New Challenges and Trends in Biomedical Big Data Development: Postprint

Authors: Zhang Guoqing, Li Yixue, Wang Zefeng, Zhao Guoping

Date: 2023-03-19T00:00:00+00:00

Abstract

Biomedical data has transitioned from the petabyte-scale omics era to the exabyte-scale multi-dimensional big data era, triggering a profound transformation of biomedical research toward the data-intensive fourth scientific paradigm. How to achieve high-dimensional and multi-level convergence and sharing of clinical data and research data, realize comprehensive management and utilization of biomedical big data spanning from "omics" to clinical and healthy population data, and thereby rapidly transform big data into new knowledge, has become the challenge confronting biomedical big data. Developing submission-based, integration-oriented data storage technologies, topic-based, interaction-oriented data sharing technologies, and traditional information technology-based, cutting-edge information technology-oriented data analysis and mining technologies, while simultaneously conducting research on standards and quality control, represents a novel approach for the storage, sharing, and translation of biomedical big data, and constitutes the technical cornerstone and future trend in building next-generation biomedical big data research centers.

Full Text

New Challenges and Trends in the Development of Bio-Medical Big Data

ZHANG Guoqing[12]*, LI Yixue[12]*, WANG Zefeng[1], ZHAO Guoping[1]

[1]Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2]Shanghai Center for Bioinformation Technology, Shanghai 201203, China

Abstract

Bio-medical data has transitioned from the petabyte-scale omics era to the exabyte-scale era of multi-dimensional big data, catalyzing a profound transformation in biomedical research toward a data-intensive "fourth paradigm" of scientific discovery. The critical challenge now lies in how to effectively aggregate and share clinical and research data across high dimensions and multiple levels, achieve comprehensive management and utilization of biomedical big data spanning from "omics" to clinical and population health data, and rapidly transform massive datasets into new knowledge. Addressing these challenges requires developing submission-based yet integration-oriented data storage technologies, subject-based yet interaction-oriented data sharing technologies, and traditional IT-based yet frontier IT-oriented data mining technologies, while simultaneously conducting research on standards and quality control. These approaches represent new strategies for the storage, sharing, and translation of biomedical big data and constitute the technical cornerstone and future trends for constructing next-generation biomedical big data research centers.

Keywords: bio-medical, big data, integration, interaction, data mining

Introduction

Since the launch of the Human Genome Project, the rapid advancement of various omics technologies—represented by next-generation sequencing and mass spectrometry—has driven exponential growth in massive life science omics data, including genomics, transcriptomics, epigenomics, proteomics, and metabolomics [1,2]. On one hand, machine learning and artificial intelligence technologies have significantly enhanced medical imaging and molecular imaging capabilities. On the other hand, population cohort studies and molecular epidemiology research have generated extensive longitudinal and spatial data, while phenomics has described high-dimensional data across multiple levels from molecules, cells, tissues, and organs to individuals. Real-world data has retrospectively aggregated vast amounts of clinical information [3,4], collectively forming complex high-dimensional biomedical big data. Breakthroughs in high-throughput experimental technologies have directly propelled biomedical data from the petabyte-scale genomics era into the exabyte-scale era of multi-omics integration.

We have entered an era of biomedical big data with considerable depth and breadth. Clinical biomedical data is characterized by enormous volume, rapid growth, difficult quality control, diverse and complex sources, and challenges in standardization and structuring. Research data in biomedicine exhibits wide variety, high-dimensional and complex internal structures, rich connotations, relative dispersion, and difficulties in high-dimensional, multi-level integration and sharing. Overall, biomedical data remains fragmented and distributed, making effective integration and analysis difficult and hindering the mining of

potential high-value insights. For China's biomedical research, the absence of data submission mechanisms has resulted in fragmented storage, decentralized management, and severe loss and degradation of data. The lack of security safeguards and international exchange platforms has compelled China to remain the world's largest exporter of omics data. The absence of sharing platforms, coupled with chaotic standardized management and uneven quality, has restricted open sharing due to both international and domestic policy and technical limitations.

Biomedical research is undergoing a profound transformation toward the data-intensive fourth scientific paradigm. The challenge we face is how to achieve the convergence, comprehensive management, utilization, and sharing of biomedical big data from "omics" to clinical and population health data, and to conduct deep mining and high-dimensional, all-round organic integration of multi-level clinical and research data to rapidly transform big data into new knowledge. This requires identifying the key elements for constructing next-generation biomedical big data storage, sharing, and translation centers (Figure 1 [Figure 1: see original paper]).

Three Technical Pillars for Next-Generation Platforms

1. Submission-Based, Integration-Oriented Data Storage As early as the 1980s–1990s, the United States, Europe, and Japan established the world's three major bioinformatics data centers: the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ). After nearly 30 years of development, these centers have formed complete data submission technical systems and exert significant influence in genomics, transcriptomics, proteomics, and other fields [5-11]. Domestic institutions have also begun constructing omics data centers by data type, including GSA [12], iPROX , and WDCM [13] for genomics, proteomics, and microbial resources, respectively. China's "1+5+X" plan for health and medical big data centers has been implemented, establishing a national data center and five regional centers in Jiangsu, Fujian, Shandong, Anhui, and Guizhou to accommodate health and medical big data for all citizens.

While these established and emerging life science and health data centers have greatly enriched data collection capabilities, the challenge lies in how to more effectively utilize data as scale increases. Traditional data models and organization methods can no longer satisfy management demands for massive data with rapidly growing volumes and constantly evolving structures, making dynamic adjustment according to actual conditions difficult. For existing or planned comprehensive biomedical big data platforms, it is necessary to break through the traditional model of building one database per data type and adopt a new warehouse-style data repository model. At the underlying data structure level, this approach should be integration-oriented, reserving connections between different data types based on sample, host, environment, temporal, and spatial information to form an elastic data structure that supports dynamic adjust-

ment, laying a solid foundation for subsequent data integration.

2. Subject-Based, Interaction-Oriented Data Sharing NCBI and EBI have aggregated vast data resources through submission services and provide data sharing via the web. As of July 2018, NCBI and EBI offered over 60 shared data resources including biological sequences, molecular structures, genetic information, and phenotype data [7], which have greatly facilitated life science and biomedical research. In addition to sharing third-party submitted data resources, alternative models have emerged, such as the database established by the U.S. National Cancer Institute's The Cancer Genome Atlas (TCGA) [14] and the UK's UK Biobank (UKB) , which provide tiered sharing based on data generated from large scientific projects to meet diverse research needs. Between these extremes, numerous small and medium-sized research teams have established a wide variety of databases and knowledge bases with varying scales and quality, offering data query, browsing, and download services, with some providing online analysis. *Nucleic Acids Research* publishes a database issue annually, which has become the most influential special issue in the biomedical database field, having published 1,737 database-related papers to date [15].

These databases—constructed by data type (e.g., genomics, transcriptomics, proteomics), species (e.g., human, non-human, vertebrates, invertebrates, microorganisms), or research purpose (e.g., genetic variation, transcription factors, regulatory networks)—have played enormous roles in promoting data sharing. However, as data types and scales expand, storing, organizing, and accessing different types of biomedical data across platforms has become a new challenge. Researchers have proposed the FAIR principles: Findable, Accessible, Interoperable, and Reusable [16]. Based on these principles, platforms such as BD2K [17] and OmicsDI [18] employ search engine technologies to overcome limitations of traditional subject-based databases, providing unified retrieval services for data resources from EBI, NCBI, and other data centers to achieve search engine-centered cross-database integration and better meet users' one-stop data sharing needs.

Beyond search technology, data visualization and online analysis are crucial means for data utilization. New visualization technologies, including HTML5 and JavaScript web display technologies, are increasingly applied in data platforms for macromolecule display, molecular imaging, and genome browsers [19-21]. Furthermore, databases have integrated online analysis tools such as sequence alignment, multiple sequence alignment, structural similarity comparison, and network structure analysis based on molecular sequences, structures, and interaction networks, greatly enhancing data interactivity.

3. Traditional IT-Based, Frontier IT-Oriented Data Mining From an analytical perspective, biomedical big data encompasses life science research data and clinical medical data. Supported by bioinformatics, computational biology, and systems biology, analysis methods for omics data—including ge-

nomics, transcriptomics, proteomics, and metabolomics—have matured, with analysis pipelines becoming increasingly commonplace and gradually evolving into traditional information technologies. Clinical medical data analysis has also widely adopted statistical modeling and machine learning technologies, with tools such as SAS, MATLAB, and R gaining extensive application.

However, data mining capabilities, particularly for omics data, are increasingly unable to meet the rapidly growing data output. The main challenges include: ever-increasing data volumes requiring faster data compression, transmission, and analysis methods [22,23]; and increasingly high data dimensions requiring more accurate dimensionality reduction methods [24]. Hardware technologies based on GPU (Graphics Processing Unit) and FPGA (Field-Programmable Gate Array) are being applied to optimize rate-limiting steps in traditional bioinformatics analysis methods, with increasing applications in sequence alignment and molecular docking [25,26]. Meanwhile, deep neural network-based artificial intelligence technologies have seen explosive growth in applications such as medical image processing and high-dimensional data dimensionality reduction, including auxiliary diagnosis of blinding retinal diseases, pneumonia, Alzheimer's disease, skin cancer, and meningioma from medical images [27-30]. Additionally, blockchain technology, due to its decentralized characteristics, is beginning to be applied in biomedical data sharing [31,32].

The application of frontier information technologies in biomedical big data will cover data preprocessing, transmission, analysis, and sharing, enhancing data mining capabilities.

Data Standards and Quality Control

Data standards for biomedical big data include terminology sets, data standards, and comprehensive standards. Typical terminology sets include Gene Ontology (GO) [33] and Human Phenotype Ontology (HPO) [34]. Minimum information standards for sequences include MIxS and MIGS [35-37], while medical data standards include ICD10 and SNOMED-CT . Life science data standards are mostly initiated by internationally influential institutions or societies, gradually gaining academic recognition alongside supporting data parsing or analysis software. For example, "The DDBJ/ENA/GenBank Feature Table Definition" [8] established by the International Nucleotide Sequence Database Collaboration (INSDC) is the earliest nucleic acid sequence data standard for NCBI, EBI, and other data centers, as well as the genome assembly data standard. MIAME [38] and GEO [39] for gene chip experiment data, MINSEQE for next-generation sequencing data defined by FGED, file formats such as BAM for alignment files, VCF for variation files, and GFF3 for genetic feature description, are all widely adopted. The most widely recognized standard in medical imaging is the DICOM standard .

Medical domain standards are far more complex than life science data standards, with higher degrees of normalization. Most medical standards undergo

formal stages including project initiation, drafting, and release, gaining broader recognition, such as the ISO/TC 215 series standards from the International Organization for Standardization's Technical Committee for Health Informatics , HL7 (Health Level Seven, the seventh layer in ISO's seven-layer information exchange protocol specification) , and the Clinical Data Interchange Standards Consortium (CDISC) . The scope of medical standards is also far more complex than life science data standards, covering vocabulary terminology, data description, technical operations, application services, and healthcare management.

Life science standards primarily focus on terminology sets and data standards, with relatively independent standards and less specification for data generation and analysis processes. Medical data standards emphasize interoperability and interconnectivity, forming self-contained systems, but with less description for research-supporting data standards. Therefore, biomedical big data urgently needs to strengthen the construction of data standard systems for clinical research and standards related to data analysis processes.

Data quality control is influenced by data generation and analysis, with different quality controls for different data types. For microarray and genomics data, the MAQC, MAQC-II, and MAQC-III projects led by the U.S. Food and Drug Administration (FDA) [40-44] have gained relatively broad recognition due to their independence from specific technology systems. For proteomics data quality control, lacking large projects comparable to MAQC, quality is primarily ensured through quality control tools from data submission platforms such as PRIDE and iPROX [45,46]. Data quality control requires reference datasets as benchmarks, including the agreement between raw data from experimental methods and reference datasets, as well as agreement between analysis results and reference datasets. Therefore, constructing reference datasets and reference data analysis pipelines for widely or critically used data types is a key component of data quality control and an important element in building biomedical big data platforms.

Practical Implementation and Considerations

We are constructing an open foundational platform represented by the Encyclopedia of Omics Data (NODE), which has already reached a considerable data scale. In terms of integrated storage, the platform includes domain-specific demonstration platforms such as the microbiome big data platform, as well as specialized databases such as the camel genome variation database and the translatable transcriptome RNA database. For interactive sharing, we are integrating conventional omics data analysis pipelines for whole genome, exome, and transcriptome data into the NODE system, along with domain-specific analysis pipelines for microbiome 16S RNA, metagenomics, and microbial functional annotation. Regarding frontier information technologies, we are optimizing high-resource-consumption steps such as transcriptome and metagenome assembly and mapping using GPU technology. For standards and quality control, we have implemented quality control for both descriptive information and raw

data, establishing automated quality control workflows that will enable automatic quality control assessment upon data submission.

Conclusion

In facing the challenges of biomedical big data, establishing a comprehensive technical and resource system to support the submission, management, sharing, and mining of life science research data and health medical big data is essential. Forming a data storage center that is submission-based yet integration-oriented, a data sharing center that is subject-based yet interaction-oriented, and a next-generation life science data translation center that is traditional IT-based yet frontier IT-oriented will effectively support basic research, applied research, and industrial demonstration in biomedicine and healthcare.

References

1 Bourne P E, Lorsch J R, Green E D. Perspective: Sustaining the big-data ecosystem. Nature, 2015, 527(7576): S16-17.

2 Perez-Riverol Y, Alpi E, Wang R et al. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics, 2015, 15(5-6): 930-949.

3 Argyropulo-Palmer M, Jenkins A, Theti D S, et al. Sunitinib in Metastatic Renal Cell Carcinoma: A Systematic Review of UK Real World Data. Front Oncol, 2015, 5: 195.

4 Berger ML, Lipset C, Gutteridge A, et al. Optimizing the leveraging of real-world data to improve the development and use of medicines. Value Health, 2015, 18(1): 127-130.

5 Benson D A, Cavanaugh M, Clark K, et al. GenBank. Nucleic Acids Res, 2018, 46(D1): D41-D47.

6 Cochrane G, Karsch-Mizrachi I, Takagi T, et al. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res, 2018, 46(D1): D48-D51.

7 The European Bioinformatics Institute in 2017: data coordination and integration. Nucleic Acids Res, 2018, 46(D1): D21-D29.

8 The DDBJ/ENA/GenBank Feature Table Definition. http://www.insdc.org/files/feature_{table}.html.

9 Kodama Y, Mashima J, Kosuge T et al. DNA Data Bank of Japan: 30th anniversary. Nucleic Acids Res, 2018, 46(D1): D30-D35.

10 Silvester N, Alako B, Amid C, et al. The European Nucleotide Archive in 2017. Nucleic Acids Res, 2018, 46(D1): D36-D40.

11 Vizcaino J A, Csordas A, Del-Toro N, et al. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res, 2016, 44(22): 11033.

12 Wang Y, Song F, Zhu J, et al. GSA: Genome Sequence Archive. Genomics, Proteomics Bioinformatics, 2017, 15(1): 14-18.

13 Wu L, Sun Q, Sugawara H, et al. World Data Centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. Nucleic Acids Res, 2017, 45(D1): D611-D618.

Author Biographies

ZHANG Guoqing is Deputy Director and Principal Investigator of the Bio-Med Big Data Center at the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences (CAS). His main research interests include bioinformatics databases and knowledge bases, with a long-term focus on integrating and mining omics data, literature data, and clinical data in precision medicine, large population cohorts, personalized drug development, microbiome, and synthetic biology. E-mail: gqzhang@picb.ac.cn

LI Yixue is Director and Principal Investigator of the Bio-Med Big Data Center at the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences (CAS). His main research interest is bioinformatics. E-mail: yxli@sibs.ac.cn

Note: Figure translations are in progress. See original paper for figures.