, Kolmogorov M, Antipov D, Pevzner PA. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. Nature Biotechnology. 2022;**40**(7):1075-1081. DOI: 10.1038/s41587-022-01220-6

[97] Rautiainen M, Marschall T. MBG: Minimizer-based sparse de Bruijn graph construction. Bioinformatics. 2021;**37**(16):2476-2478. DOI: 10.1093/bioinformatics/btab004

[98] Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Research. 2020;**30**(9):1291-1305. DOI: 10.1101/gr.263566.120

[99] Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods. 2021;**18**(2):170-175. DOI: 10.1038/s41592-020-01056-5

[100] Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nature Biotechnology. 2023;**41**(10):1474-1482. DOI: 10.1038/s41587-023-01662-6

[101] Wang X, Sun Z, Qi F, Zhou Z, Du P, Shi L, et al. A telomere-to-telomere genome assembly of the cultivated peanut. Molecular Plant. 2025;**18**(1):5-8. DOI: 10.1016/j.molp.2024.12.001

[102] Liu S, Li K, Dai X, Qin G, Lu D, Gao Z, et al. A telomere-to-telomere genome assembly coupled with multi-omic data provides insights into the evolution of hexaploid bread wheat. Nature Genetics. 2025;**57**(4):1008-1020. DOI: 10.1038/s41588-025-02137-x

[103] Garg V, Bohra A, Mascher M, Spannagl M, Xu X, Bevan MW, et al. Unlocking plant genetics with telomere-to-telomere genome assemblies. Nature Genetics. 2024;**56**(9):1788-1799. DOI: 10.1038/s41588-024-01830-7

[104] Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al. Building pangenome graphs. Nature

Methods. 2024;**21**(11):2008-2012. DOI: 10.1038/s41592-024-02430-3

[105] Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph construction from genome alignments with Minigraph-cactus. Nature Biotechnology. 2024;**42**(4):663-673. DOI: 10.1038/s41587-023-01793-w

[106] Chin C-S, Behera S, Khalak A, Sedlazeck FJ, Sudmant PH, Wagner J, et al. Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. Nature Methods. 2023;**20**(8):1213-1221. DOI: 10.1038/s41592-023-01914-y

[107] Wang J, Yang W, Zhang S, Hu H, Yuan Y, Dong J, et al. A pangenome analysis pipeline provides insights into functional gene identification in rice. Genome Biology. 2023;**24**(1):19. DOI: 10.1186/s13059-023-02861-9

[108] Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular Biology and Evolution. 2021;**38**(10):4647-4654. DOI: 10.1093/molbev/msab199

[109] Zhang Y, Lu HW, Ruan J. GAEP: A comprehensive genome assembly evaluating pipeline. Journal of Genetics and Genomics = Yi chuan xue bao. 2023;**50**(10):747-754. DOI: 10.1016/j.jgg.2023.05.009

[110] Manchanda N, Portwood JL, Woodhouse MR, Seetharam AS, Lawrence-Dill CJ, Andorf CM, et al. Genome QC: A quality assessment tool for genome assemblies and gene structure annotations. BMC Genomics. 2020;**21**(1):193. DOI: 10.1186/s12864-020-6568-2

[111] Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology. 2020;**21**(1):245. DOI: 10.1186/s13059-020-02134-9

[112] Chen Q, Yang C, Zhang G, Wu D. GCI: A continuity inspector for complete genome assembly. Bioinformatics. 2024;**40**(11):1-10. DOI: 10.1093/bioinformatics/btae633

[113] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;**29**(8):1072-1075. DOI: 10.1093/bioinformatics/btt086

[114] Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: A universal tool for genome assembly evaluation. Genome Biology. 2013;**14**(5):R47. DOI: 10.1186/gb-2013-14-5-r47

[115] MacDonald ML, Lee KH. EvalDNA: A machine learning-based tool for the comprehensive evaluation of mammalian genome assembly quality. BMC Bioinformatics. 2021;**22**(1):570. DOI: 10.1186/s12859-021-04480-2

[116] Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR assembly index (LAI). Nucleic Acids Research. 2018;**46**(21):e126. DOI: 10.1093/nar/gky730

[117] Padovani de Souza K, Setubal JC, de Leon FdCAC P, Oliveira G, Chateau A, Alves R. Machine learning meets genome assembly. Briefings in Bioinformatics. 2019;**20**(6):2116-2129. DOI: 10.1093/bib/bby072

Chapter 3

Multi-Omics in Biological System: Harnessing the Power of Multi-Omics-Applications, Challenges and the Path Forward

Kinjalben Suthar and Nisha Singh

Abstract

The study of complex biological systems is undergoing a transformation through multi-omics integration, which offers a comprehensive insight into the interplay between molecular layers such as genomics, transcriptomics, proteomics, and metabolomics. This chapter provides an introduction to the field of multi-omics, defining its scope and highlighting the growing need for integrating diverse omics data to achieve a holistic understanding of biological phenomena. The chapter goes over the types of omics data and their significance, starting from single-cell multi-omics data, which offer unprecedented resolution to examine cellular heterogeneity. Data integration is elaborated further and discussed concerning difficulties in data source combination with respect to scalability concerns. Some of the approaches and computational workflows in the integration of multi-omics data are provided with an overview of the tools and databases used to support the same. Real applications of multi-omics are described and explained across disciplines that include the search for biomarkers in disease discovery, personalised medicine, agriculture, and plant sciences, as well as in systems biology. Finally, technical, computational, and analytical challenges arising from research work in multi-omics and suggested solutions to the challenges will be described, after which this chapter will outline potential future directions in multi-omics that would evolve its full and transformative impact for opening up new frontiers of biological science discovery.

Keywords: computational tools, data integration, multi-omics integration, single-cell omics, system biology

1. Introduction

According to ancient scriptures, the entire universe is said to have been created by Lord Brahma, and he was the first to utter the sacred word "Om". Composed of three Sanskrit letters-aa, au, and ma-these sounds blended together to form "Aum" or "Om", representing completeness and a divine presence that pervades the entire universe.

In modern science, the term omics comes from the suffix -ome, which represents totality or completeness. The term "genomics" originated in a 1986 conversation between Dr. Thomas H. Roderick, who was proposing the journal *Genomics*. The term genomics came from the earlier term genome that Hans Winkler introduced in 1920 for the haploid set of chromosomes. While the suffix -ome lacks a confirmed Greek origin, scholars speculate that -ome may represent wholeness, based on terms such as chromosome. The holistic nature of omics reflects the comprehensive study of biological systems [1].

In an analogous manner, "multi-omics" is a combination of "multi", which means more than one, and "omics", a suffix used in the life sciences to signify large-scale studies aimed at understanding complex biological systems. In order to have a thorough understanding of biological processes, "multi-omics" refers to the integration of several large-scale datasets. A variety of bioinformatics tools have been developed which encompasses the synergistic analysis of genomes, transcriptomics, proteomics, and metabolomics, among other omics domains, to provide a comprehensive understanding of complex biological systems [2]. These omics layers each provide distinct perspectives on molecular and cellular mechanisms that together influence phenotypic outcomes.

Multi-omics integration covers domains ranging from biotechnology to environmental sciences, agriculture, and personalised medicine. It allows researchers to identify new biological pathways, discover biomarkers of disease, and predict how organisms will respond to changes in their environment by linking multiple omics layers. Shift from studying the individual elements to examining the complex relationship is very progressive for both basic and applied life sciences. All these areas have benefited much from multi-omics understanding of the influence of microbiome in health and illness, as well as to increase crop immunity against unfavourable conditions.

Domains, such as biotechnology, environmental sciences, agriculture, and personalised medicine, are all covered by multi-omics integration. Researchers are discovering new biological pathways, finding disease biomarkers, and forecasting how organisms will be affected by environmental changes by connecting several omics layers. Just by shifting the focus from examining individual elements to analysing the complex relationship, this approach is very progressive in both basic and applied life sciences. Understanding the dynamics of the microbiome in health and illness, increasing crop yields through stress-resistant cultivars, and clarifying cancer heterogeneity have all benefited greatly from multi-omics. Multi-omics has wide opportunity in the future of systems biology and will serve as a foundation for groundbreaking discoveries [3].

This chapter will give insights multi-omics role in biological systems and its applications in various domains of biology. The available tools, techniques, and databases will help to address specific biological questions, challenges related to multi-omics integrations, and different approaches in multi-omics data integration. Also, we will discuss about current evolving technologies such as single-cell multi-omics, spatial multi-omics, and the role of AI-ML. and future prospect of multi-omics.

1.1 Types of omics data

Omics data can be broadly categorised into four major types, each representing a distinct molecular layer.

A. Genomics: the study of an organism's entire genetic material, genomics gives insights into the genetic variations, mutations, and structural features. Next-generation sequencing (NGS) has revolutionised genomics, enabling the identification of mutations and evolutionary patterns [4].

B. Transcriptomics: transcriptome deciphers gene expression patterns under various conditions, focusing on RNA molecules. The RNA sequencing technologies (RNA-seq) reveal how genes are regulated and how they respond to environmental or pathological stimuli.

C. Proteomics: proteins are the functional executors of cellular processes; they represent a critical layer of information. It is a comprehensive study of proteins within a cell, tissue, or organism, including their abundance, modifications, functions, and interactions. Mass spectrometry (MS)-based approaches are used extensively in proteomics, providing data on protein abundance, modifications, and interactions [5].

D. Metabolomics: metabolites, small molecules that represent the end product of cellular processes. Metabolomics employing techniques such as nuclear magnetic resonance (NMR) and gas chromatography-mass spectrometry (GC-MS) uncovers metabolomics and reveals metabolic pathways and their alterations in response to stress or disease [6].

Emerging disciplines like epigenomics, lipidomics, and glycomics supplement the integration of these omics layers by contributing more layers to the study of biological systems. Together, these datasets provide a comprehensive understanding of biological processes and establish the groundwork for an investigation at the system level. The challenging part is harmonising these diverse datasets to extract meaningful insights, it calls for advanced computational techniques and interdisciplinary knowledge [7].

1.2 Need for multi-omics integration

The biological system is very complex and demands an extensive and integrative approach that transcends single-omics studies [8]. The dynamic and complex regulation of genes and their functional outputs are not captured by genomics despite the fact that it offers the blueprint of life. Similarly, each of the omics layers—transcriptome, proteome, and metabolome—captures a distinct aspect of the biological jigsaw but is incapable of doing so in isolation. This limitation can be well addressed by multi-omics integration. It combines diverse datasets and enables the reconstruction of intricate biological pathways and networks [9].

In clinical settings, the requirement for multi-omics integration is evident in precision medicine. Cancer is a multifactorial disease influenced by genetic mutations, epigenetic modifications, transcriptomic alterations, and metabolic reprogramming. These data type integrations can reveal biomarkers for early diagnosis and patient-specific treatment strategies [10].

Plants are a very diverse and complex class; there are a lot of variations that can be observed not only from one plant to another plant but also in the same species. The implication of only one omics approach also gives insights into the plant biological system; to understand the holistic view of plant function, multi-omics integration can help [11].

When the gene is inserted into the genome of the biological system to improve the yield or to develop resistance, however, significant changes can be observed in the gene expression; many factors play a part, such as epigenetic modification due to stress, insufficient nutrient availability or cofactor needed for product formation, suboptimal culture conditions, protein aggregation etc. multi-omics integration can provide a holistic view to understand this processes. Multi-omics has great potential in understanding global challenges, including food security and climate change. By integrating genomics and environmental data, researchers can develop crop resilient to stressors and understand the economics dynamics. It does not only bridge the gap between data and actionable insights but also fosters inter disciplinary collaborations driving innovation across fields.

2. Single-cell multi-omics

The single-cell multi-omics integrates various molecular layers derived from individual cells, such as transcriptomics, proteomics, epigenomics, and genomics. This provides a much deeper understanding of biological processes such as development, ageing, and the progression of disease [12]. By offering a complete view of cellular heterogeneity, it helps provide an overall understanding of the various biological processes involved. These technologies have demonstrated significant potential in identifying the complex molecular interactions and cellular networks that characterise states of phenotype. However, there are still a few challenges. From a technical perspective, the small quantities of biological material that are obtained from individual cells demand sensitive and accurate techniques for detection and amplification that sometimes cause inefficiencies [13]. Integrating data from many omic levels is also computationally challenging and requires complex methods to filter out biological signals from noise. The high cost of data storage, analysis, and sequencing also poses a problem [14].

Cell isolation is the first step, using methods like laser capture microdissection, microfluidics, and fluorescence-activated cell sorting (FACS) in order to save phenomenological and spatial information. After this, molecules get uniquely barcoded, thus allowing the amalgamation of data and later separation upon sequencing. Methods of data production are mainly specialised to particular molecular layers; for example, mass cytometry, ATAC-seq, and single-cell RNA sequencing are associated with scRNA-seq. To find patterns, correlations, and regulatory processes, integrating and analysing multi-omics data requires complex computational pipelines and machine learning methods [15]. In plant science, it has the potential to give insights into cellular heterogeneity and plant development. The process of organogenesis, tissue patterning, and cell differentiation elucidates how specific types of cells drive processes like root adaptation and photosynthetic optimisation. The pathways in plants are so interconnected that cross-interactions occur; it is a bottleneck in the phytopharmaceutical industry. Alterations are observed in plant cell lines by identifying specific cells responsible for generating secondary metabolites of our interest. It also gives new insights into cellular interactions between pathogens and plant cells [16]. With continuous technical developments targeted at improving resolution, sensitivity, and the simultaneous profiling of many omic layers, single-cell multi-omics has a bright future. Cellular data will be further contextualised within their natural surroundings through integration with spatial omics. Applications in developmental biology, such as tracking lineage trajectories, and in precision medicine, like the

development of treatments tailored to tumour or immune profiles, could revolutionise both research and clinical practice, in plant biology toward innovations in sustainable agriculture and crop enhancement.

3. Data integration in multi-omics

The core concept of system biology is data integration in multi-omics, which combines a range of datasets to provide insights into intricate biological processes [17]. To explain the mechanism behind the differences in phenotypic and disease pathogenesis, multi-omics integration aims to comprehend linkages across the omics layers of transcriptomics, proteomics, metabolomics, and genomics. However, processing diverse and high-dimensional data, variations in nomenclature, and the absence of common methods are some of the major obstacles that arise with the integration of multi-omics data [17].

3.1 Challenges in data integration

The integration of multi-omics data is a promising approach to understanding complex biological systems; however, there are several challenges that need to be addressed. The main problem lies in data heterogeneity across the omics layers, which include genomics, transcriptomics, proteomics, and metabolomics, as each layer varies in type, scale, and format because of differences in sample collection, analytical platforms, and pre-processing protocols [18, 19]. Moreover, issues such as overfitting and computational inefficiency are frequently caused by the high dimensionality of multi-omics datasets, which have a large number of variables in comparison to the number of samples, particularly in statistical and machine learning models [20].

In addition, incompleteness of the dataset is a major challenge, as comprehensive multi-omics profiles are seldom feasible due to financial constraints, sampling issues, or technological constraints. Partially overlapping datasets thereby result in more difficult integration attempts. Also, lack of uniform data pre-processing and integration protocols impedes reproducibility and comparison among studies [21]. Combining partially overlapping datasets from different sources complicates the integration process, but new methods like covariance-based approaches for merging relationship matrices promise a bright future for improving genomic prediction and inferences of trait relationships [22]. However, data heterogeneity remains the greatest obstacle to effective big-data analytics and calls for interdisciplinary cooperation as well as data sharing across institutional boundaries [23].

The biological complexity of interactions across omics layers complicates the interpretation of integrated data. Interactions are driven by both genetic and environmental factors, which introduce feedback loops that are hard to disentangle. Tools and methods for integrating multi-omics data have been developed for various applications. However, there are challenges in data cleaning, normalisation, dimensionality reduction, and biological contextualisation [7]. The integration of diverse omics datasets aims to bring unprecedented insights into cellular systems. However, there are limitations put forth by the integration, interpretation, and insight generation [7].

Identifier mapping and annotation issues are worst with regard to lipidomics and metabolomics. Statistical analysis is a concern for integration since every layer of omics and analytical methods has varying noise and variability [24]. Metabolomics focuses on metabolite analysis and is especially useful for enlightening cellular

phenomena, but only if integrated with other data to acquire a comprehensive understanding [25]. Some tools exist for metabolomic data analysis and integration, including MetaCore, MetaboAnalyst, InCroMAP, and 3Omics, which have different strengths and weaknesses [25].

Some layers, like transcriptomics and metabolomics, are extremely susceptible to environmental and temporal influences. As a result, they need to be sampled carefully in order to collect significant insights [26]. Last but not least, despite the rise in popularity of data-driven methods like machine learning, their usefulness in reaching biological findings that can be put into practice is significantly limited by the inability to comprehend the results that are frequently obtained from these methods [17].

To fully realise the potential of multi-omics integration, these obstacles must be overcome by a mix of data-driven and knowledge-based methodologies, strong computational frameworks, and interdisciplinary partnerships.

3.2 Scalability and scopes of multi-omics integration

Integrating multi-omics data has now become essential in advancing our understanding of complex biological systems and enhancing clinical outcomes [27]. It will help the researchers in integrating the complementary information from diverse omics layers to gain a holistic view of diseases and cancers in medical science, crop yield, disease resistance, tolerance against abiotic and biotic stresses, and how environmental changes such as global warming are affecting overall biodiversity. The applications of multi-omics are discussed in depth later in this book chapter.

Scalability of multi-omics integration:

A. Integration strategies: Strategies in combination have been developed early, mixed, intermediate, late, and hierarchical to integrate omics datasets with different advantages on the scalability of the data [28]. Researchers are able to handle large-dimensional data in an effective way across different biological domains.

B. Technological advancements: High-throughput technologies have now advanced to incorporate sophisticated sequencing platforms and mass spectrometry, allowing large-scale omics datasets to be generated. Challenges include, however, bottlenecks such as sequencing depth, variability in sample preparation, and limitations in measurements [29].

C. Algorithms and tools for computations: More tools and algorithms have been designed to deal with the scalability challenges of multi-omics integration. The applications of such tools are, for example, in the prediction of biomarkers, data visualisation, and disease subtyping, making it easier for researchers to handle complexity when integrating various datasets [2].

3.3 Limitations and future directions in multi-omics data integration

Despite technological progress, challenges persist. These include the need to align gene and protein expression rates, handle heterogeneous data types, and meet the computational demands of large datasets. Additionally, the lack of standardised workflows and protocols further complicates scalability [29]. The field is moving toward the development of more sophisticated integration methods that can fully

leverage multi-omics data. These include hybrid computational frameworks that combine knowledge-based and data-driven approaches, as well as distributed computing and federated learning systems to handle large datasets. Enhanced interdisciplinary collaboration will also help overcome scalability barriers and expand the applications of multi-omics in precision medicine [27].

3.4 Data integration approaches in multi-omics

In multi-omics data integration, approaches are divided into strategic approaches and mathematical approaches. These integration approaches are somewhat similar, but for detailed understanding, they are classified in strategic approaches and mathematical approaches.

3.5 Strategic approaches

3.5.1 Knowledge-based approach

Knowledge-based approaches depend on carefully curated external information, such as pathway databases, protein-protein interaction networks, and regulatory maps, to interpret and integrate multi-omics datasets. These approaches contextualise data within established biological frameworks, which enables meaningful insights into complex systems. Resources such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG), which links genes and proteins to metabolic and signalling pathways, and Reactome, which organises human biological processes into temporally structured reactions. Tools such as STRINGS offer confidence score-based protein-protein interaction networks based on experimental data, co-expression analysis, and computational predictions. Set-based enrichment analysis maps omics entities to functional annotations, identifying pathways or processes enriched in significant genes or metabolites. Techniques such as Over Representation Analysis (ORA) and Functional Set Enrichment Analysis (FSEA) extend these capabilities through the inclusion of statistical tests and quantitative data. Secondly, olic modelling (CBMM) relies on genome-scale metabolic models (GEMs), which are fine-tuned through experimental omics data. Challenges exist, such as identifier iplete database annotations. However, a biologically founded strategy for integrating multi-omics is provided through knowledge-based approaches. Many applications are there, not just limited to disease research, metabolic engineering, and personalised medicine, so there is no possible way to miss the importance in interpreting complex biological data [17].

3.5.2 Data-driven approach

Data-driven approaches discover patterns, relationships, and dependencies in multi-omics datasets without being dependent on much prior biological knowledge. Such approaches are useful in exploring high-dimensional heterogeneous data to reveal novel insights or to handle poorly annotated datasets. Techniques like clustering (k-means, hierarchical clustering) group entities according to shared molecular patterns, thereby uncovering co-expression modules and functional similarities. Dimensionality reduction techniques, including principal component analysis (PCA) and t-SNE, reduce the complexity of large datasets while retaining important relationships, which helps in visualisation and the identification of contributors to biological variation.

Network-based approaches construct interaction networks where nodes represent biological entities, and edges reflect statistical or functional relationships. The developed algorithm takes further steps to refine these networks and identify key regulatory interactions, integrating specific layers of omics, such as transcriptomics and proteomics, to detail molecular pathways. Machine learning models extend data-driven methods further for prediction of phenotype, biomarker discovery, and pathway identification [17].

However, the technique faces challenges posed by heterogeneity across omics layers, mainly in scale and variability, but also in interpreting output from machine learning models, whose functioning often borders on the mystery of "black boxes." Apart from these problems, constant increases in computational facilities and big-data infrastructure improve the feasibility and scalability of these approaches. Data-driven discoveries are critical toward revealing new system insights and precision medicine to lead targeted interventions for systems biology research [30].

3.5.3 Hybrid approach

A sophisticated way for integrating data from several omics, the hybrid approach combines data-driven and knowledge-based methods to reveal complex biological connections across multiple omics layers. This hybrid paradigm combines data-driven methods to find innate patterns in datasets, such as correlation and statistical linkages, with knowledge-based methods that use curated biological data from databases like KEGG, Reactome, and STRING. Hybrid approaches create unified, multi-layer networks containing biological entities as nodes and their interactions as edges by bridging information from various technologies or research, even if their profiles are partially overlapping or discontinuous. As a result, regulatory, enzymatic, and co-expression interactions can be fully represented and enhanced by both data-driven linkages and previously known pathway information [17].

3.6 Methodological approaches

3.6.1 Element-based integration

Element-based integration is one of the foundational approaches in multi-omics data integration that focuses on identifying direct relationships between individual elements, such as genes, proteins, and metabolites, across different omics layers. Normally it relies on correlation analysis, clustering, and multivariate statistical techniques to find interactions and shared patterns. As a specific example, correlation analysis typically relies on methods such as Pearson or Spearman coefficients for determining the strength and directions of associations among datasets, like transcript-protein interactions. Often, correlations may be weak simply because biological processes are so complex, such as post-transcriptional modifications. The analysis of clustering by using tools like k-means or random forest will group elements based on similar expression patterns, which would allow the identification of stage-specific molecular clusters or regulatory mechanisms, such as stress responses in plants. Multivariate analysis, which includes methods like PCA or OPLS-DA, helps reduce the dimensionality of the data and is useful for visualising patterns across omics layers. These approaches are crucial for exploring dynamic relationships and networks between omics elements, such as connecting proteomics and metabolomics to understand metabolic adjustments under environmental stress [29].

The element-based integrative approach is intuitive and straightforward; its primary limitation is that it is unable to capture the complexity of molecular interactions across systems, and therefore, it is most effective when used as a complement to more advanced integration strategies [29].

3.6.2 Pathway-based integration

The pathway-based integration involves the use of known biological pathways to integrate and interpret multi-omics data through a biologically intrinsic framework used to study the complex interactions that occur at a molecular level [31]. Such a reconstruction occurs by mapping the omics data on pathway databases, such as KEGG, MapMan, or Cytoscape. Researchers can visualise interactions among genes, proteins, and metabolites to analyse. Pathway mapping is a popular method where omics data are aligned with predefined pathways to identify enrichment patterns and agitation in specific biological processes. For example, integrating transcriptome and metabolome data can be used to reveal pathway disruptions in stress-tolerant plants or disease states. Co-expression analysis, using tools such as Weighted Gene Co-expression Network Analysis (WGCNA), builds networks that reveal key modules and hubs, deepening the understanding of molecular mechanisms and regulatory hierarchies. Pathway-based integration is particularly effective in identifying functional relationships and interpreting multi-omics data in the context of known biology. However, for novel or poorly characterised systems, it entirely depends on pathway annotations that may be incomplete and inaccurate. Pathway-based integration provides a robust platform for generating biologically relevant insights, mainly in areas of systems biology, plant science, and medical research [32].

3.6.3 Mathematical-based integration

It is a highly sophisticated approach in multi-omics, which makes use of mathematical models to stimulate, analyse, and predict biological processes by integrating datasets from multiple omics layers. It is a hypothesis-driven approach and aims to unveil dynamic relationships and causal interactions among genes, proteins, and metabolites. The most notable method is differential equation modelling, in which equations describe changes in the system components with respect to time and allow the prediction of molecular behaviour under certain conditions. For instance, ordinary differential equations have been used to predict protein dynamics and metabolic fluxes and provide insights into biological responses. Genome-Scale Modelling (GSM), another advanced method, where a comprehensive metabolic network is constructed to simulate an entire cellular system. Tools such as the PlantSEED model metabolic flux to predict the impact of genetic or environmental uncertainty on pathways. This approach enables researchers to investigate complex biological mechanisms, such as lignin biosynthesis or metabolic adaptation in plants, with very high precision. However, integration *via* mathematical bases is computationally resource-intensive and requires a deep understanding of biological systems, as well as time to develop and validate the model. Despite these challenges, it offers unparalleled insights into the functional integration of omics layers and works as a valuable tool for predictive and systems biology [11].

3.7 Artificial intelligence and machine learning

Artificial intelligence and machine learning integration into multi-omics data analysis has transformed the field by enabling the combination and interpretation

of complex molecular datasets, including genomics, transcriptomics, proteomics, metabolomics, and epigenomics. It does provide a comprehensive view of biological systems but is inherently challenging due to the complexity, high dimensionality of the data, and data heterogeneity. AI-ML approaches are addressing and are potent to address these challenges by facilitating data integration, dimensionality reduction, feature selection, and predictive modelling. Techniques such as similarity network fusion (SNF) and multi-omics factor analysis (MOFA) enable the flowless integration of contrasting omics layers, while dimensionality reduction methods such as PCA, t-SNE, and autoencoders reduce complexity while preserving meaningful biological patterns. ML-based clustering algorithms, such as k-means and hierarchical clustering, help in the identification of disease subtypes and patient stratification, while predictive models like random forests, support vector machines (SVMs), and deep neural networks (DNNs) give insights into clinical outcomes and therapeutic targets. Additionally, advanced imputation strategies, such as generative adversarial networks (GANs) and denoising autoencoders, address missing data, an important issue in multi-omics.

Several challenges still remain, despite this advancement. The heterogeneity of omics datasets, ranging from binary genomics to continuous omics data, complicates multi-omics integration. High dimensionality introduces risks of overfitting, while missing values and batch effects further hinder data harmonisation. Many ML models, particularly deep learning approaches, are considered "black boxes," making it complex to interpret results in a biologically meaningful way. Moreover, the computational resources and expertise required for these methods can limit accessibility, mainly for smaller research groups.

The future of AI/ML in multi-omics is promising, with several emerging trends. Explainable AI models are being developed to enhance the interpretability of ML outputs, using techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). Single-cell and spatial multi-omics integration will enable deeper insights into cellular heterogeneity and tissue microenvironments. Advancing fields such as cancer research and developmental biology. Real-time analysis of dynamic systems may improve in areas such as dynamical modelling using generative models like GANs and variational autoencoders (VAEs) in creating synthetic data as a form of imputation. Generative models stimulate multi-omics data in attempts to cover up missing values to generate new artificial datasets. Democratising access to cloud-based systems and accessible platforms is foreseen to help access AI-ML applications in scale through a scientific crowd. Furthermore, AI-ML is going to be the enabler of breakthroughs in precision medicine and drug discovery by integrating patient-specific multi-omics data with clinical outcomes that would help in identifying personalised therapies and novel drug targets.

AI-ML has dramatically altered multi-omics integration, enabling it to bypass the in-built complexities. It possessed and finally unlocked the hidden possibilities for holistic biological analysis. Even so, problems related to data heterogeneity, model interpretability, and missing values persist. Future advancement will eliminate these multidisciplinary multi-omics, fuelled by AI-ML, which will lead to system biology, crop improvement, and precision medicine with the capability of changing the current paradigm in understanding complex biological processes.

3.8 Workflow pipeline

A multi-omics pipeline can be customised to analyse a single-omics layer across different conditions or integrate several omics layers for a holistic view of one

DOI: http://dx.doi.org/10.5772/intechopen.1009591

condition. In studying one layer of omics across several conditions, the pipeline starts with standardised sample harvesting across different experimental settings, for example, stress versus control or conditions treated versus nontreated. After the generation of data, the next steps are quality control and normalisation in order to reduce experimental biases. Differential expression analysis does statistical analyses that single out changes specifically conditioned for their genes, proteins, or metabolites. These are in turn mapped to biological pathways or networks using databases like KEGG or Reactome [26]. Visualisation tools such as heatmaps, together with clustering methods, can further help depict the features distinguishing conditions so one can better understand condition-specific molecular responses [33].

In contrast, bringing together data from multiple layers of omics under the same condition entails data generation from various omics platforms that involve genomics, transcriptomics, proteomics, and metabolomics with the use of identical biological samples. To ensure uniformity, quality control and normalisation are applied separately to each omics dataset. Feature extraction methods include principal component analysis and machine learning, which identify important components in datasets. For data integration, advanced computational methods such as MOFA or DIABLO are used. These methods align the omics layers either vertically or horizontally to bring out molecular linkages [33]. A full understanding of the situation is obtained through pathway enrichment and network studies, which correlate the integrated data to biological processes. The last steps include validation using experimental or database references and result visualisation by using integrative tools like Circos or Sankey diagrams. Combining these processes offers an extremely

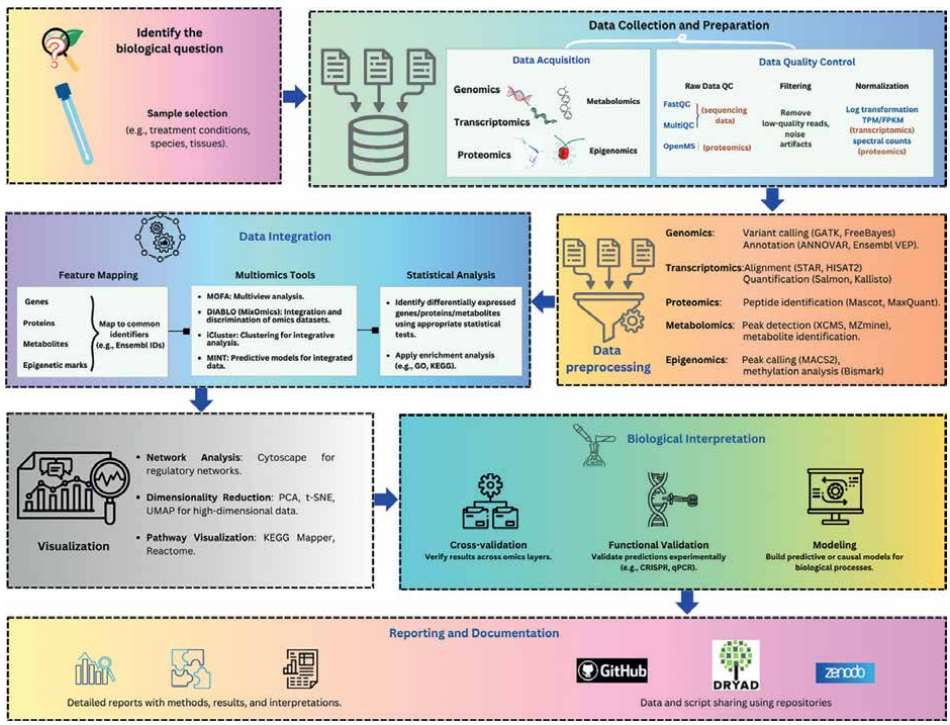

Figure 1.
Generic pipeline of multi-omics.

potent framework to identify system-wide interactions in a given context or molecular alterations across situations. The generic pipeline of multi-omics study has been illustrated in **Figure 1**.

4. Computational tools and databases

Recent developments in multi-omics technology have revolutionised biological research by providing previously unheard-of insights into disease causes and cellular processes. To evaluate and combine these intricate datasets, a plethora of databases and computational tools have been created. An overview of more than 4400 web-accessible tools for multi-omic data processing may be found in OMICtools, a manually curated meta database [33]. The knowledge about cellular dynamics and heterogeneity has been further improved by the development of single-cell multi-omics [12]. Nevertheless, the integration of multiple omics data is still a difficult task, which has led to the creation of numerous computational techniques that make use of network theory and machine learning [14]. Methodologies are crucial for developing precision medicine and system biology, providing researchers and clinicians with new ways to analyse and apply the abundance of multi-omics data [12, 34]. **Table 1** lists the many types of tools and resources that help with the execution of a multi-omics workflow, along with examples for each category. Apart from these datasets, there are many more, and according to need, they are carefully chosen.

Sr. no	Tool/software	Purpose	Link
Popular/emerging multi-omics tools			
1	mixOmics	A tool with a framework that provides wide range of multivariate statistical methods for exploratory data analysis (EDA). This involves feature identification, extraction and selection.	http://mixomics.org/
2	MOFA	A probabilistic multi-omics factor analysis-based framework that involves EDA and data integration. (Unsupervised)	https://github.com/bioFAM/MOFA
3	SNF	3 A multi-view network and fusion analysis framework for feature extraction, pairwise similarity, clustering, classification, etc.	https://cran.r-project.org/web/packages/SNFtool/index.html
4	miodin	A multi-level statistical framework involving vertical and horizontal integration of multi-omics data.	https://algoromics.gitlab.io/miodin/
5	Paintomics	A web-based systems biology tool for multi-omic integration and visualisation across multi-species.	www.paintomics.org
6	3Omics	A web-based application for integration and analysis of multi-omics data.	https://3omics.cmdm.tw/
Data sharing			
7	OmicsDI	An aggregated database facilitating the discovery of heterogenous published omics datasets across studies.	http://www.omicsdi.org
8	Zenodo	A general-purpose open-access data, softwares, etc., repository that allows user to obtain a citable DOI.	https://zenodo.org/
9	OSF	An open platform to enable collaboration by registering research projects, materials, data and documentation.	https://osf.io/

Sr. no	Tool/software	Purpose	Link
Code Sharing			
10	GitHub	A version-controlled code sharing and collaborative platform.	https://github.com/
11	BitBucket		https://bitbucket.org/
12	GitLab		https://about.gitlab.com/
Workflow Sharing			
13	Common Workflow Language (CWL)	An open standard for describing analysis workflows which makes them portable and scalable across a variety of software and hardware environments.	https://www.commonwl.org/
14	Nextflow	An enterprise level workflow language for writing scalable and reproducible scientific pipelines.	https://www.nextflow.io/ Di
15	Snakemake	A workflow language for writing scalable and reproducible scientific pipelines.	https://snakemake.readthedocs.io/en/stable/
Environment Sharing			
16	Conda	A package manager and computation environment management system.	https://docs.conda.io/en/latest/
17	Bioconda	A channel for the conda package manager specialising in bioinformatics software.	https://bioconda.github.io/
18	Docker	A container platform that provided OS-level virtualisation for providing reproducible computation environment.	https://www.docker.com/
19	BioContainers	A community-driven project that provides docker-based containerised bioinformatics software.	https://biocontainers.pro/
20	renv	A R-package that helps create reproducible environments for R-based projects.	https://rstudio.github.io/renv/
Data Visualisation			
21	Shiny	A framework in R for doing GUI-based interactive applications.	https://shiny.rstudio.com/
22	Plotly	A cross-language interactive plot library.	https://plotly.com/
23	bokeh	A Python library for Interactive data visualisation in browser.	https://bokeh.org/
24	D3,js	A JavaScript library for producing dynamic, interactive data visualisations in web browsers.	https://d3js.org/
25	Cytoscape	A platform for network data integration, analysis, and visualisation.	https://cytoscape.org/

Table 1.
This is a list of emerging and popular multi-omics tools and platforms for data visualisation, workflow, environment sharing, code sharing, and data sharing. The table is adapted from Jamil et al. [11].

5. Application: the transformative potential of multi-omics

Multi-omics is the integration of several high-throughput technologies for the analysis of intricate biological systems at various molecular levels. It can be used in a wide range of life science fields. Several of its primary uses are illustrated in several domains:

5.1 Agriculture and crop improvement

Since multi-omics integration has made it possible to understand the molecular mechanisms underlying significant traits, scientists can find genes, proteins and metabolites linked to disease resistance, abiotic and biotic stress tolerance, and yield enhancement by integrating different omics layers. These discoveries contribute to the creation of crops that can withstand climate change and biofortified crops that have more nutritional value. By clarifying the processes of heterosis and locating indicators for superior hybrids, multi-omics also helps with hybrid breeding. Additionally, it speeds up genetic prediction and marker-assisted selection, reducing the time and expanse. Sustainable agriculture can be supported by the integration of multi-omics with soil and microbiome study, which improves nutrient uptake and stress management. Secondary metabolism applications enhance crop quality by maximising natural pest control, flavour, and aroma. Additionally, multi-omics informs conservation strategies by characterising genetic resources and domestication processes. By integrating multi-omics with advanced computational tools and precision phenotyping, plant breeding can achieve breakthroughs in productivity, sustainability, and food security.

5.2 Health and medical science

Multi-omics integration is revolutionising personalised medicine because it provides comprehensive understanding of molecular mechanisms behind health and disease. It discovers diagnostic, prognostic, and predictive biomarkers and therapeutic targets by integrating genomics, transcriptomics, proteomics, metabolomics, and epigenomics. Multi-omics can be used for the prediction of drug response [12]. Optimised treatment regimens tailored to the individual patient would be possible in cancer precision medicine, characterising tumour heterogeneity, and guiding targeted therapies or immunotherapies [34]. It also aids in management of metabolic disorders by uncovering disrupted pathways for personalised dietary and pharmacological interventions. Epigenomics adds insights into DNA methylation and chromatin dynamics, thus supporting the development of patient-specific epigenetic therapies. There are many diseases whose mechanisms are unknown even to scientific communities and integrating multi-omics will give insights into disease epidemiology. In neuroscience, it unwinds the genetic, molecular, and metabolic basis of neurodegenerative diseases such as Alzheimer's and Parkinson's and thus aids in the identification of biomarkers and drug targets. For regenerative medicine, breakthroughs gained by multi-omics promise betterments in basic research on stem cells, tissue regeneration, and organ transplantation by insights into cellular differentiation and tissue repair with consequent improvement in the therapeutic index. Multi-omics integration accelerates the process of vaccine development; in the disease outburst, it plays a crucial role to find the most potent medicine in case of antimicrobial resistance. In addition, single-cell multi-omics increases precision by revealing cellular heterogeneity and individual treatment responses. With the integration of AI and machine learning, multi-omics further refines predictive models, paving the way for personalised interventions that improve outcomes and reduce adverse effects across diverse medical conditions (**Figure 2**).

5.3 Systems biology

Multi-omics integration is the cornerstone of systems biology, which enables the comprehensive understanding of complex biological systems by integrating data from

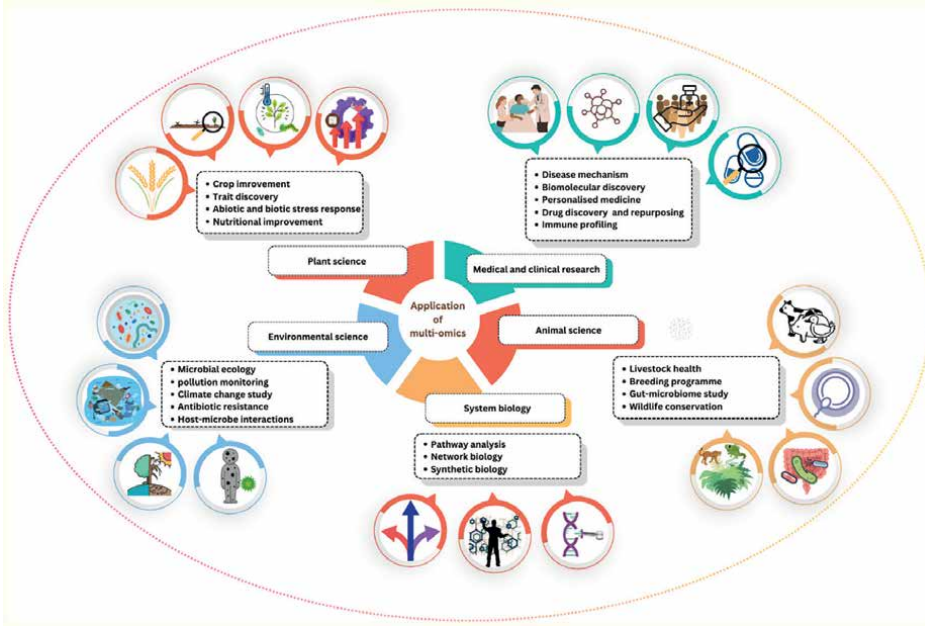

Figure 2.
Application of muti-omics in plant science, medical and clinical science, animal science, system biology, and environment science.

genomics, transcriptomics, proteomics, metabolomics, and epigenomics. It facilitates the reconstruction of gene regulatory networks, protein-protein interactions, and metabolic pathways and provides insights into cellular processes and disease mechanisms. Multi-omics has driven drug development discoveries, especially in target identification, predicting toxicity, and modelling drug metabolism, while pushing forward synthetic biology with the design of optimised biological circuits. In ecology and microbiome research, it shows interactions between microbial communities and their environments. In addition, multi-omics increases the annotation of poorly understood genes and pathways, models stress responses, and supports evolutionary studies by decoding changes in regulatory networks. By integrating diverse data layers, multi-omics empowers systems biology to address complex challenges in medicine, agriculture, and biotechnology.

5.4 Environmental and ecological studies

Multi-omics technologies have transformed environmental and ecological sciences by providing a comprehensive understanding of organisms and ecosystems at the molecular level. By identifying their functions in biogeochemical cycles and the breakdown of pollutants in soil, water, and air, these methods aid in the characterisation of microbial populations. Additionally, multi-omics helps conserve biodiversity and restore ecosystems by revealing the molecular mechanisms behind organismal responses to climate stressors such as drought, salt, and temperature fluctuations. It advances research on nutrient cycle, soil health, and host-microbiome interactions, supporting environmentally resilient practices and sustainable agriculture. Multi-omics clarifies the functions of microbial and plankton communities in carbon sequestration and nutrient cycling in marine and

aquatic environments. By locating and refining pathways for pollution breakdown, it also promotes bioremediation. Multi-omics offers vital insights for ecological conservation and sustainable environmental management by combining many biological layers.

5.5 Metabolic engineering and biotechnology

In metabolic engineering, multi-omics integration to identify key genes, enzymes, and pathways that control the production of desired metabolites. This enables optimisation of microbial and cellular factories for efficient production of biofuels, pharmaceuticals, food additives, and industrial chemicals. Understanding the interactions between different molecular layers allows researchers to design strategies to enhance metabolic fluxes, eliminate bottlenecks, and improve yield and sustainability. Synthetic biology involves multi-omics in the rational design of synthetic circuits and biological systems through insights into natural interactions and regulatory mechanisms of organisms. Genomic and transcriptomic data inform the choice and modification of genetic parts, while proteomic and metabolomic analyses are helpful in validating functional performance and assessing the impact of synthetic constructs on host metabolism. In addition, multi-omics helps in the design of minimal genomes and chassis organisms tailored toward specific applications while achieving stability and performance in engineered systems.

Multi-omics-based approaches are also crucial in troubleshooting and optimisation of engineered pathways. The interaction of data from different layers of omics can reveal unintended metabolic byproducts, off-target effects, and regulatory feedback loops. Thus, data iteration is possible for refinement of designs. This systems-level understanding accelerates the development of robust and scalable synthetic biology applications, including biosensors, bioremediation tools, and novel therapeutics. Multi-omics thus serves as a cornerstone for advancing the fields of metabolic engineering and synthetic biology, bridging the gap between design and functionality in engineered biological systems.

6. Challenges and limitations

Multi-omics research is associated with many challenges, most of which are related to the challenges of individual omics analyses [7]. These challenges are compounded when integrating multiple omics datasets, introducing additional complexities such as data fusion, clustering, visualisation, and functional interpretation [9, 11].

6.1 Pre-integration challenges

Before integrating several omics datasets, harmonisation issues need to be addressed. Most datasets require specific pre-processing steps, such as normalisation, scaling, and transformation, that are unique to their characteristics. These differences can make integration challenging and time-consuming. For instance, large-scale studies with thousands of samples have constraints in terms of the data volume, which increases computational demands and storage requirements.

6.2 Technical challenges

6.2.1 Mapping and annotation difficulties

The primary challenge involves mapping identifiers across datasets, for example, mapping genes to their corresponding transcripts or proteins. In most cases, such mappings are not one-to-one. This mapping becomes very difficult for some of the omics, especially for metabolomics [9]. It usually demands the usage of external databases such as RefSeq or KEGG. These, however, may have low coverage for specific omics. For instance, metabolites are not found in RefSeq, and in some cases, their IDs could be outdated since sources update late, as seen in KEGG that is dependent on older RefSeq data. Also, the depth of molecular identification differs significantly: although genomic annotations are comprehensive, the phosphoproteome is not yet complete, and identification of metabolites is still the major bottleneck. These are being addressed through new tools and approaches, including metabolite set enrichment analysis, pathway visualisation, and text mining [26]. Improvements in metabolite annotation in pathway databases, like WikiPathways, have led to a substantial increase in the number of annotated metabolite nodes and interoperability through FAIR principles and web services [35].

6.2.2 Data heterogeneity

The variability in data types across omics (e.g., transcriptomic counts vs. metabolomic concentrations) and the complexity of their relationships highlight the inherent challenges of integrating multi-omics data [36]. These difficulties are especially evident in biomedical studies focusing on humans, where heterogeneity is amplified by diverse data sources [36]. By addressing these technical and computational barriers, multi-omics integration can become a more robust tool for understanding complex biological systems and advancing precision medicine.

6.3 Computational and analytical challenges

Multi-omics research presents significant computational and analytical problems in addition to technical ones. The intricacy of combining sizeable, high-dimensional datasets from many omics layers and extracting significant biological insights from them is the main cause of these difficulties [2, 11].

6.4 Dimensionality reduction and representation

Multi-omics datasets are typically high-dimensional, meaning they contain many variables (features) that may not all be relevant to the biological question at hand. Dimensionality reduction is a key step used to reduce the number of features while preserving the essential relationships between them. To visually depict these relationships and to find patterns, outliers, and sources of variation (batch effects), methods such as principal component analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) are frequently employed [2, 11]. The variations in the nature of data across many omics layers, including transcriptomics, proteomics, metabolomics, and genomes,

make it difficult to apply dimensionality reduction to multi-omics data. Determining the optimal representation technique is challenging since feature relationships might be either linear or non-linear. Furthermore, no approach has been shown to be effective for all kinds of biological data. Therefore, the best option for dimensionality.

6.5 Multicollinearity and non-linear relationships

Another significant computational challenge is multicollinearity. Which occurs when several features in the datasets have a strong correlation with one another. This problem is especially prevalent in multi-omics studies because distinct omics layers may capture related biological signals. For example, there could be a correlation between standard statistical methods that sometimes assume that features are independent of one another; they might find it much harder since complex biological interactions can be hard to capture by traditional methods such as PCA or linear regression. Relationships between DNA, RNA, proteins, and metabolites are frequently better represented in the setting of gene regulatory networks (GRNs) by non-linear techniques like mutual information (MI)-based networks, which have been demonstrated to function better in these situations than linear models. GRN inference has been further transformed by single-cell multi-omics data, which enables more thorough and accurate network reconstructions. But there are still difficulties in creating techniques that can completely utilise these intricate datasets potential [37]. Advanced statistical and machine learning techniques are currently being developed and optimised.

6.6 Integration of different data layers

One of the most difficult parts of multi-omics analysis data from several omics layers is the integration of different multi-omics data layers. Data integration can be many-to-many (e.g., genes interacting with metabolites) or one-to-one (e.g., gene-to-gene). It is challenging to combine data in a way that effectively depicts the underlying biological processes because of its complexity. Further challenges arise when integrating non-omics and multi-omics data, such as phenotypic and clinical [36]. The ability of current statistical techniques to capture intricate relationships across several data layers is frequently constrained, and techniques for handling the integration of such datasets are developing. By employing factor analysis and network fusion approaches to extract significant characteristics across many omics types, tools such as multi-omics factor analysis (MOFA) and mix omics seek to overcome these issues [38]. But even with these developments, multi-omics research still faces the constant difficulty of creating robust integration.

In conclusion, continuous improvement in data integration techniques, statistical tools, and computer resources continues to enhance the capacity to analyse and interpret complex biological data, despite the significant technical and computational obstacles in multi-omics research. As these techniques advance, multi-omics may offer more profound understanding of the molecular processes that underlie both health and illness.

6.7 Statistical and technical limitations

Multi-omics analysis requires sophisticated statistical methods to account for high-dimensional data and complex interactions between variables. Multicollinearity,

in which many omics layers may correlate with one another, is one of the main obstacles that makes the analysis more difficult. These relationships may be too complicated for standard statistical approaches to handle, which requires the use of more sophisticated strategies like feature extraction and dimensionality reduction. Furthermore, specialised algorithms, like mutual information (MI)-based networks or other machine learning models, which are still being developed to better analyse complex multi-omics data, are frequently required to capture non-linear interactions between data layers [11].

7. Future trends

Multi-omics is set to undergo revolutionary developments in the future, especially in single-cell and spatial multi-omics, which will allow for a more comprehensive comprehension of organismal function and cell biology. Significant advancements in multi-omics technologies' sensitivity, specificity, and throughput are expected, as well as cost savings and the incorporation of additional modalities into a single test. Comprehensive somatic mutation profiling and phylogenetic lineage reconstruction are limited by the difficulty of fully and accurately characterising genetic variations at the genome level. Similar improvements are needed for epigenomic measurements, including histone post-translational modifications (PTMs), in order to integrate with other omics layers and enable the co-detection of numerous marks. Expanding to total RNA, which includes non-coding RNAs and isoform detection, might improve transcriptomic profiling, which is now restricted to poly(A) RNA. Multimodal integration presents difficulties for proteomic and metabolomic investigations, as antibody-based proteomics limits the ability to profile proteins simultaneously. Although mass spectrometry provides objective substitutes, it is incompatible with other omics tests. Multimodal integration of unexplored modalities, such as the epitranscriptome—which includes transcript base changes influencing gene expression—may be part of future advancements. Significant progress in spatial multi-omics is anticipated, enabling the simultaneous analysis of extrinsic and cell-intrinsic properties without tissue separation. Our knowledge of organismal development, cell migration, and stem cell biology will be revolutionised by the combination of phylogenetic lineage information with spatial and single-cell multi-omics data. Additionally, the discipline will make strides in capturing ancestral cellular states, facilitating live cell serial measurements, and enhancing computational tools for integrative studies and data extraction. Overcoming present obstacles, such as improving low-input methods like amplification-free long-read sequencing and enhancing the capabilities of single-cell and spatial assays, is necessary to accomplish such holistic profiling. These assays' commercial availability will increase their accessibility to researchers and promote their broad use. These developments will enhance disease stratification, broaden our knowledge of development, ageing, and illness causation, and provide guidance for precision medicine and novel therapeutic strategies.

The dark genome is a huge, unexplored area in biological research, including non-coding regions, regulatory elements, and poorly annotated genetic sequences. The hidden genomic landscape of such areas has immense potential for the discovery of complex biological functions in plants and other organisms. Multi-omics integration, combining genomics, transcriptomics, epigenomics, proteomics, and metabolomics, provides a powerful approach in deciphering the whole of the dark genome in cellular processes and phenotypic traits. For instance, the integration of transcriptomics and

epigenomics can provide insights into how non-coding RNAs and regulatory elements control gene expression and stress responses, while proteomics and metabolomics link these regulatory features like DNA methylation and histone modifications. Poorly annotated regions, most especially in plant species that are large, redundant, and loaded with transposons, limit the association of these elements to biological functions. Functional validation is a costly and time-consuming exercise for dark genome elements such as IncRNA and enhancers, and the ability of the presently available sequencing technology and multi-omics platforms is quite limited to decipher repetitive regions or low-abundance molecules. Standardised pipelines on computational levels were also not there for a better lead. However, the revolution of single-cell and spatial multi-omics, long-read sequencing, and AI-driven data integration is opening the doors to the dark genome. Enhancement in annotation tools, collaborative efforts, and innovative experimental methods for functional validation will be critical in translating multi-omics findings into actionable insights. By addressing these challenges, the dark genome will emerge as a critical focus in biological research and hopefully open transformative opportunities for crop improvement. Evolutionary studies or environmental resilience.

However, the revolution of single-cell and spatial multi-omics, long-read sequencing, and AI-driven data integration is opening the doors to the dark genome. Enhancements in annotation tools, collaborative efforts, and innovative experimental methods for functional validation will be critical in translating multi-omics findings into actionable insights. By addressing these challenges, the dark genome will emerge as a critical focus in biological research and hopefully open transformative opportunities for crop improvement, evolutionary studies, or environmental resilience.

Author details

Kinjalben Suthar and Nisha Singh*
Gujarat Biotechnology University (GBU), Gandhinagar, Gujarat, India

*Address all correspondence to: nisha.singh@gbu.edu.in;
singh.nisha88@gbu.edu.in

IntechOpen

References

[1] Yadav SP. The wholeness in suffix -omics, −omes, and the word om. Journal of Biomolecular Techniques. 2007;**18**(5):277

[2] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinformatics and Biology Insights. 2020;**14**. DOI: 10.1177/1177932219899051

[3] Giri K, Bisht VS, Maity S, Ambatipudi K. Integrated omics technology for basic and clinical research. In: Biotechnological Advances for Microbiology, Molecular Biology, and Nanotechnology. Toronto, Canada: Apple Academic Press; 2022. pp. 351-407

[4] Bragg JG, Supple MA, Andrew RL, Borevitz JO. Genomic variation across landscapes: Insights and applications. New Phytologist. 2015;**207**(4):953-967

[5] Angel TE, Aryal UK, Hengel SM, Baker ES, Kelly RT, Robinson EW, et al. Mass spectrometry-based proteomics: Existing capabilities and future directions. Chemical Society Reviews. 2012;**41**(10):3912-3928

[6] Yousf S, Chugh J. Nuclear magnetic resonance spectroscopy and mass spectrometry: Complementary approaches to analyze the metabolome. Journal of Endocrinology and Reproduction. 2020;**24**(1):21-30

[7] Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: Tools, advances and future approaches. Journal of Molecular Endocrinology. 2019;**62**(1):R21-R45

[8] Hill MM, Gerner C. Integrative multi-omics in biomedical research. Biomolecules. 2021;**11**(10):1527

[9] Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, et al. Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. Metabolites. 2019;**9**(4):76

[10] Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nature Reviews Genetics. 2018;**19**(5):299-310

[11] Jamil IN, Remali J, Azizan KA, Nor Muhammad NA, Arita M, Goh HH, et al. Systematic multi-omics integration (MOI) approach in plant systems biology. Frontiers in Plant Science. 2020;**11**:944

[12] Li Y, Ma L, Wu D, Chen G. Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. Briefings in Bioinformatics. 2021;**22**(5):bbab024

[13] Chappell L, Russell AJ, Voet T. Single-cell (multi) omics technologies. Annual Review of Genomics and Human Genetics. 2018;**19**(1):15-41

[14] Stanojevic S, Li Y, Ristivojevic A, Garmire LX. Computational methods for single-cell multi-omics integration and alignment. Genomics, Proteomics & Bioinformatics. 2022;**20**(5):836-849

[15] Hu Y, An Q, Sheu K, Trejo B, Fan S, Guo Y. Single cell multi-omics technology: Methodology and application. Frontiers in Cell and Developmental Biology. 2018;**6**:28

[16] Yu X, Liu Z, Sun X. Single-cell and spatial multi-omics in the plant sciences: Technical advances, applications, and perspectives. Plant Communications. 2023;**4**(3):1-14

[17] Wörheide MA, Krumsiek J, Kastenmüller G, Arnold M. Multi-omics integration in biomedical research: A metabolomics-centric review. Analytica Chimica Acta. 2021;**1141**:144-162

[18] Lotta LA, Scott RA, Sharp SJ, Burgess S, Luan JA, Tillin T, et al. Genetic predisposition to an impaired metabolism of the branched-chain amino acids and risk of type 2 diabetes: A Mendelian randomisation analysis. PLoS Medicine. 2016;**13**(11):e1002179

[19] Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nature Genetics. 2014;**46**(6):543-550

[20] Feldner-Busztin D, Firbas Nisantzis P, Edmunds SJ, Boza G, Racimo F, Gopalakrishnan S, et al. Dealing with dimensionality: The application of machine learning to multi-omics data. Bioinformatics. 2023;**39**(2):btad021

[21] Huang Y, Gottardo R. Comparability and reproducibility of biomedical data. Briefings in Bioinformatics. 2013;**14**(4):391-401

[22] Akdemir D, Knox R, Isidro y Sánchez, J. Combining partially overlapping multi-omics data in databases using relationship matrices. Frontiers in Plant Science. 2020;**11**:947

[23] Prokosch HU. Data integration in life sciences. IT-Information Technology. 2017;**59**(4):159-160

[24] Arakawa K, Tomita M. Merging multiple omics datasets in silico: statistical analyses and data interpretation. In: Systems Metabolic Engineering: Methods and Protocols. Totowa, NJ, USA: Humana Press; 2013. pp. 459-470

[25] Cambiaghi A, Ferrario M, Masseroli M. Analysis of metabolomic data: Tools, current strategies and future challenges for omics data integration. Briefings in Bioinformatics. 2017;**18**(3):498-510

[26] Barupal DK, Fan S, Fiehn O. Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. Current Opinion in Biotechnology. 2018;**54**:1-9

[27] Huang S, Chaudhary K, Garmire LX. More is better: Recent progress in multi-omics data integration methods. Frontiers in Genetics. 2017;**8**:84

[28] Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal. 2021;**19**:3735-3746

[29] Graw S, Chappell K, Washam CL, Gies A, Bird J, Robeson MS, et al. Multi-omics data integration considerations and study design for biological systems and disease. Molecular Omics. 2021;**17**(2):170-185

[30] Hernández-Lemus E, Ochoa S. Methods for multi-omic data integration in cancer research. Frontiers in Genetics. 2024;**15**:1425456

[31] Wieder C, Cooke J, Frainay C, Poupin N, Bowler R, Jourdan F, et al. PathIntegrate: Multivariate modelling approaches for pathway-based multi-omics data integration. PLoS Computational Biology. 2024;**20**(3):e1011814

[32] Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. Briefings in Bioinformatics. 2016;**17**(5):891-901

[33] Henry VJ, Bandrowski AE, Pepin AS, Gonzalez BJ, Desfeux A. OMICtools: An informative directory for multi-omic data analysis. Database. 2014;**2014**:1-5

[34] Kaur H, Kumar R, Lathwal A, Raghava GP. Computational resources for identification of cancer biomarkers from omics data. Briefings in Functional Genomics. 2021;**20**(4):213-222

[35] Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research. 2018;**46**(D1):D661-D667

[36] Krassowski M, Das V, Sahu SK, Misra BB. State of the field in multi-omics research: From computational needs to data mining and sharing. Frontiers in Genetics. 2020;**11**:610798

[37] Kim D, Tran A, Kim HJ, Lin Y, Yang JYH, Yang P. Gene regulatory network reconstruction: Harnessing the power of single-cell multi-omic data. npj Systems Biology and Applications. 2023;**9**(1):51

[38] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-omics factor analysis—A framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology. 2018;**14**:e8124. DOI: 10.15252/msb.20178124

Section 2

Bioinformatics Applications

Chapter 4

Protein Sequence Feature Extraction Techniques: Research Advances, Progress, and Applications

Xiaogeng Wan

Abstract

Proteins are intimately involved in transmitting and expressing genetic information and actively engage in various life activities. Since protein sequences encode their structures, and these structures determine the function of the proteins, therefore sequence feature extraction by converting an amino acid sequence to numerical vectors is a very important process in exploring proteins in greater depth. This chapter presents a systematic review of the protein sequence feature extraction literature. The existing features are classified into different categories with respect to their definitions and properties. These include fundamental features that describe the composition and arrangement, physicochemical properties, and features that are based on local sequence units and similarity scores. Advanced features in recent progress are categorized into graphical features, numerical coding features, and probabilistic and information-based features, as well as machine learning features and features obtained via other techniques. Deep learning and language model features are particularly introduced as recent advances. Typical feature-generation platforms are also summarized, and hybrid features are discussed. Finally, popular feature classifiers and areas of applications for the features are outlined as an application guidance.

Keywords: protein sequence, feature extraction, amino acid, physicochemical properties, evolutionary information, protein classification

1. Introduction

Proteins actively involved in life activities have a critical role in transmitting and expressing genetic information [1]. The biological function of proteins is determined by their structure, which is encoded by their sequence. Thus, deep insights into the amino acid features will facilitate the analyses of protein structures and functions and hence benefit related biological research.

Primary protein sequence analyses are based on the similarity between protein sequences. These methods are usually alignment-based, resulting in high computational complexity and unsatisfactory accuracy when dealing with low-identity

sequences [2, 3]. Alignment-free feature extraction approaches are preferred to compensate for these deficiencies. These alignment-free feature extraction methods convert the amino acid sequence into vectors or matrices to represent the various characteristics of proteins, including the composition, order of arrangement, and physicochemical properties of the amino acids. The feature representations can be used as inputs to classification algorithms for classification applications. The quality of the feature representation determines the accuracy of protein classification.

In recent decades, protein sequence feature extraction research has flourished. Diverse techniques have emerged to optimize feature extraction problems. Although diverse approaches have been innovated, converting the arithmetic characteristics of protein sequences into numeral feature vectors is still a challenging problem [3, 4].

In this chapter, we systematically review the existing feature methods, where the different approaches are grouped into different categories concerning their definitions and techniques. The fundamental features refer to the evolutionary features, including the composition and arrangement features, physicochemical property features, features based on local sequence units, and alignment-based similarity scoring features. The advanced features refer to features based on various techniques, including the graphical, numerical coding, information-based, machine learning, and hybrid features. Popular feature-generation platforms, feature classifiers, and classification application areas are also introduced for the convenience of analysis.

2. Fundamental features based on evolutionary information

The fundamental features are the basic features that capture the amino acid compositions and arrangement features, physicochemical property features, and alignment-based similarity scoring features. This section recalls a series of fundamental features that are often used in protein analysis.

2.1 Composition and arrangement features

The composition and order of arrangement of amino acids are the essential characteristics of proteins. The most basic composition and arrangement features are, for instance, the amino acid composition (AAC) feature [5], the pseudo amino acid composition (PseAAC) feature [6], the natural vector (NV) [7], and the quasi-sequence-order (QSOrder or QSO) features [8].

The amino acid composition (AAC) refers to the amino acid frequencies in a given protein sequence:

$$V_{AAC} = (v_1, v_2, v_3, \cdots, v_{20}) \tag{1}$$

where v_i notates the proportion of amino acid type i [5] in the given sequence and i stands for one of the 20 common amino acids [5]. This feature is used in most protein classification/prediction studies [8–17] and is also often used for composing hybrid features.

When considering the categories of functional groups, the frequency of functional groups feature [18] divides the amino acids into ten functional groups: amido (QN), carboxyl (DE), guanidino (R), hydroxyl (ST), imidazole (H), nonpolar (AGILVP), phenyl (FWY), primary amine (K), sulfur (M) and thiol (C) with respect to the side chain chemical groups [18, 19].

The pseudo amino acid composition (PseAAC) [6] is a 20 + λ-dimensional vector [17, 18, 20–22]:

$$V_{PseAAC} = (x_1, \cdots, x_{20}, x_{20+1}, \cdots, x_{20+\lambda}), \tag{2}$$

in which

$$x_\kappa = \begin{cases} \dfrac{f_\kappa}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (1 \le \kappa \le 20) \\ \dfrac{\omega \theta_{\kappa-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (20+1 \le \kappa \le 20+\lambda) \end{cases} \tag{3}$$

where f_i notates the normalized proportion for the 20 different amino acids [23]; θ_j is the j-th tier autocorrelation for the hydrophobicity, hydrophilicity, and side chain mass property sequences; w is the weight coefficient usually assigned as w = 0.05, representing the sequence-order effect [23]; and $\lambda \in \mathbb{N}^+$ (no larger than the sequence length) adjusts the sequence-order effects [23]. The PseAAC reduces to the amino acid frequencies when $\lambda = 0$.

In the grouped amino acid composition (GAAC) [14, 16, 24], the 20 different amino acids are split into five classes (aliphatic: AGILMV, aromatic: FWY, positively charged: HKR, negatively charged: DE, and uncharged: CNPQST), and the GAAC for amino acid type i is formulated as the percentage of the amino acid type i in its class [14]: $v_i = \frac{N_i}{N}$, in which N_i is the count of amino acid type i and N is the amount of amino acids in the protein sequence belonging to its class [14]. The amphiphilic pseudo amino acid composition (APAAC) [10, 11] incorporates the correlation functions and partial arrangement effects into the PseAAC by using the hydrophobicity and hydrophilicity properties [8, 16, 21, 25]. The enhanced amino acid composition (EAAC) [8], the PseAAC of distance pairs [14] and the pseudo amino acid based on K-nearest neighbors (K-PseAA) [26, 27] are all related derivative features.

The natural vector (NV) is a 60-dimensional feature vector proposed by Yu et al. [7] that accounts for both the amino acid composition and sequence order in a protein sequence [7]:

$$\langle n_A, n_R, \ldots, n_V, \mu_A, \mu_R, \ldots, \mu_V, D_2^A, D_2^R, \ldots, D_2^V \rangle. \tag{4}$$

It is composed of three parts: the composition number for the 20 kinds of amino acids, the mean distance between each type i amino acid and the initial amino acid (regarded as the origin), and the second-order normalized central moments of the distance between every type i amino acid and the origin [7].

The accumulated natural vector (ANV) [16] method adds additional information on the covariance between the amino acids [16], where the former two parts are the composition number and mean positions of the 20 different amino acids, whereas the third and fourth parts are the variance and covariance, respectively, for each of the amino acids [7].

An analogous feature defined by Carr et al. [28] also considers the compositional, centroidal and distributional parameters of a given protein sequence [28]. The positional parameters evaluate the amino acid proportions apart from those expected,

while the centroidal parameters describe the amino acid tendency situated in a peculiar region of the sequence. The distributional parameters [28] analyze the clustering of amino acids along the sequence.

The frequency profile M [29] is a 20 × L frequency matrix $M = \left(m_{i,j}\right)_{20\times L}$, where $m_{i,j}$ denotes the probability of amino acid type i appearing at position j, i = 1,2, … ,20, and j = 1,2, … ,L (L is the sequence length). The element sum in each column is 1 [29].

The quasi-sequence-order (QSOrder or QSO) [30] quantifies the arrangement effects using the Grantham chemical [31] and Schneider-Wrede physicochemical distance matrices [32]:

$$X_a = \frac{N_a}{\sum_{a=1}^{20} N_a + w\sum_{d=1}^{30}\tau_d}, a = 1, 2, 3, \ldots, 20 \quad (5)$$

$$X_d = \frac{w\tau_{d-20}}{\sum_{a=1}^{20} N_a + w\sum_{d=1}^{30}\tau_d}, a = 21, 2, 3, \ldots, 50 \quad (6)$$

$$\tau_d = \sum_{j=1}^{L-d}\left(dist_{j,j+d}\right)^2, d = 1, 2, 3, \ldots, 30 \quad (7)$$

where N_a notates the normalized amino acid frequency, $dist_{j,j+d}$ notates the distance between the jth and the j + dth amino acids, and w denotes the weight factor and is usually equal to 1 [33]. The other arrangement features include the sequence-order-coupling number [8], the positional residue match score [34], and the distance distribution of repeats (DDR) [35].

2.2 Physicochemical property features

Physicochemical properties are crucial amino acid characteristics that have a greater influence on protein similarity analyses than simple amino acid alphabetical orders do [36]. Kidera et al. performed a comprehensive statistical analysis of hundreds of physical properties and identified 10 representative properties for the different property clusters [37]. Rackovsky used these properties to develop the average physical property feature [38], which succeeded in identifying the different structural groups in CATH. The graphical features in later Section 3.1 also utilize the physical properties to make graph representations and extract graphical features for protein classification [2].

AAindex is a famous database that stores over 566 physical property indices (in its current version, ver. 9.2) for amino acids and amino acid pairs [2]. Many studies have used the AAindex database to develop new physical property features [8, 9, 11, 12, 15, 27, 39]. Tung and Ho [39] mapped each amino acid to a 531-dimensional feature by utilizing the physical property values in the AAindex database [8]. Saha et al. innovated a 160-dimensional feature by employing eight high-quality amino acid indices [13, 40].

The averaged property factor (APF) extracts a 10-dimensional feature for a given sequence S by using ten structure-related amino acid properties [38]:

$V_S = \left(\left\langle f^{(1)}\right\rangle_S, \left\langle f^{(2)}\right\rangle_S, \ldots, \left\langle f^{(10)}\right\rangle_S\right)$, where $\left\langle f^{(m)}\right\rangle_S = \frac{1}{N_S}\sum_{n=1}^{N_S} f_n^{(m)}$ is the mth property mean, N_S notates the sequence length, and $f_n^{(m)}$ is the mth property value for the n-th amino acid, $1 \le m \le 10$.

The Z-scale encoding [8] employs five physicochemical descriptors to represent the amino acid features [8]. The amino acid scale-based features [9, 27] implement amino acid propensity scales to quantify the properties of amino acids into a quantitative value, which can be extracted by the ProFET algorithm [27]. The protein sequences are then represented as time series, which use varying size sliding windows to capture extra features.

The solvent-accessible surface area (ASA) [15], produced by the neutral network program RVP-Net [41], represents the solvent-accessible area proportions for each amino acid while accounting for the amino acid composition in the neighborhood, whose values are normalized between zero and one [15]. Composite physicochemical properties (CPPs) [42], generated by Pfeature [35], utilize 16 physicochemical properties of amino acids (acidicity, aliphaticity, aromaticity, basicity, charge, cyclic, hydrophilicity, hydrophobicity, hydroxylic, neutral (pH), nonpolarity, polarity, small and large, sulfur content, and tiny) to project a protein sequence into a 16-dimensional Pfeature [42].

The large number of physicochemical properties may result in high dimensions and high computational complexity, and critical properties can be used to classify the amino acids into reduced groups, where an average or coding strategy can be used to generate efficient features with reduced dimensions. Reduced amino acid alphabets (RAAAs) [22] categorize the amino acids into five groups in terms of their physicochemical properties [22]. The twenty-one-bit feature (TOBF) reduces the normal amino acid alphabet [9] into three clusters for each of the seven physicochemical properties. Each residue in the sequence is deciphered as a 21-bit binary vector (if the amino acid fits the cluster, then the bit is set to 1 and 0 otherwise) [9]; then, each sequence of length N is denoted as a $21 \times N$-dimensional feature [43]. Similarly, the encoding based on a grouped weight (EBGW) [44] separates the 20 amino acids into seven classes (C1 (hydrophobic): AFGILMPVW; C2 (polar): CNQSTY; C3 (positively charged): KHR; C4 (negatively charged): DE; combined H1 (C1 + C2); combined H2 (C1 + C3); and combined H3 (C1 + C4) [44]); and uses sliding windows and averages to define the features [44]. The qualitative properties of amino acids (Qualc) [45] categorize the amino acids with respect to their qualitative properties; amino acids of the same class are depicted by the same one-hot coding [45]. Similar quantitative properties of amino acids (Quanc) [45] deploy quantitative properties such as the isoelectric point, molecular mass, van der Waals volume, etc., to characterize the features. The conjoint triad (CTriad) [46] engenders a 343-dimensional feature for each protein sequence by grouping the amino acids with respect to the dipole and side chain volume properties [46].

2.3 Features based on local sequence units

The above features quantify the involvement of each amino acid. However, the dynamic influences between amino acid pairs or sequence segments are also important sequential characteristics that can be captured by features defined on local sequential units. In this subsection, features defined for the local units are divided into dipeptide and amino acid pair features and K-mer (K-gram) features.

2.3.1 Features based on dipeptide and amino acid pairs

A dipeptide consists of two linked amino acids, which admit 400 amino acid combinations. The amino acid pair composition (AAPC) [15] and dipeptide composition (DPC) [47] are the simplest dipeptide features that describe the percentage of

dipeptides divided by the number of dipeptide combinations: $V_{AAC} = (v_1, v_2, v_3, \cdots, v_{400})$, where $v_i = \frac{r_i}{N}$ is the composition percentage for dipeptide type i, r_i is the composition number of dipeptide type i in the sequence, and N denotes the amount of dipeptide combinations [10–12]. The grouped dipeptide composition (GDPC) [13] is the grouped version of the DPC; it divides the amino acids into five clusters in terms of their physicochemical properties (aromatic(g1): WYF; positive charge (g2): HKR; aliphatic (g3): AIMGLV); uncharged(g4): CTPSQN; and negatively charged(g5): DEG), generating a 25-dimensional feature profile for each protein sequence [13]: $f(i,j) = \frac{M_{ij}}{L-1}$, where M_{ij} stands for the amount of dipeptides with amino acids in groups i and j [13]. The dipeptide deviation from the expected mean (DDE) [48] is made up of three parts, namely, the dipeptide composition (DPC), theoretical mean (TM) and the theoretical variance (TV) for a given protein sequence [49]. Similar features include the pair frequencies [50] and the Grantham matrix [51]. The latter is a 20 × 20-dimensional amino acid pair matrix characterizing the differences between the composition, molecular volume and polarity properties [51].

For amino acid pairs with k-spaced intervals, common features include the g-gap dipeptide composition (g-Gap DC) [52], the composition of the k-spaced amino acid pairs (CKSAAP) [53], and the composition of k-spaced amino acid group pairs (CKSAAGP) [10]. The adaptive skip dipeptide composition (ASDC) [14] is an adapted version of the DPC; it describes all pertinent data between adjacent and intervening residues [14]. Reduced alphabet dipeptide composition (RADP) [14] is based on the k-spaced amino acid pair proportions, accompanied by the amino acid composition and distance pairs in the PseAAC [14]. The normalized Moreau-Broto (NMBroto) [14] also describes the k-spaced amino acid pair frequencies with the inclusion of the amino acid compositions [14]. Other amino acid pair features include, for example, the DPC-position-specific scoring matrix (PSSM) [54], k-separated-bigrams-PSSM [55], and the trigram-PSSM [56].

2.3.2 Features based on K-mers (K-grams)

Another sequence building block is the K-mer (a string of K adjacent amino acids in protein sequences), also known as the K-tuple or K-gram (often used in natural language models). A larger K will cause the K-mer (or K-gram) combinations to increase in power, which results in a low frequency for each K-mer combination, and three adjacent amino acids (tripeptides), that is, K = 3, are essential units in characterizing the proteins. The tripeptide composition (TPC) [8], grouped tripeptide composition (GTPC) [10], CTF [13], and tri-gram-PSSM [56] are popular tripeptide features. The CTF categorizes amino acids into seven groups (AVG, FLIP, TSYM, HQNW, RK, DE, and C) according to their physicochemical properties and treats three adjacent amino acids (traid) categorized in the same group as identical, resulting in a 343-dimensional feature accounting for all sets of triads and their frequencies in the given sequence [13].

The K-mer method [57] defines a vector consisting of the frequencies for the K adjacent amino acids. The K-string dictionary [58] also considers the probability of K-mers. Since the K-string dictionary size (20^K) grows exponentially, a new K-string probability vector is defined to reduce the high dimensions, where the size of the K-string dictionary equals the quantity of K-string combinations in the given protein sets, and a probability vector with its dimension equal to the K-string dictionary size is defined to depict the probability of appearance for each K-string [58]. The LOGO

feature encodes a sequence segment (K-mer or K-string) with respect to the amino acid frequency at each position [44]. "Mirror" K-mers [59] account for the K-mer groups in various combinations, for example, the amino acid combination arginine-lysine RK is grouped with the KR [59] for K = 2. Related features also include the overlapping K-mers [59] and K-tuple features.

The pseudo-K-tuple reduced amino acid composition (PseKRAAC) [8] deploys multiclustering methods to define the weighted frequency of the residues at each position [14]. The enhanced amino acid composition (EAAC) [49] is a window-based version of AAC that accounts for the composition of every amino acid in a size n window in a length L sequence ($n \leq L$) [60]; here, the window of length n can be viewed as a K-mer with K = n. Analogously, the segment pseudo amino acid composition (S-PseAAC) [22] is defined to depict the local regional situation by introducing segmentation in the PseAAC.

The K-mer natural vector [61] is a more delicate method; it concatenates the counting, mean distance and the distance variance of the K-mers, which results in a 3×20^K-dimensional vector. A compound K-mer natural vector is further proposed to fuse the K-mer natural vector with K = 2 and 3 [61].

Similar features are defined for n-grams (K-mers with K = n). An n-gram feature [44] is defined to record the relative frequencies of all probable n-connected amino acids in the sequence [44], which is sparse when the sequence is short. An improvement is provided by the k-skip-n-gram model (SkipF) [62], embedding distances into the n-gram model, which provides the n residue composition with distances $\leq$k in the sequence, that is, it accounts for the n residues with distances ranging between 1 and k in the given protein sequence. In this model, k evaluates the deviation between two arbitrary residues, which is usually limited to no more than 5 in applications due to high complexity. The adaptive k-skip-n-gram feature addresses this problem by setting k as each protein sequence length, which better reflects the varying distances in the datasets [63]. The adaptive skip-n-gram feature [43] accounts for the composition of all n residues within distance N in a length N sequence, and a proper choice for n is n = 2, which yields 400 dimensions for the adaptive k-skip-2-gram feature [63].

2.4 Alignment and similarity scoring features

Features can also be deduced from the position-specific scoring matrix (PSSM) profile [64]. The original PSSM for a length m sequence is an $m \times 20$ matrix [29]:

$$\mathrm{PSSM}_{\mathrm{original}} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ p_{i,1} & p_{i,2} & \cdots & p_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ p_{L,1} & p_{L,2} & \cdots & p_{L,20} \end{bmatrix}_{L\times 20} \tag{8}$$

where $p_{i,j}$ records the score of the residue at the ith position being mutated to the type j residue during evolution ($1 \leq i \leq L$, $1 \leq j \leq 20$). By summing up each row for the same kind of amino acid, the PSSM profile is transformed into a 20×20 matrix, which is further converted to a 400-dimensional feature vector, whose elements are then normalized by sequence length and $\frac{1}{1+e^{-x}}$ to scale in [0,1] [65].

Profile-based cross-covariance (CC-PSSM) [66] accounts for the property difference between every pair of residues with a lag in sequence [66] and converts the PSSM matrices of various sizes into 760-dimensional vectors. Parallel correlation pseudo amino acid composition (PC-PseAAC) [67] encodes the local and global sequence arrangements into feature vectors [17]. Related features include the general PC-PseAAC (PC-PseAAC-General) [67], series correlation pseudo amino acid composition (SC-PseAAC) [25], and general SC-PseAAC (SC-PseAAC-General) [25], as well as the normalized PSSM [68], pseudo position-specific scoring matrix (Pse-PSSM) [49], local Pse-PSSM features [68], spatial neighbor-based position-specific scoring matrix (SNB-PSSM) [69], and bigram position-specific scoring matrix (bigram PSSM) [20].

The position-specific frequency matrix (PSFM) computed by MSA (produced by PSI-BLAST search) [70] notates the frequencies of the 20 different amino acids situated at peculiar positions in the evolutionary process, with advantages in analyzing low-similarity proteins [70]. The sequence-order frequency matrix (SOFM) [71] consists of the scores calculated from the K-mer probabilities at each sequence position using MSA. SOFM variants are derived using different sequence-order extraction methods or weighted schemes [71]. The SOFM can also be combined with the top-n-gram method to convert the SOFM into a fixed-length vector [71]. The block substitution matrix (BLOSUM) and BLOSUM62 [42] are broadly utilized substitution scoring matrices for protein alignment [42]. Other derivative features include the BLOSUM matrix-derived descriptor [72], the two-dimensional features [51], position-specific symbol composition (PSSC) and the position-specific frequency matrix (PSFM).

The hidden Markov model (HMM) feature [73] consists of three local diversities, seven transition probabilities, and 20 emission frequencies [74]. The HHblits profile [64] is a statistical model derived from the hidden Markov model and it can predict the sequence mutation probability [64].

The alignment profiles can also be evaluated by using correlation measures. For example, the k-nearest neighbor (KNN) feature describes the sequence segment characteristics regarding the homology with the k samples [44]. The autocorrelation function (ACF) projects the protein sequence into numeral sequences by using amino acid indices and expresses the features as the self-correlation function eigenvector for the lagged numerical sequences [20]. The autocorrelation fold (ACC fold) [75] uses auto and cross-covariance transformations to transform the PSSMs into vectors. The covariance [50] calculates the covariance and frequency between every pair of residues. Similar features include Geary autocorrelation features, Moran autocorrelation features, and normalized Moreau-Broto autocorrelation features [8].

2.5 Related composite features

Composition, transition, and distribution (CTD) is a composite feature that divides the amino acids into three clusters with respect to their charge, hydrophobicity, normalized van der Waals volume, polarity, polarizability, secondary structure, and solvent accessibility properties [76]. The composition (C) evaluates the proportion of each of the three clusters, the transition (T) estimates the proportion of a certain amino acid property progressed by another property, and the distribution (D) component presents five values for the three clusters and records the proportions of the sequence in which 25, 50, 75 and 100% of the amino acids in a specific property are located [77]. The composition-transition-distribution-composition (CTDC) [10] can

be used to evaluate the percentage of transitions for the three kinds of residue pairs that are sorted in terms of their physicochemical properties [11]. The composition-transition-distribution-transition (CTDT) [10] and enhanced grouped amino acid composition (EGAAC) [49] are both related features.

3. Advanced features based on multidisciplinary techniques

Recent progress in feature extraction methods has adopted various kinds of techniques to improve feature presentation. In this section, popular features are summarized in terms of their techniques, for example, the graphical features, numerical coding features, probabilistic and information-based features, and machine learning features. The division is based upon feature definitions, which is not rigid because many features may employ several different techniques.

3.1 Graphical features

Graphical features refer to features generated from graphical representations. The protein map is a graphical representation feature introduced by Yu et al. [78]. It first constructs a graphical representation of a protein sequence by using vectors connected end-to-end, where each amino acid admits one vector with an x-coordinate equal to 1, and the y-coordinate is standardized in $[-1, 1]$. The y-coordinate is ranked according to the hydrophobicity scale value. A protein sequence then corresponds to a graphical curve connecting the points $(1, y_1)$, $(2, y_2)$, ..., (n, y_n). The protein map is formulated as a 2-dimensional vector (M_1, M_2) containing the first two moments defined by $M_j = \sum_{i=1}^{n} \frac{(x_i - y_i)^j}{n^j}$, where $j = 1, 2, \cdots, n$ denotes the amino acid quantity in the protein sequence and (x_i, y_i) is the i-th amino acid coordinate in the graphical curve. A new protein map feature was later introduced by using 10 graphical representations for ten important physical properties of amino acids; the new protein map is a 30-dimensional feature defined by the weighted first three moments of the ten graphical curves [79].

FEGS constructs 158 3D graphical curves for protein sequences using 158 physical property indices [2]. It is a 578-dimensional vector containing three parts, namely, the percentages of the 20 different amino acids, the frequency of the 400 combinations of amino acid pairs, and the 158 leading eigenvalues for the 158 $L \times L$-dimensional matrix (L is the protein sequence length) constructed from the 158 3D graphical curves for the 158 physical property indices in the AAindex database. The 158×158-dimensional L/L matrix is a nonnegative symmetric matrix, with the off-diagonal elements $M_{i,j}$ computed as the quotients of the Euclidean distances between points P_i and P_j in the spatial curve divided by the path length connecting P_i and P_j, and all diagonal elements vanish [2].

Another graphical feature employs nine physicochemical properties [80] (flexibility, hydrophilicity, hydrophobicity, irreplaceability, mass, pI, pK1, pK2, and rigidity) to cluster the 20 amino acids into four clusters (nonpolar residues: AVLIPFWM; polar residues: GSTCYNQ; basic residues: KRH; acidic residues: DE). Amino acids belonging to each group are mapped to a 3D point, where the points in the same property group lie on the same conical surfaces. An 18-dimensional feature is formulated to characterize the protein sequence, which contains 6 sets of 3D coordinate averages; the first

four sets are the 3D coordinate averages for the four different amino acid classes; and the last two sets are the averages and the accumulated averages for the 3D coordinates of all amino acids in the sequence [80].

3.2 Numerical coding features

Popular coding features include binary one-hot representations, the numerical conversion of amino acids via nonnegative integers, and the numerical coding of amino acids with respect to their property groups.

Numerical representation of amino acids (NUM) [49] involves reversing the amino acid sequences into numerical sequences using integers from 1 to 20, and the dummy amino acid O is converted to 21 [49]. The one-hot encoding [74] and binary profile feature (BPF) [45] are similar features, where every amino acid is denoted by a 20-dimensional binary vector with only one element equal to 1 and the others equal to 0. For a sequence of L residues, one-hot encoding has admits $L \times 20$ dimensions [74]. When analyzing sequence segments, the binary encoding of amino acids (BINA) [49] transforms each amino acid in a given segment into a 21-dimensional binary vector [44], which results in a 525-dimensional feature for a segment of length 25.

The positional weighted matrix (PWM) [15] characterizes the sequence fragments by using sliding windows of size 21 and one-hot coding [15]. The multiscale local descriptor (MLD) feature decomposes the protein sequence into multiple sequence segments of different lengths and employs a binary coding scheme to depict local overlapping regions [81]. Protein sequence encoding (PSE) [82] encodes the 20 different amino acids by using six exchange groups (E1: HRK, E2: DENQ, E3: C, E4: STPAG, E5: MILV, and E6: FYW) [82]. The overlapping property features (OPF) [43] classify the 20 standard amino acids into ten groups (aliphatic: IVL, aromatic: FYWH, charged: KHRDE, hydrophobic: AGCTIVLKHFYWM, negative: DE, polar: NQSDECTKRHYW, positive: KHR, proline: P, small: PNDTCAGSV, and tiny: ASGC) and encode every amino acid by a 10-dimensional vector of 0 and 1, indicating whether this amino acid is in the group [7]. The qualitative properties of amino acids (Qualc) [45], the encoding based on the grouped weight (EBGW) [44], and the twenty-one-bit features (TOBF) [43] mentioned earlier are all coding-based features.

3.3 Probabilistic and information-based features

Information theory provides useful concepts for describing amino acid distributions in protein sequences, which contain uncertain information due to evolution. Many methods take the advantage of various entropy concepts to characterize protein sequences into probabilistic and statistical features.

The information-based statistics [59] is an information-entropy-based feature that aims to depict the amino acid distribution in the sequence; it includes three parts, namely, the total entropy per letter, the binary autocorrelation and the features for the selected letters. The information gain score evaluates the information transformation between the states by means of the information gain score [43]. The information entropy feature is formulated as the Shannon entropy of the amino acid and dipeptide compositions [18]. Similarly, the Shannon entropy and the relative Shannon entropy features quantify the amino acid distributions from the Shannon entropy and the relative Shannon entropy aspects [43]. The sequence entropy [83] feature is defined by the normalized Shannon entropy of the residue type frequencies that favor a given position in the sequence. Other information-based features include the entropy

density (Den) [39], the Shannon entropy of residues (SER) and the Shannon entropy of properties (SEP) [46]. Kolmogorov complexity is also an information concept used to describe the complexity of word strings. When the protein sequences are treated as word strings, the Kolmogorov complexity (KOL) can also be applied to quantify the amino acid distributions in a given sequence [83].

3.4 Machine learning features and feature representation learning techniques

In recent years, artificial intelligence algorithms flourish, bringing fresh ideas to every discipline. In bioinformatics research, a new approach, feature representation learning, emerges, reforming feature extraction studies from new perspectives.

Feature representation learning approaches use machining learning techniques to find feature representations from computational aspects. Wei et al. reported a new feature learning strategy with various successful prediction applications [9]. This method constructs an initial feature pool from eight traditional feature encoding schemes, such as AAC and CTD. Then, a feature learning model is developed on the basis of the extremely randomized tree (ERT) using a benchmarking dataset and tenfold cross-validation (CV), where 51 prediction models are generated as the baseline model. For every peptide, each baseline model predicts its probability for antihypertensive peptides (AHTPs), which is used as a feature. The probabilities from all the prediction models form a new feature vector, FVERT [9]. Another model, mACPpred 2.0, employs 16 pretrained NLP-based embeddings and 15 conventional features for feature representations [10].

Convolutional neural networks (CNNs) and recursive neural networks (RNNs) are primary models for protein feature extraction. CNNs process data in matrix form and generate features *via* feature mapping [84]. The sequence embedding module introduced in [85] converts the original sequence into a digital sequence and then encodes it as picture-like data, which can be further processed by the TextCNN model with eight different-shaped kernels [85]. Unlike CNNs, RNNs emphasize sequence information but suffer from time lag problems in training. The long short-term memory recurrent neural networks (LSTM-RNNs) target this problem in RNNs [84].

Word embedding algorithms such as word2vec [11] are efficient methods for feature representation [11]. The word embedding-based features treat protein sequences as sentences and engender a K-mer corpus by sliding windows. Each word is then embedded into an N-dimensional vector by combining word2vec with a skip-gram model. Thus, each sequence is depicted by the mean corpora in the sequence [86]. This word2vec representation for proteins can be generated by the Gensim library [11]. The flexible HMSRN [34] model employs a two-layer CNN module and language model embeddings to extract amino acid features [34]. In the UniRep [87] method, the one-hot coding of amino acids is input to the LSTM, and a mean pooling operation is employed to generate a 1900-dimensional feature for the sequence [88]. The SeqVec and PortTrans models are also protein language models for amino acid feature extraction [34]. The SeqVec model is pretrained on the UniRef50 dataset, which produces a 1024-dimensional vector for each amino acid [34].

The BiLSTM embedding, SSA embedding and UniRep embedding are three popular deep learning embedded representation methods [89]. For BiLSTM embedding, which consists of two reversed unidirectional LSTMs, a linear transformation is performed on the combination of the forwards and backward hidden states, which produces an $L \times 3605$ dimensional matrix for the protein sequence with length L. The pretrained SSA embedding [89] uses three bidirectional LSTM layers, which produce

a 121-dimensional embedding vector [90]. The pretrained protein unified representation (UniRep) embedding method [91] implements the hybrid structures of multiplicative RNNs (mRNNs) and the LSTM architecture [92] to process the 10-dimensional one-hot coding of amino acids, which produces a 1900-dimensional embedding vector. In [89], pretrained BiLSTM embedding, pretrained SSA embedding, and pretrained UniRep embedding were integrated to extract protein sequence features [89]. DeepPPISP utilizes sliding windows and deep learning models to learn a 49-dimensional feature (20 dimensions for PSSM, 9 dimensions for secondary structure, and 20 dimensions for raw protein sequences) of protein sequences [93].

The transformer encoder [94] utilizes a self-attention mechanism to account for the positional relationships of amino acids within the sequence [94]. The pretrained protein language model ProtT5-XL-UniRef50 (ProtT5) [42] projects protein sequences into feature vectors using a transformer-based 24-layer language model in a self-supervised manner. PSPred-ALE transforms each amino acid into a unique dense vector by employing self-attention mechanisms and adaptive learning techniques, and protein segments are converted into a unique matrix, where the absolute and relative positional information of amino acids are encoded by positional embedding [95]. Other feature methods include AESNN3 [96], multisense-scaled embedding [97], and learnable position embedding [98].

3.5 Features based on other techniques

Other techniques, such as the discrete wavelet transform and k-nearest neighbor methods, can be used for sequence feature extraction. In [18], a discrete wavelet transform was applied to convert numerical physicochemical property sequences, and the statistical features at each scale were calculated as protein sequence features [18]. The term frequency-inverse document frequency (TF-IDF) [49] separately computes the term frequency (TF) and the inverse document frequency (IDF), which are multiplied as a product feature [49]. Other feature extraction techniques include the functional domain (FunD) composition [99], the KNN protein [8] and KNN peptide [100] features using the k-nearest neighbor method, and the residue fluctuation dynamics feature based on the Gaussian network model (GNM) [69].

4. Feature-generation platforms and hybrid features

In this subsection, popular protein feature-generation platforms such as the Pse-in-One, PseAAC-General, iFeature, pFeature, modlAMP, and POSSUM are introduced.

Pse-in-One [17] (http://bioinformatics.hitsz.edu.cn/Pse-in-One/home/) is a web platform for biological sequence analysis that can generate desired pseudo protein sequence features such as the abovementioned PseAAC and PC-PseAAC. The PseAAC-General [101] program is able to generate eight PseAAC vector modes, including the autocovariance and basic K-mer features, cross covariance and autocross covariance features, the PC-PseAAC and SC-PseAAC, as well as PC-PseAAC-General and SC-PseAAC-General features [101]. iFeature [8] (http://iFeature.erc.monash.edu/) is a Python toolkit that has 18 major sequence coding strategies encompassing 53 different features, which cover the majority of the features mentioned in Section 2. Pfeature [35] (https://github.com/dataprofessor/Pfeature) software contains five main modules, namely, the composition, binary configuration file, evolutionary information, structure, and pattern modules. Pfeature [35] can produce simple amino acid, dipeptide and

tripeptide features, along with combined features based on the physicochemical properties, repetition and distribution of amino acids, Shannon entropy features and pseudo amino acid features, and other features, such as autocorrelation, trimesters, quasi-sequence-order (QSO), and binary configuration profiles. ModlAMP [16] (https://pypi.org/project/modlamp/) is a Python software mainly used for peptide sequence generation, peptide feature calculation, and sequence analysis. The peptide feature calculation module is able to calculate various features of peptides, such as the molecular formula, charge, isoelectric point, and instability index. The sequence analysis module performs various statistical and analytical analyses of peptide sequences, which can facilitate researchers gain a better understanding of proteins. The POSSUM [102] (http://possum.erc.monash.edu/) toolkit generates numerical sequence features according to PSSM transformations. The row transformations produce seven features, such as the AAC-PSSM, D-FPSSM and smoothed-PSSM. The column transformations generate five features: DPC-PSSM, EEDP, k-separated-bigrams-PSSM, TPC, and trigram-PSSM. The mixed row and column transformations produce six features, including the EDP, DP-PSSM, Pse-PSSM, and RPSSM. The POSSUM can also generate combinations of the above features [103]. Other protein feature extraction platforms include PROFEAT [104], PseAAC-Builder [102], Propy [105], Protr/ProtrWeb [106], Rcpi [107] and PseKRAAC [108].

In practice, individual features usually cannot cover all aspects of protein sequences. A better method is to concatenate different features to form composite or hybrid features. Hybrid features tend to outperform individual features by integrating comprehensive information from various aspects [9]. Hybrid features can be generated by feature fusion or concatenation. Manavalan et al. used a linear combination of five traditional features to generate hybrid features [9]. Weng et al. used cross-validation to assess the predictive power for each individual feature and combined all the different features for prediction improvement [15]. Wei et al. used a multiple-feature fusing strategy to combine the ASDC, PC-PseAAC, physicochemical properties (PPs), and SC-PseAAC features [109]. Other fusion techniques may rank the different features by their accuracy scores and use these scores to either select the most efficient features or fuse all the features by using score weights.

Hybrid features may attain high computational complexity because of their high dimensionality. Feature optimization or selection approaches can be implemented to reduce feature dimensions. In particular, the iFeature toolkit provides a series of feature clustering, dimensional reduction, and selection algorithms [8] that can be used for reducing the dimensionality of hybrid features.

5. Applications

In this section, we introduce some popular protein classifiers and typical areas of protein feature applications.

5.1 Methods for protein classification using feature vectors

Since proteins with strong sequence homology may have similar structures or functions, protein sequence feature vectors can be used to perform protein classification or predictions. A simple classification method involves treating the high-dimensional features as points in high-dimensional real spaces and calculating the Euclidean distances or norms between the feature points, where a short distance

indicates close relations and strong sequence homology, and a large distance implies far relations and low sequence similarity. Distance methods are often used in protein evolutionary classification [7], where the taxa of bacteria or viruses can be correctly classified *via* distance.

Geometric methods such as the convex hull [110] and mean-square-error (MSE) hyperplane [111] are often used for feature point classification. These geometric classifiers use geometric boundaries such as convex hulls or hyperplanes to separate high-dimensional feature points into disjoint regions. Similar classifications can be performed by clustering methods such as the k-nearest neighbor (KNN) method, fuzzy-C-means (FCM), and the clustering methods (hierarchical clustering, mean shift, DBSCAN, affinity propagation, etc.) that are provided in the iFeature toolkit [8].

In recent analyses, machine learning methods have become prevalent. Machine learning methods use neural networks or large language models to train and test feature datasets. Training data with class labels are employed to train the neural network coefficients, which are applied to testing datasets to predict the classes. Typical machine learning approaches encompass random forest (RF), support vector machine (SVM), and ensemble deep learning methods. In practice, ensemble classifiers often outperform individual classifiers with higher classification accuracy. Recently, large language models (LLM) have emerged. As mentioned in Subsection 3.4, protein sequences can be deemed as sentences with amino acid characters as words [11], from which large language models (LLM) learn evolutionary information from the semantic contexture of protein sequences.

5.2 Applications

Implementing the various protein sequence features can realize many applications. Here, we generalize the protein feature applications from three aspects: protein evolutionary classification, protein structural prediction and fold recognition, and protein functional prediction.

Protein evolutionary classification is one of the essential applications of the protein sequence features. For example, the classification of subtypes of viruses, such as Zika, Ebola, influenza A virus, coronavirus, and human rhinovirus, has been performed. The taxonomic classification of animal species or bacterial lineages from the protein sequences that are translated from genomes, for example, mammalian mitochondrial proteins [112], beta-globins of animal species [2] and SAR11 clade marine bacteria [112], and the classification of subtypes for enzymes and proteins with particular functions, for example, the classification of different subtypes of protein kinase C [112], and the classification of transferrin sequences and antifreeze proteins [2].

Since protein sequence decides the protein structure, protein sequence features can be used for protein secondary or spatial structure prediction [83, 96] or fold recognition [71]. For example, the HMM and UniRep features can be used for peptide structure prediction [34, 112], and the PSSM and Kolmogorov complexity (KOL) features, as well as the combination of the SOFM and top-n-grams, can be implemented for protein fold recognition [71, 83].

Protein sequence features can also be implemented in protein functional predictions. These include the prediction of certain sites in the protein sequence. These methods include the prediction of protein-protein interaction (PPI) sites [93]; the prediction of phosphorylation [43], malonation [49], malonylation [44] and carbonylation sites [54]; and the prediction of intrinsically disordered regions in yeast [113].

These features can also be employed to distinguish epitopes from nonepitopes in a cell [49]. Other applications include protein-protein interaction (PPI) prediction [93] and the prediction of ligand-binding proteins, DNA-binding proteins [84], membrane proteins [49], DNA-binding proteins [68], anticancer peptides [114], quorum-sensing peptides [115], cell-penetrating peptides [63] and extracellular matrix (ECM) proteins [18].

6. Conclusion

The feature extraction methods reviewed in this chapter depict the protein sequence features from diverse aspects. Feature learning models, including deep learning and large language models, are new emerging methods implementing artificial intelligence. Various features can be generated from feature-generation platforms and combined to form hybrid features for better classification accuracy in real practice. General feature classification methods and areas of protein classification applications are also presented to better understand and apply the features.

Acknowledgements

We acknowledge the library facilities the Beijing University of Chemical Technology provided that support this review study.

Conflict of interest

The author declares no conflict of interest.

Author details

Xiaogeng Wan
College of Mathematics and Physics, Beijing University of Chemical Technology, Beijing, China

*Address all correspondence to: wxgbj88@sina.com

References

[1] Wang J, Wang Z, Tian X. Bioinformatics: Fundamentals and Applications. Beijing: Tsinghua University Press; 2014

[2] Mu Z, Yu T, Liu X, Zheng H, Wei L, Liu J. FEGS: A novel feature extraction model for protein sequences and its applications. BMC Bioinformatics. 2021; **22**(1):297. DOI: 10.1186/s12859-021-04223-3

[3] Wan X, Tan X, Cao J. A study on novel amino acid pair features for protein evolutionary classifications. Computational Biology and Bioinformatics. 2024;**12**(1):18-31. DOI: 10.11648/j.cbb.20241201.13

[4] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: Benefits, applications, and tools. Genome Biology. 2017;**18**(1):186. DOI: 10.1186/s13059-017-1319-7

[5] Liu B, Wu H, Zhang DY, Wang XL, Chou KC. Pse-analysis: A python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. Oncotarget. 2017;**8**(8):13338-13343. DOI: 10.18632/oncotarget.14524

[6] Shen H, Chou K. Pse AAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. Analytical Biochemistry. 2008;**373**(2):386-388. DOI: 10.1016/j.ab.2007.10.012

[7] Yu C, Deng M, Cheng SY, Yau SC, He RL, Yau SST. Protein space: A natural method for realizing the nature of protein universe. Journal of Theoretical Biology. 2013;**318**:197-204. DOI: 10.1016/j.jtbi.2012.11.005

[8] Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. iFeature: A python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018;**34**(14): 2499-2502. DOI: 10.1093/bioinformatics/bty140

[9] Manavalan B, Basith S, Shin TH, Wei L, Lee G. mAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics. 2019;**35**(16):2757-2765. DOI: 10.1093/bioinformatics/bty1047

[10] Sangaraju VK, Pham NT, Wei L, Yu X, Manavalan B. mACPpred 2.0: Stacked deep learning for anticancer peptide prediction with integrated spatial and probabilistic feature representations. Journal of Molecular Biology. 2024;**436** (17):168687. DOI: 10.1016/j.jmb.2024.168687

[11] Wang C, Zou Q. Prediction of protein solubility based on sequence physicochemical patterns and distributed representation information with DeepSoluE. BMC Biology. 2023;**21** (1):1-11. DOI: 10.1186/s12915-023-01510-8

[12] Rajput A, Gupta AK, Kumar M. Prediction and analysis of quorum sensing peptides based on sequence features. PLoS One. 2015;**10**(3): e0120066. DOI: 10.1371/journal.pone.0120066

[13] Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-fuse: Improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics. 2020;**36**

(11):3350-3356. DOI: 10.1093/bioinformatics/btaa160

[14] Shaon MSH, Karim T, Sultan MF, Ali MM, Ahmed K, Hasan MZ, et al. AMP-RNNpro: A two-stage approach for identification of antimicrobials using probabilistic features. Scientific Reports. 2024;**14**(1):12892. DOI: 10.1038/s41598-024-63461-6

[15] Weng SL, Huang KY, Kaunang FJ, Huang CH, Kao HJ, Chang TH, et al. Investigation and identification of protein carbonylation sites based on positionspecific amino acid composition and physicochemical features. BMC Bioinformatics. 2017;**18**(3):66. DOI: 10.1186/s12859-017-1472-8

[16] Yu JC, Ni K, Chen CT. ENCAP: Computational prediction of tumor T cell antigens with ensemble classifiers and diverse sequence features. PLoS One. 2024;**19**(7):e0307176. DOI: 10.1371/journal.pone.0307176

[17] Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-one: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Research. 2015; **43**(W1):W65-W71. DOI: 10.1093/nar/gkv458

[18] Yang R, Zhang C, Gao R, Zhang L. An ensemble method with hybrid features to identify extracellular matrix proteins. PLoS One. 2015;**10**(2): e0117804. DOI: 10.1371/journal.pone.0117804

[19] Pugalenthi G, Kumar KK, Suganthan PN, Gangal R. Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. Biochemical and Biophysical Research Communications. 2008;**367**(3):630-634. DOI: 10.1016/j.bbrc.2008.01.038

[20] Yu B, Qiu W, Chen C, Ma A, Jiang J, Zhou H, et al. SubMito-XGBoost: Predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. Bioinformatics. 2020;**36**(4): 1074-1081. DOI: 10.1093/bioinformatics/btz734

[21] Ariaeenejad S, Mousivand M, Dezfouli PM, Hashemi M, Kavousi K, Salekdeh GH. A computational method for prediction of xylanase enzymes activity in strains of *Bacillus subtilis* based on pseudo amino acid composition features. PLoS One. 2018;**13**(10): e0205796. DOI: 10.1371/journal.pone.0205796

[22] Sikander R, Ghulam A, Ali F. XGB-DrugPred: Computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. Scientific Reports. 2022;**12**(1):1-9. DOI: 10.1038/s41598-022-09484-3

[23] Chou KC. Prediction of protein cellular attributes using pseudo-amino-acid-composition. Proteins: Structure, Function, and Genetics. 2001;**43**(3):246-255. DOI: 10.1002/prot.1035

[24] Lee TY, Lin ZQ, Hsieh SJ, Bretaña NA, Lu CT. Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. Bioinformatics. 2011;**27**(13):1780-1787. DOI: 10.1093/bioinformatics/btr291

[25] Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005;**21**(1):10-19. DOI: 10.1093/bioinformatics/bth466

[26] Cong H, Liu H, Chen Y, Cao Y. Self-evoluting framework of deep convolutional neural network for multilocus protein subcellular

localization. Medical & Biological Engineering & Computing. 2020;**58**(12): 3017-3038. DOI: 10.1007/s11517-020-02275-w

[27] Cong H, Liu H, Cao Y, Liang C, Chen Y. Protein–protein interaction site prediction by model ensembling with hybrid feature and self-attention. BMC Bioinformatics. 2023;**24**(1):456. DOI: 10.1186/s12859-023-05592-7

[28] Carr K, Murray E, Armah E, He RL, Yau SST. A rapid method for characterization of protein relatedness using feature vectors. PLoS One. 2010;**5** (3):e9550. DOI: 10.1371/journal.pone.0009550

[29] Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2014;**30**(4):472-479. DOI: 10.1093/bioinformatics/btt709

[30] Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochemical and Biophysical Research Communications. 2000;**278**(2):477-483. DOI: 10.1006/bbrc.2000.3815

[31] Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974;**185**(4154):862-864. DOI: 10.1126/science.185.4154.862

[32] Schneider G, Wrede P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophysical Journal. 1994;**66**(2 Pt 1):335-344. DOI: 10.1016/S0006-3495(94)80782-9

[33] Chou KC, Cai YD. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochemical and Biophysical Research Communications. 2004;**320**(4):1236-1239. DOI: 10.1016/j.bbrc.2004.06.073

[34] Chen N, Yu J, Zhe L, Wang F, Li X, Wong K. TP-LMMSG: A peptide prediction graph neural network incorporating flexible amino acid property representation. Briefings in Bioinformatics. 2024;**25**(4):bbae308. DOI: 10.1093/bib/bbae308

[35] Pande A, Patiyal S, Lathwal A, Arora C, Kaur D, Dhall A, et al. Pfeature: A tool for computing wide range of protein features and building prediction models. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology. 2023;**30**(2):204-222. DOI: 10.1089/cmb.2022.0241

[36] Randic M. 2-D graphical representation of proteins based on physico-chemical properties of amino acids. Chemical Physics Letters. 2008; **440**(4–6):291-295. DOI: 10.1016/j.cplett.2007.04.037

[37] Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. Journal of Protein Chemistry. 1985;**4**(1):23-55. DOI: 10.1007/BF01025492

[38] Rackovsky S. Sequence physical properties encode the global organization of protein structure space. Proceedings of the National Academy of Sciences of the United States of America. 2009;**106**(34):14345-14348. DOI: 10.1073/pnas.0903433106

[39] Tung CW, Ho SY. Computational identification of ubiquitylation sites from protein sequences. BMC Bioinformatics. 2008;**9**(1):1-15. DOI: 10.1186/1471-2105-9-310

[40] Saha I, Maulik U, Bandyopadhyay S, Plewczynski D. Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino Acids. 2012;**43**(2):583-594. DOI: 10.1007/s00726-011-1106-9

[41] Ahmad S, Gromiha MM, Sarai A. RVP-net: Online prediction of real valued accessible surface area of proteins from single sequences. Bioinformatics. 2003;**19**(14):1849-1851. DOI: 10.1093/bioinformatics/btg249

[42] Fang Y, Xu F, Wei L, Jiang Y, Chen J, Wei L, et al. AFP-MFL: Accurate identification of antifungal peptides using multi-view feature learning. Briefings in Bioinformatics. 2023;**24**(1):1-12. DOI: 10.1093/bib/bbac606

[43] Wei L, Xing P, Tang J, Zou Q. PhosPred-RF: A novel sequence-based predictor for phosphorylation sites using sequential information only. IEEE Transactions on Nanobioscience. 2017;**16**(4):240-247. DOI: 10.1109/TNB.2017.2661756

[44] Zhang Y, Xie R, Wang J, Leier A, Marquez-Lago TT, Akutsu T, et al. Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. Briefings in Bioinformatics. 2019;**20**(6):2185-2199. DOI: 10.1093/bib/bby079

[45] Han B, Zhao N, Zeng C, Mu Z, Gong X. ACPred-BMF: Bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction. Scientifc Reports. 2022;**12**:21915. DOI: 10.1038/s41598-022-24404-1

[46] Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein-protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America. 2007;**104**(11):4337-4341. DOI: 10.1073/pnas.0607879104

[47] Dhanda SK, Usmani SS, Agrawal P, Nagpal G, Gautam A, Raghava GPS. Novel in silico tools for designing peptide-based subunit vaccines and immunotherapeutics. Briefings in Bioinformatics. 2017;**18**(3):467-478. DOI: 10.1093/bib/bbw025

[48] Saravanan V, Gautham N. Harnessing computational biology for exact linear B-cell epitope prediction: A novel amino acid composition-based feature descriptor. OMICS A Journal of Integrative Biology. 2015;**19**(10):648-658. DOI: 10.1089/omi.2015.0095

[49] Mesrabadi HA, Faez K, Pirgazi J. Drug–target interaction prediction based on protein features, using wrapper feature selection. Scientific Reports. 2023;**13**(1):3594. DOI: 10.1038/s41598-023-30026-y

[50] Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics. 2018;**34**(19):3308-3315. DOI: 10.1093/bioinformatics/bty341

[51] Rifaioglu AS, Atalay RC, Kahraman DC, Dogan T, Martin M, Atalay V. MDeePred: Novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery. Bioinformatics. 2021;**37**(5):693-704. DOI: 10.1093/bioinformatics/btaa858

[52] Ding H, Li DM. Identification of mitochondrial proteins of malaria parasite using analysis of variance. Amino Acids. 2015;**47**(2):329-333

[53] Chen K, Kurgan LA, Ruan J. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. BMC Structural Biology. 2007;**7**(1):25. DOI: 10.1186/1472-6807-7-25

[54] Liu TG, Zheng XQ, Wang J. Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. Biochimie. 2010;**92**(10):1330-1334. DOI: 10.1016/j.biochi.2010.06.013

[55] Saini H, Raicar G, Lal S. Protein fold recognition using genetic algorithm optimized voting scheme and profile bigram. Journal of Software. 2016;**11**(8): 756-767. DOI: 10.17706/jsw.11.8.756-767

[56] Paliwal KK, Sharma A, Lyons J. A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. IEEE Transactions on Nanobioscience. 2014;**13**(1):44-50. DOI: 10.1109/TNB.2013.2296050

[57] Liu B, Wang S, Wang X. DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. Scientific Reports. 2015; **5**:15479. DOI: 10.1038/srep15479

[58] Yu C, He RL, Yau SST. Protein sequence comparison based on K-string dictionary. Gene. 2013;**529**(2):250-256. DOI: 10.1016/j.gene.2013.07.092

[59] Ofer D, Linial M. ProFET: Feature engineering captures high-level protein functions. Bioinformatics. 2015;**31**(21): 3429-3436. DOI: 10.1093/bioinformatics/btv345

[60] Wang M, Cui X, Shan L, Yang X, Ma A, Zhang Yusen YB. DeepMal: Accurate prediction of protein malonylation sites by deep neural networks. Chemometrics and Intelligent Laboratory Systems. 2020;**207**:104175. DOI: 10.1016/j.chemolab.2020.104175

[61] Zhang Y, Wen J, Yau SST. Phylogenetic analysis of protein sequences based on a novel k-mer natural vector method. Genomics. 2019; **111**(6):1298-1305. DOI: 10.1016/j.ygeno.2018.08.010

[62] Guthrie D, Allison B, Liu W, Guthrie L, Wilks Y. A closer look at skipgram modelling. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006); 22-28 May 2006; Genoa, Italy. Paris: ELRA; 2006. pp. 1-4

[63] Wei L, Tang J, Zou Q. SkipCPP-Pred: An improved and promising sequence-based predictor for predicting cell-penetrating peptides. BMC Genomics. 2017;**18**(7):1-11. DOI: 10.1186/s12864-017-4128-1

[64] Zhang W, Pan X, Shen H. Signal-3L3.0: Improving signal peptide prediction through combining attention deep learning with window-based scoring. Journal of Chemical Information and Modeling. 2020;**60**(7): 3679-3686. DOI: 10.1021/acs.jcim.0c00401

[65] Altschul SF, Madden TL, Schaffer AA, Zhang J. Gapped blast and psi-blast: A new generation of protein database search programs. Nucleic Acids Research. 1997;**25**(17):14. DOI: 10.1093/nar/25.17.3389

[66] Yu S, Peng D, Zhu W, Liao B, Wang P, Yang D, et al. Hybrid_DBP: Prediction of DNA-binding proteins using hybrid features and convolutional neural networks. Frontiers in Pharmacology. 2022;**13**:1031759. DOI: 10.3389/fphar.2022.1031759

[67] Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001;**44** (1):60. DOI: 10.1002/prot.1072

[68] Wei L, Tang J, Zou Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. Information Sciences: An International Journal. 2017; **384**:135-144. DOI: 10.1016/j.ins.2016.06.026

[69] Liu Y, Gong W, Zhao Y, Deng X, Zhang S, Li C. PRBind: Protein-RNA interface prediction by combining sequence and I-TASSER model based structural features learned with convolutional neural networks. Bioinformatics. 2021;**37**(7):937-942. DOI: 10.1093/bioinformatics/btaa747

[70] Liu B, Li C, Yan K. DeepSVM-fold: Protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. Briefings in Bioinformatics. 2020;**21**(5):1733-1741. DOI: 10.1093/bib/bbz098

[71] Liu B, Chen J, Guo M, Wang X. Protein remote homology detection and fold recognition based on sequence-order frequency matrix. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019;**16**(1):292-300. DOI: 10.1109/TCBB.2017.2765331

[72] Lee TY, Chen SA, Hung HY, Ou YY. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. PLoS One. 2011;**6**(3):e17331. DOI: 10.1371/journal.pone.0017331

[73] Hanson J, Peliwal K, Litfin T, Yang YD, Zhou YQ. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics. 2018;**34**(23):4039-4045. DOI: 10.1093/bioinformatics/bty481

[74] Feng S, Xia C, Shen H. CoCoPRED: Coiled-coil protein structural feature prediction from amino acid sequence using deep neural networks. Bioinformatics. 2022;**38**(3):720-729. DOI: 10.1093/bioinformatics/btab744

[75] Mohammad SR, Mir A, Nasiri JA. A novel fusion based on the evolutionary features for protein fold recognition using support vector machines. Scientific Reports. 2020;**10**(1):14368. DOI: 10.1038/s41598-020-71172-x

[76] Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: Predicting TATA binding proteins with novel features and dimensionality reduction strategy. BMC Systems Biology. 2016;**10**(4):401-412. DOI: 10.1186/s12918-016-0353-5

[77] Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Research. 2006;**34**(2): W32-W37. DOI: 10.1093/nar/gkl305

[78] Yau SST, Yu C, He R. A protein map and its application. DNA and Cell Biology. 2008;**27**(5):241-250. DOI: 10.1089/dna.2007.0676

[79] Xia X, Li WH. What amino acid properties affect protein evolution? Journal of Molecular Evolution. 1998;**47** (5):557-564. DOI: 10.1007/PL00006412

[80] Zhang D, Xu Z, Su W, Yang YH, LvH YH, Lin H. iCarPS: A computational tool for identifying protein carbonylation sites by novel encoded features. Bioinformatics. 2021;**37**(2):

171-177. DOI: 10.1093/bioinformatics/btaa702

[81] You Z, Chan KCC, Hu P. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random Forest. PLoS One. 2015;**10**(5):e0125811. DOI: 10.1371/journal.pone.0125811

[82] Preeti J, Aruna T, Neha B, Milind R, Om PP, Nilagiri H, et al. Apache spark based scalable feature extraction approaches for protein sequence and their clustering performance analysis. International Journal of Data Science and Analytics. 2023;**15**(4):359-378. DOI: 10.1007/s41060-022-00381-6

[83] Robert PB. Prediction of protein structural features from sequence data based on Shannon entropy and Kolmogorov complexity. PLoS One. 2015;**10**(4):e0119306. DOI: 10.1371/journal.pone.0119306

[84] Hu S, Ma R, Wang H. An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences. PLoS One. 2019;**14**(11): e0225317. DOI: 10.1371/journal.pone.0225317

[85] Yang X, Jin J, Wang R, Li Z, Wang Y, Wei L. CACPP: A contrastive learning-based siamese network to identify anticancer peptides based on sequence only. Journal of Chemical Information and Modeling. 2024;**64**(7):2807-2816. DOI: 10.1021/acs.jcim.3c00297

[86] Wang C, Ju Y, Zou Q, Lin C. DeepAc4C: A convolutional neural network model with hybrid features composed of physicochemical patterns and distributed representation information for identification of N4-acetylcyti dine in mRNA. Bioinformatics. 2021;**38**(1):btab611. DOI: 10.1093/bioinformatics/btab611

[87] Lv Z, Cui F, Zou Q, Zhang L, Xu L. Anticancer peptides prediction with deep representation learning features. Briefings in Bioinformatics. 2021;**22**(5): 1-14. DOI: 10.1093/bib/bbab008

[88] Zhang W, Ding Y, Wei L, Guo X, Ni F. Therapeutic peptides identification via kernel risk sensitive loss-based k-nearest neighbor model and multi-Laplacian regularization. Briefings in Bioinformatics. 2024;**25**(6):bbae534. DOI: 10.1093/bib/bbae534

[89] Wang Y, Luo X, Zou Q. Effector-GAN: Prediction of fungal effector proteins based on pretrained deep representation learning methods and generative adversarial networks. Bioinformatics. 2022;**38**(14):3541-3548. DOI: 10.1093/bioinformatics/btac37

[90] Bepler T, Berger B. Learning protein sequence embeddings using information from structure. In: Seventh International Conference on Learning Representations (ICLR 2019); 6-9 May 2019; New Orleans, USA. Massachusetts: OpenReview; 2019. DOI: 10.48550/arXiv.1902.08661

[91] Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nature Methods. 2019;**16**(12): 1315-1322. DOI: 10.1038/s41592-019-0598-1

[92] Ben K, Liang L, Iain M, Steve R. Multiplicative LSTM for sequence modelling. Statistics. 2016;**2**:1-9. Available from: https://arxiv.org/abs/1609.07959

[93] Zeng M, Zhang F, Wu F, Li Y, Wang J, Li M. Protein-protein interaction site

prediction through combining local and global features with deep neural networks. Bioinformatics. 2020;**36**(4): 1114-1120. DOI: 10.1093/bioinformatics/btz699

[94] Wang R, Feng Y, Sun M, Jiang Y, Li Z, Cui L, et al. MVIL6: Accurate identification of IL-6-induced peptides using multi-view feature learning. International Journal of Biological Macromolecules. 2023;**246**: 125412. DOI: 10.1016/j.ijbiomac.2023.125412

[95] Jiao S, Ye X, Ao C, Sakurai T, Zou Q, Xu L. Adaptive learning embedding features to improve the predictive performance of SARS-CoV-2 phosphorylation sites. Bioinformatics. 2023;**39**(11):btad627. DOI: 10.1093/bioinformatics/btad627

[96] Lin K, May ACW, Taylor WR. Amino acid encoding schemes from protein structure alignments: Multi-dimensional vectors to describe residue types. Journal of Theoretical Biology. 2002;**216**(3):361-365. DOI: 10.1006/jtbi.2001.2512

[97] He W, Wang Y, Cui L, Su R, Wei LY. Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides. Bioinformatics. 2021;**37**(24):4684-4693. DOI: 10.1093/bioinformatics/btab560

[98] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019); 2–7 June 2019; Minneapolis, Minnesota, USA. Kerrville: ACL; 2019. DOI: 10.48550/arXiv.1810.04805

[99] Tung CH, Chien CH, Chen CW, Huang LY, Liu YN, Chu YW. QUATgo: Protein quaternary structural attributes predicted by two-stage machine learning approaches with heterogeneous feature encoding. PLoS One. 2020;**15**(4): e0232087. DOI: 10.1371/journal.pone.0232087

[100] Chen Z, Zhou Y, Song JN, Zhang ZD. hCKSAAP_UbSite: Improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochimica et Biophysica Acta. 2013;**1834**(8):1461-1467. DOI: 10.1016/j.bbapap.2013.04.006

[101] Du P, Gu S, Jiao Y. PseAAC-general: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. International Journal of Molecular Sciences. 2014;**15**(3):3495-3506. DOI: 10.3390/ijms15033495

[102] Du PF, Wang X, Xu C, Gao Y. PseAAC-builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Analytical Biochemistry: An International Journal of Analytical and Preparative Methods. 2012;**425**(2):117-119. DOI: 10.1016/j.ab.2012.03.015

[103] Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, et al. POSSUM: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics. 2017;**33**(17):2756-2758. DOI: 10.1093/bioinformatics/btx302

[104] Zhang P, Tao L, Zeng X, Qin C, Chen SY, Zhu F, et al. PROFEAT update: A protein features web server with added facility to compute network descriptors for studying omics-derived networks. Journal of Molecular Biology.

2017;**429**(3):416-425. DOI: 10.1016/j.jmb.2016.10.013

[105] Cao D, Xu Q, Liang Y. Propy: A tool to generate various modes of Chou's PseAAC. Bioinformatics. 2013;**29**(7):960-962. DOI: 10.1093/bioinformatics/btt072

[106] Xiao N, Cao DS, Zhu MF, Xu QS. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics. 2015;**31**(11):1857-1859. DOI: 10.1093/bioinformatics/btv042

[107] Cao DS, Xiao N, Xu QS, Chen AF. Rcpi:R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. Bioinformatics. 2015;**31**(2):279-281. DOI: 10.1093/bioinformatics/btu624

[108] Zuo YC, Li Y, Chen YL, Li GP, Yan ZH, Yang L. PseKRAAC: A flexible web server for generating pseudo K-tuple reduced amino acids composition. Bioinformatics. 2017;**33**(1):122-124. DOI: 10.1093/bioinformatics/btw564

[109] Wei L, Xing P, Su R, Shi G, Ma ZS, Zou Q. CPPred-RF: A sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. Journal of Proteome Research. 2017;**16**(5):2044-2053. DOI: 10.1021/acs.jproteome.7b00019

[110] Tian K, Zhao X, Yau ST. Convex hull analysis of evolutionary and phylogenetic relationships between biological groups. Journal of Theoretical Biology. 2018;**456**:34-40. DOI: 10.1016/j.jtbi.2018.07.035

[111] Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed. Beijing: China Machine Press; 2001

[112] Wan X, Tan X. A simple protein evolutionary classification method based on the mutual relation. Current Bioinformatics. 2020;**15**(10):1113-1129. DOI: 10.2174/1574893615666200305090055

[113] Lu AX, Lu AX, Pritišanac I, Taraneh Z, Forman-Kay JD, Moses AM. Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning. PLoS Computational Biology. 2022;**18**(6):e1010238. DOI: 10.1371/journal.pcbi.1010238

[114] Rao B, Zhou C, Zhang G, Su R, Wei L. ACPred-fuse: Fusing multi-view information improves the prediction of anticancer peptides. Briefings in Bioinformatics. 2020;**21**(5):1846-1855. DOI: 10.1093/bib/bbz088

[115] Wei L, Hu J, Li F, Song J, Su R, Zou Q. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Briefings in Bioinformatics. 2020;**21**(1):106-119. DOI: 10.1093/bib/bby107

Chapter 5

When Bioinformatics Meets Agent-Based Modeling: An Evolving Paradigm for Complex Biological Systems

Giulia Russo and Francesco Pappalardo

Abstract

Bioinformatics and agent-based modeling (ABM) represent a transformative integration for exploring and simulating complex biological systems. By combining computational models with diverse biological datasets, these approaches address intricate dynamic behaviors spanning molecular to population levels. This chapter delineates the foundational principles of bioinformatics and ABM, explores their integration strategies, and discusses the computational tools that facilitate this synergy. Case studies illustrate applications in immunotherapy optimization, immunotoxicant dynamics, and vaccine design, showcasing their relevance in advancing precision medicine and drug discovery. Key challenges, including data standardization, computational scalability, and model validation, are discussed alongside future directions. The chapter underscores the pivotal role of interdisciplinary collaborations and emerging technologies, such as artificial intelligence (AI) and quantum computing, in overcoming existing barriers and driving innovation in this field. Additionally, a special focus will be devoted to the evolving regulatory landscape that is starting to incorporate these innovative tools.

Keywords: bioinformatics, agent-based modeling, complex systems, computational biology, systems biology

1. Introduction

Bioinformatics has revolutionized the analysis of biological data, offering sophisticated tools to decode vast datasets from genomics, proteomics, and metabolomics [1, 2]. Concurrently, Agent-based modeling (ABM) has emerged as a powerful methodology to simulate individual entities and their interactions, elucidating emergent behaviors in complex systems [3, 4]. Integrating these fields enables innovative solutions for studying intricate biological phenomena [5].

Recent advancements in bioinformatics, including high-throughput sequencing and systems biology approaches, have enabled unprecedented exploration of molecular mechanisms underlying diseases [6, 7]. For instance, comprehensive protocols

for investigating potential vaccine targets in pathogens like SARS-CoV-2 have been developed [8], combining transcriptomics and immunoinformatics to streamline influenza vaccine design [9].

Moreover, incorporating ABM in these analyses has allowed researchers to dynamically simulate immune responses, as demonstrated in studies predicting allergic reactions to chemical sensitizers [10]. These integrated approaches not only enhance predictive power but also reduce experimental costs and ethical concerns [11].

The synergy of bioinformatics and ABM is further highlighted in the context of immune system simulations [12]. By leveraging computational tools, researchers can predict the efficacy of therapeutic interventions, as shown in the development of multi-epitope vaccines for complex diseases like influenza [13]. This paradigm shift underscores the transformative potential of combining molecular insights with agent-based methodologies to address challenges in precision medicine and systems biology [14].

2. Core concepts and framework

2.1 Bioinformatics: An overview

Bioinformatics applies computational methods to manage, analyze, and interpret biological data [15]. The field has expanded significantly with the advent of next-generation sequencing (NGS) technologies, enabling high-resolution analysis of genomic and transcriptomic datasets [6]. These data provide insights into gene expression, regulatory networks, and epigenetic modifications [16, 17], paving the way for advancements in systems biology and personalized medicine.

Bioinformatics tools such as Cytoscape [18], STRING [19], and Reactome [20] provide valuable resources for annotating and visualizing molecular interactions. Additionally, machine learning algorithms integrated into platforms like DeepCell [21] have enhanced predictive accuracy, facilitating seamless integration into ABM studies.

A notable application of bioinformatics is in the design and evaluation of vaccines [22]. For example, transcriptomic analyses have identified key immune response pathways, aiding the development of multi-epitope vaccines targeting pathogens like SARS-CoV-2 [9, 13]. These approaches have been complemented by immunoinformatics tools that predict antigenicity and population coverage, ensuring broader vaccine efficacy [23].

Additionally, bioinformatics plays a critical role in drug discovery [24]. Researchers can identify novel therapeutic targets and repurpose existing drugs by integrating gene expression data with pathway analysis [25]. Studies on differential gene expression during disease progression have uncovered biomarkers that inform diagnostic and prognostic strategies [26].

Another area of focus is immunotoxicology [27], where bioinformatics methods are used to assess the impact of environmental chemicals on immune systems [28]. Tools like the Universal Immune System Simulator (UISS) [29, 30] have been employed to model immunotoxicity induced by skin sensitizers [10, 31], demonstrating the capability of bioinformatics to bridge data-driven insights and mechanistic understanding [32].

These advancements underscore the versatility of bioinformatics in addressing complex biological questions [33], from understanding molecular mechanisms to optimizing therapeutic strategies [34–38].

2.2 Agent-based modeling in biomedicine

ABM is a computational approach used to simulate the actions and interactions of individual agents within a system to assess their effects on the system as a whole [3, 4, 39–41]. Agents can represent cells, molecules, or even entire organisms in biological contexts. By modeling interactions at the individual level, ABM can capture emergent phenomena that are difficult to predict using traditional modeling techniques [5].

ABM is particularly valuable in studying systems with heterogeneous components and stochastic behaviors [42, 43], such as immune responses [44, 45], disease spread [46], and cellular dynamics [47, 48]. For instance, ABM has been utilized to model tumor growth and its microenvironment [49], capturing the complex interplay between cancer cells, stromal cells, and the immune system [50]. This has provided insights into tumor evolution [51] and potential therapeutic strategies [52].

Another application of ABM is in epidemiology, where it has been used to simulate the spread of infectious diseases and evaluate intervention strategies [53, 54]. By incorporating spatial and temporal dynamics, ABM can predict the impact of vaccination campaigns or quarantine measures, offering a flexible tool for public health planning [55–57].

The STriTuVaD project is an excellent example of ABM's utility. It has been applied to simulate the immune response to tuberculosis vaccines [45, 58, 59]. These models incorporate individual variability, enabling predictions of vaccine efficacy across diverse populations. Additionally, ABM has proven instrumental in ecological studies, modeling the interactions within microbial communities to understand their responses to environmental changes [60, 61].

As computational power grows, ABM becomes increasingly sophisticated, incorporating more detailed biological data and complex interaction rules [62, 63]. This advancement allows for more accurate and predictive simulations, bridging the gap between theoretical models and real-world applications [64].

2.3 Integrating bioinformatics with ABM

Integrating bioinformatics and ABM enables researchers to bridge molecular-level insights and system-level behaviors [65]. By combining large-scale omics data with dynamic modeling [66], this approach provides a comprehensive framework for understanding biological complexity and predicting outcomes of therapeutic interventions [67–69].

One key advancement in this integration is using multiscale models that incorporate genomic, proteomic, and metabolic data into ABM frameworks [40, 70]. For example, a recent study demonstrated the potential of combining transcriptomic data with ABM to predict the efficacy of drug treatments in heterogeneous populations [42]. Such models account for individual variability, enabling the design of personalized medicine strategies.

Educational initiatives have also emerged to promote integrating bioinformatics and ABM approaches. These efforts aim to train researchers in combining data-driven insights with mechanistic modeling, ensuring the next generation of scientists can leverage these tools effectively [71].

Another significant development is the application of integrated bioinformatics-ABM frameworks to study immune system dynamics [72]. By simulating immune responses using agent-based approaches informed by bioinformatics data, researchers can predict the efficacy of vaccines and identify potential side effects [22].

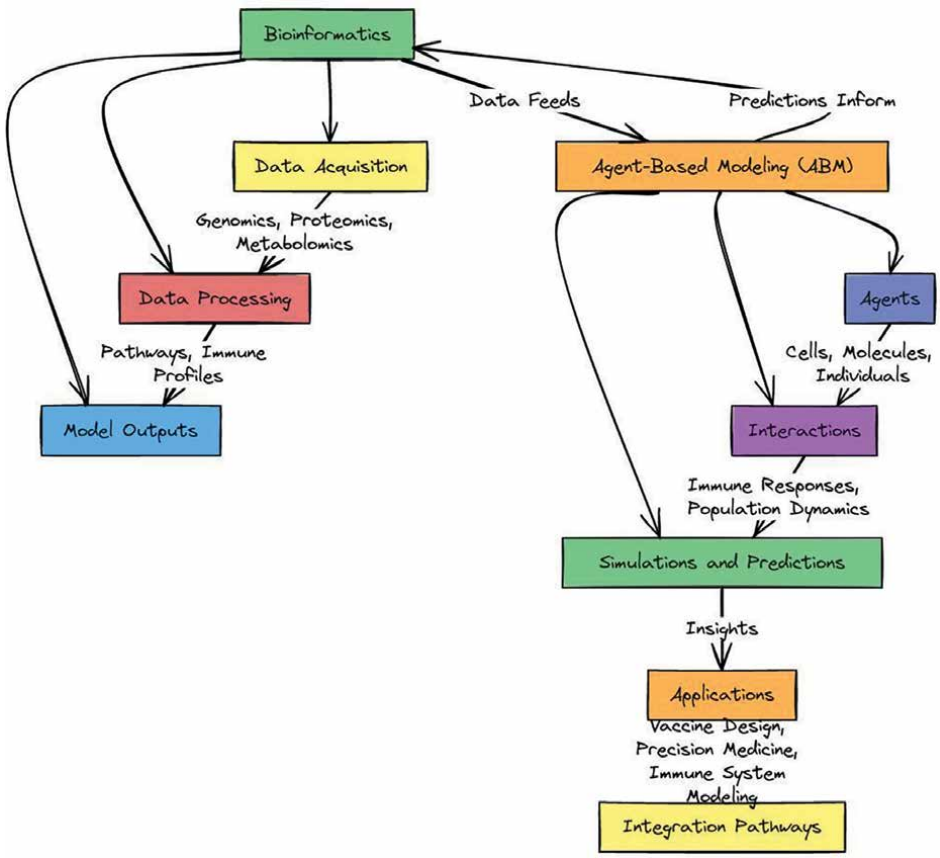

Figure 1.
Integration of bioinformatics and agent-based modeling (ABM), demonstrating the workflow from data acquisition to simulation outcomes for applications in precision medicine and vaccine design.

This methodology has been pivotal in accelerating the development of COVID-19 vaccines, demonstrating the real-world impact of this integration [73].

The combination of bioinformatics and ABM is also facilitating advancements in synthetic biology. For instance, models integrating gene regulatory networks and cellular dynamics have been used to design synthetic circuits with desired behaviors [74]. These innovations highlight the potential of integrated modeling approaches to drive fundamental research and applied sciences breakthroughs.

The integration of bioinformatics and ABM is visually summarized in the schematic diagram and flowchart provided below (**Figure 1**). These figures illustrate the data flow from acquisition through processing, leading to simulation and predictions, highlighting the dynamic interplay between data-driven bioinformatics and the mechanistic insights offered by agent-based modeling.

3. Case studies

The practical application of bioinformatics and ABM offers transformative insights into complex biological phenomena. This section highlights three case studies

demonstrating this integration's utility: simulating the immune response to influenza, exploring allergic reactions to chemical sensitizers, and developing SARS-CoV-2 vaccine targets. These examples emphasize the versatility of these approaches in addressing diverse challenges in systems biology and medicine.

3.1 Modeling influenza: Merging bioinformatics and ABM for precision insights

This case study explores how the Universal Immune System Simulator (UISS) was utilized to simulate immune responses to influenza, as detailed in Pappalardo et al. [68]. UISS, an advanced ABM platform, incorporated multiscale data to model the intricate dynamics of host-pathogen interactions during influenza infection.

Combining bioinformatics with ABM protocols, the simulations provided insights into the effectiveness of different vaccine strategies by modeling antigen presentation, T-cell activation, and antibody production. For instance, the study demonstrated how adjuvants influenced the strength and duration of the immune response. Additionally, the model predicted population-level impacts of vaccination under various scenarios, helping to optimize immunization strategies [9, 30]. These findings highlight the synergistic role of bioinformatics and ABM in guiding influenza vaccine development and public health decision-making.

3.2 Simulating allergy pathways: Fusion of data-driven and agent-based models

A novel study [10] utilized ABM and bioinformatics to investigate skin and respiratory allergic reactions to chemical sensitizers. By integrating transcriptomic and proteomic data, the researchers modeled the activation of immune cells in response to chemical exposure, focusing on dendritic cell maturation and T-cell proliferation.

Using bioinformatics and ABM protocols, the simulations explored how different chemical structures triggered varying levels of immune sensitization, leading to allergic reactions [75]. The study also assessed the efficacy of potential therapeutic interventions, such as cytokine inhibitors, in mitigating these responses. By predicting the allergenic potential of new chemicals, this approach provides a valuable tool for regulatory agencies and industries aiming to develop safer consumer products. This case study underscores the critical role of integrated bioinformatics and ABM in advancing personalized approaches to managing and preventing allergic diseases [31, 76].

3.3 Targeting SARS-CoV-2 with combined bioinformatics and ABM approaches

The final case study examines how integrated approaches have streamlined vaccine development processes specifically targeting SARS-CoV-2. Researchers have accelerated the timeline for vaccine candidates by utilizing bioinformatics to identify immunodominant epitopes and ABM to predict population-level efficacy [8].

The combined protocols of bioinformatics and ABM proved invaluable in simulating herd immunity thresholds and optimizing adjuvant formulations. These methodologies demonstrated the potential to reduce dependency on animal testing, align with ethical considerations, and lower development costs. Moreover, integrating large-scale genomic and immunological data allowed for the design of vaccines that target specific viral vulnerabilities, advancing the precision of SARS-CoV-2 vaccine development [77, 78].

4. The regulatory aspect

The regulatory landscape surrounding the integration of bioinformatics and agent-based modeling (ABM) into biological research and clinical applications is evolving, driven by the increasing recognition of these tools as essential components of modern biomedical innovation. Regulatory agencies, such as the European Medicines Agency (EMA) and the U.S. Food and Drug Administration (FDA), are progressively embracing computational models for applications ranging from drug development to personalized medicine [79–81]. This shift has been supported by frameworks like the FDA's "Model-Informed Drug Development (MIDD)" initiative [82] and the EMA's efforts to incorporate in silico methods into the approval processes for new therapies [83]. However, these advancements bring unique challenges, particularly for ABMs, which differ from traditional modeling approaches due to their stochastic nature and the high granularity of individual-agent interactions [84].

To achieve regulatory acceptance, ABMs must undergo rigorous validation and standardization [85]. This involves demonstrating their reproducibility and alignment with experimental or clinical data and ensuring transparency in their assumptions, parameterization, and decision-making processes [86]. Recent state-of-the-art advancements have proposed Good Simulation Practices (GSP) as a pathway to guide the validation of computational models [87], including ABMs, in a manner that meets regulatory expectations [88]. These practices prioritize traceability and robustness and employ benchmark datasets to authenticate predictive outputs [89]. For bioinformatics data, integration into ABM frameworks raises additional concerns, particularly around data provenance, quality, and compliance with privacy regulations such as the General Data Protection Regulation (GDPR) [90]. Ensuring the ethical use of patient-derived data and the ability to audit and replicate simulations are increasingly critical for regulatory approval [91].

Emerging technologies are also reshaping the regulatory landscape. For instance, artificial intelligence (AI) and machine learning are leveraged to enhance model validation processes, optimize parameter fitting, and assess predictive accuracy [92]. Recent advances in explainable AI (XAI) offer tools to address the "black-box" nature of complex ABMs, making their predictions and underlying mechanisms more interpretable and aligned with regulatory requirements [93]. Furthermore, efforts to create harmonized regulatory guidelines for in silico trials, such as those spearheaded by the Virtual Physiological Human (VPH) Institute [94] and renowned research group, provide a foundation for systematically integrating bioinformatics and ABMs into decision-making frameworks for new therapies and interventions [58, 95, 96].

A critical frontier lies in establishing regulatory acceptance of multiscale ABMs incorporating bioinformatics data at molecular, cellular, and population levels. These models aim to bridge the gap between genomic insights and system-wide dynamics and are increasingly recognized for their potential to accelerate drug development and refine clinical trial designs. For instance, in silico trials that use integrated bioinformatics-ABM approaches to simulate patient populations can reduce the dependency on animal testing and lower development costs while maintaining ethical and scientific rigor [97].

The future of regulatory engagement with bioinformatics and ABM will depend heavily on fostering interdisciplinary collaboration among computational scientists, biologists, clinicians, and regulators. Transparent reporting standards, clear documentation of model assumptions and limitations, and continuous dialog with regulatory bodies will be essential to ensure these tools meet both scientific and regulatory

expectations. By aligning advancements in computational modeling with regulatory requirements, the bioinformatics-ABM paradigm is poised to become a cornerstone of precision medicine, synthetic biology, and other fields where complex biological systems demand innovative and ethical solutions.

5. Challenges and future directions

Integrating bioinformatics and agent-based modeling (ABM) offers transformative potential for studying complex biological systems, yet it confronts significant challenges that must be surmounted to harness its impact [98] fully. These challenges include data standardization, computational scalability, and model validation [99, 100].

One major obstacle is the lack of data standardization and interoperability [101]. Integrating heterogeneous datasets from diverse sources—spanning genomic, proteomic, and metabolic information—frequently encounters issues of inconsistent annotations and variable experimental conditions [102]. This variability complicates their incorporation into ABM frameworks and limits the reproducibility of results [103]. Efforts to harmonize omics data with ABM frameworks are imperative, necessitating universal standards in bioinformatics tools to streamline data preprocessing and enhance model accuracy [104].

Moreover, the computational demands of ABM, especially when modeling large-scale systems or incorporating multiscale data, are substantial. These simulations often require significant memory and processing power, potentially hindering scalability [105] and the broader adoption of these methods. Recent advancements in cloud-based platforms and parallel computing [106], such as leveraging high-performance computing clusters [107], have begun to address these challenges by enabling the simulation of large-scale agent interactions, as demonstrated in studies of immune response dynamics [108].

Robust validation methods are required to ensure the reliability of integrated bioinformatics-ABM models [109]. Transparent reporting of model assumptions, parameter values, and performance metrics ensures reproducibility and encourages broader adoption.

Further addressing these computational and analytical demands, advances in bioinformatics have been instrumental. Enhanced by advanced machine learning techniques and high-performance computing [110], these improvements enable the development of more robust and interpretable models. Collaborative platforms that foster interdisciplinary efforts are also crucial for future advancements [111]. As we look forward, interdisciplinary collaborations will be pivotal in overcoming existing challenges. Additionally, emerging technologies such as quantum computing and AI-driven optimization offer promising solutions to computational bottlenecks [1, 92, 112]. Expanding the scope of applications to include areas like synthetic biology [113] and ecosystem modeling [114] could also unlock new frontiers for integrating bioinformatics and ABM.

6. Conclusions

Integrating bioinformatics and ABM has revealed a robust framework for unraveling the complexities of biological systems, providing multiscale insights by merging data-driven methodologies with dynamic simulations. The chapter has highlighted

various applications, including immunotherapy optimization, host-microbiome interactions, and vaccine development, illustrating the practical impact of this integrated approach in medicine and biology. Despite its vast potential, the field faces significant challenges. Data standardization remains a major hurdle, complicating the integration of heterogeneous datasets and impacting reproducibility. The computational intensity of ABM also poses scalability challenges, although recent advances in cloud computing and high-performance computing have begun to mitigate these issues. Furthermore, ensuring the validity of models through robust validation protocols is crucial for their broader acceptance and application. This necessity aligns with the evolving regulatory perspective that increasingly acknowledges the importance of computational models in biomedical applications, as highlighted by initiatives like the FDA's Model-Informed Drug Development (MIDD). The future of bioinformatics and ABM integration looks promising, driven by advancements in machine learning, quantum computing, and AI optimization. These technologies are poised further to enhance models' computational capabilities and interpretability. Continued efforts in fostering interdisciplinary collaborations will be vital in overcoming existing challenges and unlocking new potential. By leveraging these advancements, bioinformatics and ABM can significantly contribute to transformative research in medicine, ecology, and beyond, paving the way for innovative solutions to complex biological challenges while adhering to regulatory standards.

Conflict of interest

The authors declare no conflict of interest.

Notes/thanks/other declarations

None.

Author details

Giulia Russo and Francesco Pappalardo*
Department of Drug and Health Sciences, University of Catania, Catania, Italy

*Address all correspondence to: francesco.pappalardo@unict.it

References

[1] Jamialahmadi H, Khalili-Tanha G, Nazari E, Rezaei-Tavirani M. Artificial intelligence and bioinformatics: A journey from traditional techniques to smart approaches. Gastroenterology and Hepatology from Bed to Bench. 2024;**17**(3):241-252. DOI: 10.22037/ghfbb.v17i3.2977

[2] Clark AJ, Lillard JW. A comprehensive review of bioinformatics tools for genomic biomarker discovery driving precision oncology. Genes. 2024;**15**(8):1036

[3] Bonabeau E. Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences. 2002;**99**(Suppl. 3):7280-7287

[4] Cosgrove J, Butler J, Alden K, Read M, Kumar V, Cucurull-Sanchez L, et al. Agent-based modeling in systems pharmacology: Agent-based modeling in systems pharmacology. CPT: Pharmacometrics & Systems Pharmacology. 2015;**4**(11):615-629

[5] Merelli E, Armano G, Cannata N, Corradini F, d'Inverno M, Doms A, et al. Agents in bioinformatics, computational and systems biology. Briefings in Bioinformatics. 2006;**8**(1):45-59

[6] Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, et al. Next-generation sequencing technology: Current trends and advancements. Biology. 2023;**12**(7):997

[7] Vitorino R. Transforming clinical research: The power of high-throughput omics integration. Proteomes. 2024;**12**(3):25

[8] Russo G, Di Salvatore V, Sgroi G, Parasiliti Palumbo GA, Reche PA, Pappalardo F. A multi-step and multi-scale bioinformatic protocol to investigate potential SARS-CoV-2 vaccine targets. Briefings in Bioinformatics. 2022;**23**(1):bbab403

[9] Maleki A, Russo G, Parasiliti Palumbo GA, Pappalardo F. In silico design of recombinant multi-epitope vaccine against influenza A virus. BMC Bioinformatics. 2021;**22**:617. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124104006&doi=10.1186%2fs12859-022-04581-6&partnerID=40&md5=a0bb68053aa51ee6054a41ffe0614fdf

[10] Russo G, Crispino E, Casati S, Corsini E, Worth A, Pappalardo F. Pioneering bioinformatics with agent-based modelling: An innovative protocol to accurately forecast skin or respiratory allergic reactions to chemical sensitizers. Briefings in Bioinformatics. 2024;**25**(6):bbae506

[11] Tsou AY, Graf WD, Russell JA, Epstein LG, On behalf of the Ethics, Law, and Humanities Committee, a joint committee of the American Academy of Neurology (AAN), American Neurological Association (ANA), et al. Ethical perspectives on costly drugs and health care: AAN position statement. Neurology. 2021;**97**(14):685-692

[12] Pappalardo F, Russo G, Reche PA. Toward computational modelling on immune system function. BMC Bioinformatics. 2020;**21**(S17):546, s12859-020-03897-5

[13] Momajadi L, Khanahmad H, Mahnam K. Designing a multi-epitope influenza vaccine: An immunoinformatics approach. Scientific Reports. 2024;**14**(1):25382

[14] Ponnarengan H, Rajendran S, Khalkar V, Devarajan G, Kamaraj L. Data-driven healthcare: The role of computational methods in medical innovation. Computer Modeling in Engineering and Sciences. 2025;**142**(1):1-48

[15] Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002;**324**(7344):1018-1022

[16] Sonawane AR, DeMeo DL, Quackenbush J, Glass K. Constructing gene regulatory networks using epigenetic data. npj Systems Biology and Applications. 2021;**7**(1):45

[17] Fritz AJ, El Dika M, Toor RH, Rodriguez PD, Foley SJ, Ullah R, et al. Epigenetic-mediated regulation of gene expression for biological control and cancer: Cell and tissue structure, function, and phenotype. In: Kloc M, Kubiak JZ, editors. Nuclear, Chromosomal, and Genomic Architecture in Biology and Medicine, Results and Problems in Cell Differentiation. Vol. 70. Cham: Springer International Publishing; 2022. pp. 339-373. Available from: https://link.springer.com/10.1007/978-3-031-06573-6_12

[18] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research. 2003;**13**(11):2498-2504

[19] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research. 2021;**49**(D1):D605-D612

[20] Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: A database of reactions, pathways and biological processes. Nucleic Acids Research. 2011;**39**(Database):D691-D697

[21] Greenwald NF, Miller G, Moen E, Kong A, Kagel A, Dougherty T, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nature Biotechnology. 2022;**40**(4):555-565

[22] OladipoEK,AdeniyiMO,OgunlowoMT, Irewolede BA, Adekanola VO, Oluseyi GS, et al. Bioinformatics designing and molecular modelling of a universal mRNA vaccine for SARS-CoV-2 infection. Vaccine. 2022;**10**(12):2107

[23] Sarvmeili J, Baghban Kohnehrouz B, Gholizadeh A, Shanehbandi D, Ofoghi H. Immunoinformatics design of a structural proteins driven multi-epitope candidate vaccine against different SARS-CoV-2 variants based on fynomer. Scientific Reports. 2024;**14**(1):10297

[24] Somda D, Wilson Kpordze S, Jerpkorir M, Chantelle Mahora M, Wanjiru Ndungu J, Wambui Kamau S, et al. The role of bioinformatics in drug discovery: A comprehensive overview. In: Rudrapal M, editor. Pharmaceutical Science. IntechOpen; 2024. Available from: https://www.intechopen.com/chapters/88596

[25] Grimes T, Potter SS, Datta S. Integrating gene regulatory pathways into differential network analysis of gene expression data. Scientific Reports. 2019;**9**(1):5479

[26] Ajadee A, Mahmud S, Hossain MB, Ahmmed R, Ali MA, Reza MS, et al. Screening of differential gene expression patterns through survival analysis for diagnosis, prognosis and therapies of clear cell renal cell carcinoma. PLoS ONE. 2024;**19**(9):e0310843

[27] Germolec D, Luebke R, Rooney A, Shipkowski K, Vandebriel R, Van Loveren H. Immunotoxicology: A brief history, current status and strategies for future immunotoxicity assessment. Current Opinion in Toxicology. 2017;**5**:55-59

[28] Vos J, Van Loveren H, Wester P, Vethaak D. Toxic effects of environmental chemicals on the immune system. Trends in Pharmacological Sciences. 1989;**10**(7):289-292

[29] Pappalardo F, Pennisi M, Motta S. Universal immune system simulator framework (UISS). In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology. Niagara Falls New York: ACM; 2010. pp. 649-650. Available from: https://dl.acm.org/doi/10.1145/1854776.1854900

[30] Russo G, Crispino E, Maleki A, Di Salvatore V, Stanco F, Pappalardo F. Beyond the state of the art of reverse vaccinology: Predicting vaccine efficacy with the universal immune system simulator for influenza. BMC Bioinformatics. 2023;**24**(1):231

[31] Russo G, Crispino E, Corsini E, Iulini M, Paini A, Worth A, et al. Computational modelling and simulation for immunotoxicity prediction induced by skin sensitisers. Computational and Structural Biotechnology Journal. 2022;**20**:6172-6181

[32] Olson RS, La Cava W, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. arXiv. 2017. Available from: https://arxiv.org/abs/1708.05070

[33] Oulas A, Minadakis G, Zachariou M, Sokratous K, Bourdakou MM, Spyrou GM. Systems bioinformatics: Increasing precision of computational diagnostics and therapeutics through network-based approaches. Briefings in Bioinformatics. 2019;**20**(3):806-824

[34] Casotti MC, Meira DD, Alves LNR, Bessa BGDO, Campanharo CV, Vicente CR, et al. Translational bioinformatics applied to the study of complex diseases. Genes. 2023;**14**(2):419

[35] Van Camp PJ, Haslam DB, Porollo A. Bioinformatics approaches to the understanding of molecular mechanisms in antimicrobial resistance. International Journal of Molecular Sciences. 2020;**21**(4):1363

[36] Ramarajyam G, Murugan R, Rajendiran S. Network pharmacology and bioinformatics illuminates punicalagin's pharmacological mechanisms countering drug resistance in hepatocellular carcinoma. Human Genetics. 2024;**42**:201328

[37] Jiménez-Santos MJ, García-Martín S, Fustero-Torre C, Di Domenico T, Gómez-López G, Al-Shahrour F. Bioinformatics roadmap for therapy selection in cancer genomics. Molecular Oncology. 2022;**16**(21):3881-3908

[38] Ashik MAR, Hossain MA, Rahman SA, Akter MS, Zaman NN, Uddin MH, et al. Bioinformatics and system biology approaches for identifying potential therapeutic targets for prostate cancer. Informatics in Medicine Unlocked. 2024;**47**:101488

[39] Laubenbacher R, Hinkelmann F, Oremland M. Agent-based models and optimal control in biology: A discrete approach. In: Mathematical Concepts and Methods in Modern Biology. United States: Elsevier;

2013. pp. 143-178. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780124157804000053

[40] An G, Mi Q, Dutta-Moscato J, Vodovotz Y. Agent-based models in translational systems biology. WIREs Systems Biology and Medicine. 2009;**1**(2):159-171

[41] Cogno N, Axenie C, Bauer R, Vavourakis V. Agent-based modeling in cancer biomedicine: Applications and tools for calibration and validation. Cancer Biology & Therapy. 2024;**25**(1):2344600

[42] Yu JS, Bagheri N. Agent-based models predict emergent behavior of heterogeneous cell populations in dynamic microenvironments. Frontiers in Bioengineering and Biotechnology. 2020;**8**:249

[43] Bauer AL, Beauchemin CAA, Perelson AS. Agent-based modeling of host–pathogen systems: The successes and challenges. Information Sciences. 2009;**179**(10):1379-1389

[44] Xu Q, Ozturk MC, Cinar A. Agent-based modeling of immune response to study the effects of regulatory T cells in type 1 diabetes. Processes. 2018;**6**(9):141

[45] Russo G, Sgroi G, Parasiliti Palumbo GA, Pennisi M, Juarez MA, Cardona PJ, et al. Moving forward through the in silico modeling of tuberculosis: A further step with UISS-TB. BMC Bioinformatics. 2020;**21**(S17):458

[46] Kim Y, Cho N. A simulation study on spread of disease and control measures in closed population using ABM. Computation. 2022;**10**(1):2

[47] Dalmasso G, Marin Zapata PA, Brady NR, Hamacher-Brady A. Agent-based modeling of mitochondria links sub-cellular dynamics to cellular homeostasis and heterogeneity. PLoS ONE. 2017;**12**(1):e0168198

[48] Pleyer J, Fleck C. Agent-based models in cellular systems. Frontiers of Physics. 2023;**10**:968409

[49] Van Genderen MNG, Kneppers J, Zaalberg A, Bekers EM, Bergman AM, Zwart W, et al. Agent-based modeling of the prostate tumor microenvironment uncovers spatial tumor growth constraints and immunomodulatory properties. npj Systems Biology and Applications. 2024;**10**(1):20

[50] Gong C, Milberg O, Wang B, Vicini P, Narwal R, Roskos L, et al. A computational multiscale agent-based model for simulating spatio-temporal tumour immune response to PD1 and PDL1 inhibition. Journal of the Royal Society Interface. 2017;**14**(134):20170320

[51] Colyer B, Bak M, Basanta D, Noble R. A seven-step guide to spatial, agent-based modelling of tumour evolution. Evolutionary Applications. 2024;**17**(5):e13687

[52] Stephan S, Galland S, Labbani Narsis O, Shoji K, Vachenc S, Gerart S, et al. Agent-based approaches for biological modeling in oncology: A literature review. Artificial Intelligence in Medicine. 2024;**152**:102884

[53] Thomopoulos V, Tsichlas K. An agent-based model for disease epidemics in Greece. Information. 2024;**15**(3):150

[54] Bissett KR, Cadena J, Khan M, Kuhlman CJ. Agent-based computational epidemiological modeling. Journal of the Indian Institute of Science. 2021;**101**(3):303-327

[55] Sulis E, Terna P. An agent-based decision support for a vaccination

campaign. Journal of Medical Systems. 2021;**45**(11):97

[56] Wang P, Zheng X, Liu H. Simulation and forecasting models of COVID-19 taking into account spatio-temporal dynamic characteristics: A review. Frontiers in Public Health. 2022;**10**:1033432

[57] Cattaneo A, Vitali A, Mazzoleni M, Previdi F. An agent-based model to assess large-scale COVID-19 vaccination campaigns for the Italian territory: The case study of Lombardy region. Computer Methods and Programs in Biomedicine. 2022;**224**:107029

[58] Juárez MA, Pennisi M, Russo G, Kiagias D, Curreli C, Viceconti M, et al. Generation of digital patients for the simulation of tuberculosis with UISS-TB. BMC Bioinformatics. 2020;**21**(S17):449

[59] Pennisi M, Russo G, Sgroi G, Bonaccorso A, Parasiliti Palumbo GA, Fichera E, et al. Predicting the artificial immunity induced by RUTI® vaccine against tuberculosis using universal immune system simulator (UISS). BMC Bioinformatics. 2019;**20**(S6):504

[60] Bengtsson-Palme J. Microbial model communities: To understand complexity, harness the power of simplicity. Computational and Structural Biotechnology Journal. 2020;**18**:3987-4001

[61] Nagarajan K, Ni C, Lu T. Agent-based modeling of microbial communities. ACS Synthetic Biology. 2022;**11**(11):3564-3574

[62] Puniya BL, Verma M, Damiani C, Bakr S, Dräger A. Perspectives on computational modeling of biological systems and the significance of the SysMod community. Bioinformatics Advances. 2024;**4**(1):vbae090

[63] Sun Z, Lorscheid I, Millington JD, Lauf S, Magliocca NR, Groeneveld J, et al. Simple or complicated agent-based models? A complicated issue. Environmental Modelling and Software. 2016;**86**:56-67

[64] Jain HV, Norton KA, Prado BB, Jackson TL. SMoRe ParS: A novel methodology for bridging modeling modalities and experimental data applied to 3D vascular tumor growth. Frontiers in Molecular Biosciences. 2022;**9**:1056461

[65] Ji Z, Yan K, Li W, Hu H, Zhu X. Mathematical and computational modeling in complex biological systems. BioMed Research International. 2017;**2017**:1-16

[66] Sanches PHG, De Melo NC, Porcari AM, De Carvalho LM. Integrating molecular perspectives: Strategies for comprehensive multi-omics integrative data analysis and machine learning applications in transcriptomics, proteomics, and metabolomics. Biology. 2024;**13**(11):848

[67] Ponce-de-Leon M, Montagud A, Noël V, Meert A, Pradas G, Barillot E, et al. PhysiBoSS 2.0: A sustainable integration of stochastic Boolean and agent-based modelling frameworks. npj Systems Biology and Applications. 2023;**9**(1):54

[68] Pennisi M, Russo G, Ravalli S, Pappalardo F. Combining agent based-models and virtual screening techniques to predict the best citrus-derived vaccine adjuvants against human papilloma virus. BMC Bioinformatics. 2017;**18**:544. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85039714188&doi=10.1186%2fs12859-017-1961-9&partnerID=40&md5=79ec977bd7341043e31d717316526cbc

[69] Pennisi M, Russo G, Pappalardo F. Combining parallel genetic algorithms

and machine learning to improve the research of optimal vaccination protocols. In: Proceedings—26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP 2018. Cambridge, United Kingdom: IEEE Computer Society; 2018. pp. 399-405. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048768620&doi=10.1109%2fPDP2018.2018.00070&partnerID=40&md5=9c6406acd0accbdd0da014c8383d96e0

[70] Zhang L, Athale CA, Deisboeck TS. Development of a three-dimensional multiscale agent-based tumor model: Simulating gene-protein interaction profiles, cell phenotypes and multicellular patterns in brain cancer. Journal of Theoretical Biology. 2007;**244**(1):96-107

[71] Sivakumar N, Mura C, Peirce SM. Innovations in integrating machine learning and agent-based modeling of biomedical systems. Frontiers in Systems Biology. 2022;**2**:959665

[72] Jamali Y. Modeling the immune system through agent-based modeling: A mini-review. Immunoregulation. 2024;**6**(1):3-12

[73] Russo G, Di Salvatore V, Caraci F, Curreli C, Viceconti M, Pappalardo F. How can we accelerate COVID-19 vaccine discovery? Expert Opinion on Drug Discovery. 2021;**16**(10):1081-1084

[74] Rodrigo G, Carrera J, Jaramillo A. Computational design of synthetic regulatory networks from a genetic library to characterize the designability of dynamical behaviors. Nucleic Acids Research. 2011;**39**(20):e138

[75] Sutanto H. Mechanobiology of type 1 hypersensitivity: Elucidating the impacts of mechanical forces in allergic reactions. Mechanobiology in Medicine. 2024;**2**(1):100041

[76] Pappalardo F, Russo G, Corsini E, Paini A, Worth A. Translatability and transferability of in silico models: Context of use switching to predict the effects of environmental chemicals on the immune system. Computational and Structural Biotechnology Journal. 2022;**20**:1764-1777

[77] Ozsahin DU, Ameen ZS, Hassan AS, Mubarak AS. Enhancing explainable SARS-CoV-2 vaccine development leveraging bee colony optimised Bi-LSTM, Bi-GRU models and bioinformatic analysis. Scientific Reports. 2024;**14**(1):6737

[78] Russo G, Pennisi M, Fichera E, Motta S, Raciti G, Viceconti M, et al. In silico trial to test COVID-19 candidate vaccines: A case study with UISS platform. BMC Bioinformatics. 2020;**21**(S17):527

[79] Marques L, Costa B, Pereira M, Silva A, Santos J, Saldanha L, et al. Advancing precision medicine: A review of innovative in silico approaches for drug development, clinical pharmacology and personalized healthcare. Pharmaceutics. 2024;**16**(3):332

[80] Musuamba FT, Skottheim Rusten I, Lesage R, Russo G, Bursi R, Emili L, et al. Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility. CPT: Pharmacometrics & Systems Pharmacology. 2021;**10**(8):804-825

[81] Karanasiou G, Edelman E, Boissel FH, Byrne R, Emili L, Fawdry M, et al. Advancing in silico clinical trials for regulatory adoption and innovation. IEEE Journal of Biomedical and Health Informatics. 2024;**99**:1-15

[82] Madabushi R, Benjamin J, Zhu H, Zineh I. The US Food and Drug Administration's model-informed drug development meeting program: From pilot to pathway. Clinical Pharmacology and Therapeutics. 2024;**116**(2):278-281

[83] Giannuzzi V, Bertolani A, Torretta S, Reggiardo G, Toich E, Bonifazi D, et al. Innovative research methodologies in the EU regulatory framework: An analysis of EMA qualification procedures from a pediatric perspective. Frontiers in Medicine. 2024;**11**:1369547

[84] Curreli C, Pappalardo F, Russo G, Pennisi M, Kiagias D, Juarez M, et al. Verification of an agent-based disease model of human mycobacterium tuberculosis infection. International Journal of Numerical Methods in Biomedical Engineering. 2021;**37**(7):e3470. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105674679&doi=10.1002%2fcnm.3470&partnerID=40&md5=302faed2cac478c377211534fc61d1a0

[85] Viceconti M, Pappalardo F, Rodriguez B, Horner M, Bischoff J, Musuamba TF. In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. Methods. 2021;**185**:120-127

[86] Russo G, Parasiliti Palumbo GA, Pennisi M, Pappalardo F. Model verification tools: A computational framework for verification assessment of mechanistic agent-based models. BMC Bioinformatics. 2021;**22**:626. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130388346&doi=10.1186%2fs12859-022-04684-0&partnerID=40&md5=2a18b4cb393b67b89965e08585764ea7

[87] Viceconti M, Juárez M, Loewe A, Calvetti D, Somersalo E, Geris L, et al. Theoretical foundations of good simulation practice. In: Viceconti M, Emili L, editors. Toward Good Simulation Practice, Synthesis Lectures on Biomedical Engineering. Cham: Springer Nature Switzerland; 2024. pp. 9-23. Available from: https://link.springer.com/10.1007/978-3-031-48284-7_2

[88] Viceconti M, Serigado A, Rousseau CF, Voisin EM. Possible qualification pathways for in silico methodologies. In: Viceconti M, Emili L, editors. Toward Good Simulation Practice, Synthesis Lectures on Biomedical Engineering. Cham: Springer Nature Switzerland; 2024. pp. 67-72. Available from: https://link.springer.com/10.1007/978-3-031-48284-7_5

[89] Courcelles E, Horner M, Afshari P, Kulesza A, Curreli C, Vaghi C, et al. Model credibility. In: Viceconti M, Emili L, editors. Toward Good Simulation Practice, Synthesis Lectures on Biomedical Engineering. Cham: Springer Nature Switzerland; 2024. pp. 43-66. Available from: https://link.springer.com/10.1007/978-3-031-48284-7_4

[90] Alkhatib R, Gaede KI. Data management in biobanking: Strategies, challenges, and future directions. Biotech. 2024;**13**(3):34

[91] Khatiwada P, Yang B, Lin JC, Blobel B. Patient-generated health data (PGHD): Understanding, requirements, challenges, and existing techniques for data security and privacy. Journal of Personalized Medicine. 2024;**14**(3):282

[92] Huanbutta K, Burapapadh K, Kraisit P, Sriamornsak P, Ganokratanaa T, Suwanpitak K, et al. Artificial intelligence-driven pharmaceutical industry: A paradigm shift in drug discovery, formulation development,

manufacturing, quality control, and post-market surveillance. European Journal of Pharmaceutical Sciences. 2024;**203**:106938

[93] Sharma NA, Chand RR, Buksh Z, Ali ABMS, Hanif A, Beheshti A. Explainable AI frameworks: Navigating the present challenges and unveiling innovative applications. Algorithms. 2024;**17**(6):227

[94] Viceconti M, Clapworthy G, Jan SVS. The virtual physiological human—A European initiative for in silico human modelling. The Journal of Physiological Sciences. 2008;**58**(7):441-446

[95] Cassidy R, Singh NS, Schiratti PR, Semwanga A, Binyaruka P, Sachingongu N, et al. Mathematical modelling for health systems research: A systematic review of system dynamics and agent-based models. BMC Health Services Research. 2019;**19**(1):845

[96] Dotolo S, Marabotti A, Rachiglio AM, Esposito Abate R, Benedetto M, Ciardiello F, et al. A multiple network-based bioinformatics pipeline for the study of molecular mechanisms in oncological diseases for personalized medicine. Briefings in Bioinformatics. 2021;**22**(6):bbab180

[97] Pappalardo F, Russo G, Tshinanu FM, Viceconti M. In silico clinical trials: Concepts and early adoptions. Briefings in Bioinformatics. 2019;**20**(5):1699-1708

[98] Cilfone NA, Kirschner DE, Linderman JJ. Strategies for efficient numerical implementation of hybrid multi-scale agent-based models to describe biological systems. Cellular and Molecular Bioengineering. 2015;**8**(1):119-136

[99] Yang A, Troup M, Ho JWK. Scalability and validation of big data bioinformatics software. Computational and Structural Biotechnology Journal. 2017;**15**:379-386

[100] Vermeer WH, Smith JD, Wilensky U, Brown CH. High-Fidelity agent-based modeling to support prevention decision-making: An open science approach. Prevention Science. 2022;**23**(5):832-843

[101] Lopez Poncelas M, La Barbera L, Rawlinson JJ, Crandall D, Aubin CE. Credibility assessment of patient-specific biomechanical models to investigate proximal junctional failure in clinical cases with adult spine deformity using ASME V&V40 standard. Computer Methods in Biomechanics and Biomedical Engineering. 2022;**25**(5):543-553

[102] Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. Briefings in Bioinformatics. 2018;**19**(2):286-302

[103] Fitzpatrick BG. Issues in reproducible simulation research. Bulletin of Mathematical Biology. 2019;**81**(1):1-6

[104] Sen P, Orešič M. Integrating omics data in genome-scale metabolic modeling: A methodological perspective for precision medicine. Metabolites. 2023;**13**(7):855

[105] Lorig F, Dammenhayn N, Müller DJ, Timm IJ. Measuring and comparing scalability of agent-based simulation frameworks. In: Müller JP, Ketter W, Kaminka G, Wagner G, Bulling N, editors. Multiagent System Technologies. Cham: Springer International Publishing; 2015. pp. 42-60. Available from: http://link.springer.com/10.1007/978-3-319-27343-3_3

[106] Grelck C, Niewiadomska-Szynkiewicz E, Aldinucci M, Bracciali A, Larsson E. Why high-performance modelling and simulation for big data applications matters. In: Kołodziej J, González-Vélez H, editors. High-Performance Modelling and Simulation for Big Data Applications, Lecture Notes in Computer Science. Vol. 11400. Cham: Springer International Publishing; 2019. pp. 1-35. Available from: http://link.springer.com/10.1007/978-3-030-16272-6_1

[107] Chou J, Chung WC. Cloud computing and high performance computing (HPC) advances for next generation internet. Future Internet. 2024;**16**(12):465

[108] Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, Fonner JM, et al. VDJServer: A cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. Frontiers in Immunology. 2018;**9**:976

[109] Walpole J, Papin JA, Peirce SM. Multiscale computational models of complex biological systems. Annual Review of Biomedical Engineering. 2013;**15**(1):137-154

[110] Sovis A, Patikirige C, Pandigama Y. Enhanced timetable scheduling: A high-performance computational approach. In: 2023 8th International Conference on Information Technology Research (ICITR). Colombo, Sri Lanka: IEEE; 2023. pp. 1-6. Available from: https://ieeexplore.ieee.org/document/10382749/

[111] Patel AU, Gu Q, Esper R, Maeser D, Maeser N. The crucial role of interdisciplinary conferences in advancing explainable AI in healthcare. BioMedInformatics. 2024;**4**(2):1363-1383

[112] Durant TJS, Knight E, Nelson B, Dudgeon S, Lee SJ, Walliman D, et al. A primer for quantum computing and its applications to healthcare and biomedical research. Journal of the American Medical Informatics Association. 2024;**31**(8):1774-1784

[113] Matuszyńska A, Ebenhöh O, Zurbriggen MD, Ducat DC, Axmann IM. A new era of synthetic biology—Microbial community design. Synthetic Biology. 2024;**9**(1):ysae011

[114] Petrovskii S, Petrovskaya N. Computational ecology as an emerging science. Interface Focus. 2012;**2**(2):241-254

Edited by Ibrokhim Y. Abdurakhmonov

Biological data plays a crucial role in deciphering novel mechanisms of cellular and molecular interactions, enhancing our understanding of life. Therefore, bioinformatics has emerged as a vital field in the life sciences. This new volume of *Bioinformatics - Recent Advances* presents the views and opinions of the global bioinformatics research community on recent advancements in biological data analysis, methodologies, algorithms, software tools, and applications in the era of artificial intelligence and technological advances. This book will serve as a valuable resource for all life science researchers and students, providing an understanding of current trends and methodologies in bioinformatics.

Robert Koprowski, Biomedical Engineering Series Editor

ISSN 2631-5343
ISBN 978-1-83634-627-2
9 781836 346272

Published in London, UK

IntechOpen