Bioinformatics: Advancing biomedical discovery and innovation in the era of big data and artificial intelligence

Yuan Liu, Yamei Chen, and Leng Han

1Center for Epigenetics and Disease Prevention, Institute of Biosciences and Technology, Texas A&M University, Houston, TX 77030, USA
2Department of Translational Medical Sciences, College of Medicine, Texas A&M University, Houston, TX 77030, USA
3Department of Biostatistics and Data Science, School of Medicine, Indiana University, Indianapolis, IN 46202, USA
4Brown Center for Immunotherapy, School of Medicine, Indiana University, Indianapolis, IN 46202, USA

*Correspondence: yli218@iu.edu (Y.L.); lenghan@iu.edu (L.H.)

Received: April 12, 2023; Accepted: May 16, 2023; Published Online: May 28, 2023; https://doi.org/10.5971/j.xinn-med.2023.100012

(C) 2023 The Author(s). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Citation: Liu Y., Chen Y., and Han L. (2023). Bioinformatics: Advancing biomedical discovery and innovation in the era of big data and artificial intelligence. The Innovation Medicine 1(1), 100012.

Bioinformatics made significant progress in generating, analyzing, and interpreting vast amounts of biological data in the past decades. Challenged by the vast amount of data collected from diverse sources, bioinformatics research powered by artificial intelligence has led to novel insights into the field of biomedicine and will continue to drive further discoveries.

Bioinformatics has emerged as a driving force in the realm of biology, propelling research and innovation. Recently, Genomics, Proteomics & Bioinformatics announced the top 10 national breakthroughs in bioinformatics in 2022. These breakthroughs spanned across various sub-fields such as databases, algorithms, and applications. The field of bioinformatics has rapidly evolved over the past few decades, with the introduction of largescale high-throughput multi-omics technologies, such as genomics, proteomics, and metabolomics. With the advent of these technologies, the amount of biomedical data being generated has exploded, creating a need for advanced computational tools to manage, analyze, and interpret this data. The amount of data generated is so vast that traditional methods of analysis may be no longer sufficient in the era of “big data” in biomedicine. To address this challenge, the field of bioinformatics has increasingly turned to artificial intelligence (AI) to develop novel computational tools for analyzing large-scale omics data. In this perspective, we will briefly review the current progress in 2022 and future outlook of bioinformatics in the era of big data and AI.

LEVERAGING AND INTEGRATING LARGE-SCALE MULTI-DIMENSIONAL OMICS DATA

With advancements in technology, single-cell omics and spatial omics offer several advantages over traditional bulk omics approaches in providing more comprehensive insights into complex biological systems (Figure 1). Single-cell omics enable investigations of gene expression, epigenetic modifications, and other molecular characteristics of individual cells, enabling the identification of rare or previously unknown cell populations and the characterization of cellular states and dynamics. Meanwhile, spatial omics provides molecular information with the spatially resolved tissue contexture, fine tissue dissection, and visualization of intercellular crosstalk, which can be important for understanding the development and progression of diseases. The development of high-throughput technologies for single-cell and spatial omics has enabled the generation of large-scale datasets that can capture cellular heterogeneity and spatial context with high resolution.

Microbiota residing in the many parts of the body, such as the skin and gastrointestinal system, influence numerous physiologic processes, such as digestion, metabolism, cognitive development and function, as well as immune system development and function. Microbiota may also contribute to the development and treatment of diseases, including cancer. For instance, a high response rate of patients receiving immunotherapy was identified with higher microbial diversity and relatively high abundances of Ruminococcaceae family taxa, Akkermansia muciniphila, and methanogenic archaea. The increasing recognition of the significance of microbiota in human health and disease is driving the growing interest and research in this field, and large-scale data profiling from microbiome studies has emerged as a promising approach to gain insights into the complex microbial communities that reside in various environments. The profiling of microbiome is becoming increasingly utilized as biomarkers of response and therapeutic targets in precision medicine (Figure 1).

The large-scale omics profiling generates a tremendous amount of biomedical data and presents a big data challenge. Tackling this challenge requires developing efficient algorithms and techniques to extract meaningful information from complex datasets, while also addressing issues related to data storage, management, quality, scalability, and integration.

Model-based integration of omics data is crucial for advancing our understanding of cell biology, as it enables us to investigate intermolecular dynamics between gene regulation at the epigenome level and gene expression at the transcriptome and/or proteome levels within individual cells (Figure 1). For example, a model known as Graph-Linked Unified Embedding (GLUE) was proposed to integrate multi-omics single-cell data to infer regulatory relationships. Various data integration strategies have been proposed, including horizontal, vertical, and diagonal approaches. Despite extensive research, several challenges remain for computational data integration. These strategies make implicit assumptions about the expected similarity of cellular states captured across different studies and/or omics, which can lead to overcorrection of true biological variation. Additionally, integration methods need to be developed that can not only identify the common anchors between the different samples and/or molecular layers but can also account for sample-specific and omics-specific variation in a readily interpretable way. The expansion of these models to integrate paired, unpaired, and multi-omics single-cell or spatial assays will require robust data standards and well-established benchmarks to evaluate their efficacy.

Data integration is not limited to omics data fusion but also can be extended to broader modalities (Figure 1). AI-driven integration of multi-modality data in healthcare has led to significant progress in areas such as digital clinical trials, remote monitoring, and healthcare tools, offering a more comprehensive view of individual patients and enabling precision treatment plans. For example, the system known as Free-Text Reports for Supervision (REFERENCES) can process existing radiographs and free-text radiology reports with limited supervision, offering reliable abnormality detection in new X-ray images. Digital clinical trials use wearable devices and smartphone-enabled questionnaires to optimize measurements and increase participant engagement. Remote monitoring or ‘hospital-at-home’ uses wearable and ambient wireless sensors to provide valuable data for continuous patient care. Effective infectious disease surveillance integrates diverse data sources to predict and contain outbreaks. Virtual health assistants, powered by AI, provide personalized health coaching based on individualized multiple-type profiles. To fully harness the potential of multi-modality data integration, bioinformaticians must collaborate with health systems, research groups, and industry partners, addressing technical, data-related, and privacy challenges.

ENHANCING THE BIOMEDICAL UNDERSTANDING BY ARTIFICIAL INTELLIGENCE

AI has become an integral part of bioinformatics research, having deeply embedded in many aspects of the field (Figure 1). Among various AI algorithms, Deep learning (DL) has gained significant popularity in bioinformatics and is applied more frequently in various tasks compared to other traditional...
methods. DL has revolutionized the understanding and prediction of protein structures in recent years. Parallel to predicting protein structures in silico, protein design, was also powered by DL at a rapid pace, aiming to generate proteins with defined structures and desired functions. DL has been utilized to enhance sequence space exploration during the directed evolution of protein function and to develop proteins from scratch, considering constraints associated with binding interfaces. Recently, a data-driven approach leveraging evolutionary sequence data was utilized to produce active protein, such as protein message-passing neural network (Protein-MPNN). Protein design is a continually evolving field, and it holds the potential for the discovery of novel protein geometries that have not yet been observed in nature.

The impact of AI on drug discovery has been profound, as it has significantly accelerated the drug development process, lowered costs, and improved the success rate of drug development. By utilizing powerful algorithms to analyze large and complex datasets, AI has enabled researchers to identify potential drug targets, predict the properties of drug candidates, and screen vast libraries of compounds. Data-driven model-building processes can navigate vast datasets from high-throughput screening and prioritize alternatives and this represents a partial transfer of decision-making power to machine intelligence. Currently, there are over 150 small-molecule drugs in the discovery phase and more than 15 in clinical trials that have been developed by biotech companies utilizing an AI-first approach. This demonstrates that AI has the potential to lead to the autonomous generation of new chemical entities with desired properties, while minimizing the labor-intensive full-deck high-throughput screening.

Deep learning has shown remarkable success in visual tasks such as image recognition, object detection, and segmentation, achieving state-of-the-art performance in various benchmarks. This success has also extended to biomedical image analysis such as for cancer diagnosis, risk stratification, and treatment planning. Traditional machine learning models used hand-crafted features extracted from images, whereas DL models automatically learn features and outperform traditional models. For example, Optellum announced the world’s first FDA-cleared imaging biomarker for lung cancer prediction using DL [FDA 510(k) Number:K202300]. The ability of deep learning to learn from large image datasets and generalize to new data makes it a promising technology for continuing to advance the field of medical imaging.

Networks in biomedicine can be extremely large and complex, making traditional analytical methods difficult to apply. A branch of DL, specifically graph neural networks (GNN), has emerged as a powerful tool in network biology, which involves studying complex biological systems as networks of interacting components, such as genes, proteins, and metabolites. GNNs can learn representations of molecules, genes, and other biological entities as graphs, where nodes represent atoms or genes, and edges represent bonds or functional relationships. Applications of DL in network biology can extend to predicting protein-protein interactions, identifying disease-associated genes and pathways, and predicting drug-target interactions. For example, Li et al. proposed a GNN-based application, DeepFhni, that infers transcription factor regulatory networks from single-cell ATAC-seq data. These applications are necessary to gain a deeper understanding of biomedical systems and to develop more effective treatments for diseases.

AI models have been applied to a variety of tasks in genomic analysis, including gene expression analysis, DNA sequence classification, and variant calling. Computational models that process genomics data have helped to understand the impact of genetic variants, identify dysregulated genes in specific disease and tissue contexts, and interpret disease risk beyond what is feasible with experiments alone. For example, Lin et al. leveraged DL models to investigate complex structural variants in genome. With the continued growth of genomics data, AI models are poised to become an increasingly important tool in advancing our understanding of the genome and its role in health and disease.
CHALLENGE AND FUTURE DIRECTIONS

While bioinformatics has made significant progress, there are still key challenges that need to be addressed to effectively apply these advances in clinical settings. Firstly, the generalizability and robustness of discoveries and applications should be demonstrated through validation on multiple datasets, in various clinical settings, and prospectively, because the datasets used for training may have specific characteristics that affect the performance on external datasets. Thoughtful informed reporting is essential, providing details about the performance in different environmental and/or demographical settings.

Secondly, it is crucial and challenging to make constructed AI models interpretable for biomedicine researchers. Lack of interpretability may lead to errors or biases. Methods like saliency maps help understand what the model focuses on, while reporting intermediary model steps could mimic clinical workflows. These techniques can partially provide insights into model mechanisms, but it is advisable to propose more interpretation techniques to gain a more holistic understanding of the model’s behavior in biomedicine. Pairing AI models with interpretation mechanisms improves usability and performance, making interpretation vital for implementing AI in biomedicine.

Moreover, the remarkable performance exhibited by recently released large language models (LLMs), such as ChatGPT, has generated significant enthusiasm within the biomedical research community. LLMs are frequently utilized for tasks like generating and summarizing scientific content, as well as facilitating information retrieval through question-answering approaches. LLMs have the potential to greatly contribute to clinical decision support and patient interactions. Additionally, ongoing exploration of LLMs in various other areas holds the potential to significantly advance current biomedical research practices.

CONCLUSION

The progress made in bioinformatics in 2022 in terms of big data and AI has paved the way for more efficient and effective analysis of biomedical data, leading to novel discoveries and deep insights to address biomedical challenges. In conclusion, as the field continues to advance, we expect to see more breakthroughs driven by bioinformatics that will further enhance our understanding of biological systems and lead to more precise medical interventions.

REFERENCES


ACKNOWLEDGMENTS

The work of the authors was supported by the National Institutes of Health (NIH) grants R01HG011633 and R01CA262623 to L.H. We regret that page limitations have prevented us from including all relevant studies in this perspective. The figure is created with BioRender.com.

DECLARATION OF INTERESTS

The authors declare no competing interests.