Life and Health Big Data: Current Status and Prospects (Postprint)

Authors: Bao Yiming, Xue Yongbiao

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

Abstract

Life and health big data constitute crucial foundational resources for national population health and biosafety. Currently, relevant data in China face severe loss, compromised sovereignty, inadequate security guarantees, and extremely low reuse efficiency, making it imperative to accelerate the construction of a national-level life and health big data sharing platform. Through developing novel methodologies for diversified and proactive data collection, new mechanisms for mutually beneficial data sharing, and advanced technologies for efficient and intelligent data analysis, we aim to establish a comprehensive system for the aggregation, management, sharing, and application of life and health big data, safeguard national data sovereignty, ensure data security, accelerate data utilization, serve research institutions, universities, hospitals, enterprises, and the general public, and make significant contributions to China' s economic and social development as well as the improvement of people' s well-being.

Full Text

Abstract

Life and health big data represents a critical foundational resource for national population health and biosafety. Currently, China faces severe challenges including data loss, loss of data sovereignty, inadequate security safeguards, and extremely low reuse efficiency. There is an urgent need to accelerate the construction of a national-level life and health big data sharing platform. By developing novel methods for multi-source proactive data collection, new mechanisms for mutually beneficial data sharing, and advanced technologies for efficient and intelligent data analysis, we must establish a comprehensive system for the aggregation, management, sharing, and application of life and health big data. Such a system will safeguard national data sovereignty, ensure data security, accelerate data utilization, and serve research institutions, universities, hospitals, enterprises, and the general public, thereby making significant contributions to

China' s economic and social development as well as improvements in people' s livelihoods.

Keywords: life and health, big data, current status, prospect

Challenges Facing Human Society and the Role of Life and Health Big Data

The Earth has entered the "Anthropocene" epoch, where human activities have profoundly impacted geological and ecological systems. Global warming-induced permafrost thawing threatens to revive extinct pathogens. The world population continues to grow and age, with China' s population aged 65 and older projected to reach 167 million by 2020, accounting for approximately one-quarter of the global total. Global agricultural productivity has fallen short of expectations for four consecutive years, and without improvement, will fail to meet the needs of the planet' s growing population. Local conflicts have increased refugee numbers, triggering serious social and public safety issues. Major chronic diseases pose severe threats to public health—statistics indicate that China has over 340 million patients with major chronic diseases, with an average of 8 people diagnosed with cancer and 5 dying from it every minute. Additionally, rising crime rates, terrorist attacks, and emergencies significantly threaten public safety.

Big data, particularly life and health big data, offers promising solutions to these societal challenges. Life and health big data refers to massive, complex datasets that cannot be processed within reasonable timeframes using conventional methods. Both basic life sciences research and health-related fields generate such data. In recent years, China has continuously increased its investment in life and health science and technology. The National Key R&D Program has launched major initiatives including "Precision Medicine Research," "Prevention and Control of Major Chronic Non-communicable Diseases," and "Reproductive Health and Major Birth Defect Prevention and Control," which are expected to generate over 300 PB of genomic data in the next five years. Internationally, multiple countries have launched genome sequencing projects at various scales, some reaching the million-person level. By 2025, global genomic data generation is estimated to reach 1 ZB annually. As health and medical technologies continue to advance, data output in the life and health sector is increasing exponentially. It is estimated that each hospital generates 665 TB of medical data annually, meaning that China' s more than 1,300 top-tier tertiary hospitals alone accumulate approximately 850 PB of data each year.

Rapid Growth of Life and Health Big Data

The development of health science increasingly relies on precision medicine big data. Modern medicine has entered the precision medicine era, built upon biological information big data, offering revolutionary opportunities for preventing and treating malignant tumors, cardiovascular and cerebrovascular diseases, and common diseases. Successful applications of big data are becoming increasingly

common: using whole-genome sequencing to guide type 2 diabetes treatment [?]; employing wearable devices to collect health big data [?]; utilizing deep learning and other artificial intelligence technologies to assist in skin cancer diagnosis [?]; conducting integrated multi-omics big data analysis for cancer precision classification and personalized treatment [?]; and inferring physical appearance phenotypes, ethnicity, geography, age, and lifestyle habits from DNA information [?]. These represent only a few examples among numerous successful big data applications.

Current Status of Life and Health Big Data

International Landscape

Foreign Genome Sequencing Projects Have Generated Massive Life and Health Data In 1977, Frederick Sanger's publication of the dideoxy chain termination method marked the maturation of sequencing technology. The Human Genome Project launched in 1986 and completed its draft in 2001. The emergence of the 454 sequencer in 2005 ushered in next-generation sequencing technology. Since then, large-scale sequencing projects in the life and health domain have proliferated. For example, the National Human Genome Research Institute (NHGRI) launched the ENCODE Project in September 2003 to identify and analyze all functional elements in the human genome. Complementing ENCODE, the NIH initiated the Roadmap Epigenomics Project in 2007 to create reference epigenome maps for different cell types. Almost concurrently, Europe's Wellcome Trust funded the 1000 Genomes Project [?], operated by the European Bioinformatics Institute (EMBL-EBI) from 2008-2015, with the primary goal of discovering genetic variants present at $ 1% frequency in studied human populations. Similarly, the Arabidopsis 1001 Genomes Project, launched in early 2008, aimed to identify sequence variations relative to the Arabidopsis reference genome across at least 1001 strains. The TCGA project [?], funded by the U.S. NHGRI and NIH, sequenced genomes, exomes, and transcriptomes of thousands of tumor samples to identify common gene mutations driving cancer development. The NIH-funded Human Microbiome Project (HMP) sequenced 16S rRNA amplicons of microorganisms inhabiting the human gut and skin to identify a core microbiome affecting human health. In 2012, the UK' s 100,000 Genomes Project was launched [?]. The larger U.S. government-funded Million Genomes Project, which had been in planning for three years, officially launched on May 20, 2018. This project will establish a health big data cohort of 1 million people, with an estimated cost of $1.5 billion over 10 years.

Established International Life and Health Data Center Infrastructure Developed countries have long prioritized the collection, analysis, and application of life and health big data. As early as November 1988, the U.S. National Library of Medicine (NLM) recognized the importance of "developing new information technologies to promote understanding of molecular processes controlling health and disease," and established the National Center for Biotechnology In-

formation (NCBI) by spinning off a project from the Lister Hill National Center for Biomedical Communications. From its inception, NCBI' s responsibilities included collecting biotechnology data worldwide. Over three decades, NCBI has grown from 20 to over 700 employees, with annual congressional funding increasing from $5.073 million in 1990 to a peak of $95.833 million in 2014. NCBI has accumulated the world' s largest collection of life and health databases (including GenBank, PubMed, SRA, dbGaP) and software resources (such as BLAST, e-Utilities), currently storing 30 PB of data. The website receives 4.2 million daily users, downloads over 60 TB of data daily, and handles more than 7,000 hits per second during peak times.

The European Bioinformatics Institute (EBI) traces its origins to the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database established in Heidelberg, Germany in 1980. In 1992, EMBL established EBI in Hinxton, UK. Initially, EBI maintained only two databases—the European Nucleotide Archive (ENA) and UniProt protein sequence resource—but has since built the world' s most comprehensive collection of molecular biology databases, managing 12 PB of data serving 3.2 million monthly users. EBI currently employs approximately 600 people with operating expenses of $88.2 million in 2016, primarily funded by EU governments, particularly the UK government.

In May 2005, NCBI, EBI, and DDBJ established the International Nucleotide Sequence Database Collaboration (INSDC). INSDC is one of the most prominent organizations for public domain data sharing internationally. Its members exchange data daily, hold annual meetings to discuss sequence archive establishment and maintenance, and have formulated unified standards and policies. INSDC wields enormous influence in international life and health data collection, with the convention that data must be uploaded to INSDC member databases before publication in mainstream biomedical journals.

The Swiss Institute of Bioinformatics (SIB), established in 1998, is a non-profit academic foundation that coordinates bioinformatics activities across Switzerland. SIB' s resources cover diverse life science domains, including genomics, proteomics, medical health, evolution, structural biology, and systems biology. In 2017, SIB' s core resources served approximately 6 million global users, with total managed funds reaching $26.765 million.

In the health big data sector, Epic is the largest electronic medical record vendor in the U.S., with about 190 million individual users storing their electronic medical information in Epic systems. Cerner is another major U.S. electronic medical record vendor, currently supporting 27,000 medical institutions of various sizes across 35 countries. DeepMind, a subsidiary of Google' s parent company Alphabet, is using artificial intelligence to analyze various medical images, attempting to learn the diagnostic experience that physicians acquire over years of training, thereby enabling machines to learn disease identification.

Domestic Landscape in China

Various Types of Life and Health Big Data Centers Have Been Established Representative domestic life and health big data centers include:

1. The Shenzhen National GeneBank, which focuses on self-produced data and serves as a node collecting data for EBI.
2. The Shanghai Biomedical Big Data Center, which concentrates on self-produced data from the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, supporting data submission, publication, management, and sharing.
3. The Microbial Resources and Big Data Center, which focuses on microbial resource libraries, providing microbial resource registration, query services, and microbial knowledge retrieval, with users throughout the international microbial research community.
4. The National Population and Health Science Data Sharing Service Platform, which provides access to approximately 400 medical databases, focusing on medical and health science data.
5. The National Public Security DNA Database [**?**], launched in 2004, which as of May 31, 2016, contained 44.358 million entries, including 40.719 million criminal information entries and 1.498 million crime scene evidence entries; the "Anti-Trafficking" DNA database has cumulatively entered 594,000 personnel data entries and 513,000 DNA data entries, bringing the combined total to 44.871 million entries [**?**].
6. The Beijing Institute of Genomics Life and Health Big Data Center [**?**], whose data primarily comes from user submissions and supports data submission, management, publication, sharing, retrieval, download, and online analysis. This database has over 300 data submission users from nearly 100 institutions, data access and download users from more than 70 countries and regions, is recognized by over 40 international journals, and was listed by the authoritative journal *Nucleic Acids Research* in 2018 as a global core data center alongside the U.S. NCBI and European EBI [**?**].

Existing Problems

(1) **Lack of a public platform for life and health big data management, leading to severe data loss.** Journals in the life and health field typically require authors to make published data publicly available in academically recognized databases. Due to the absence of unified national-level deployment and planning, China suffers serious data resource loss. Statistics show that in 2016, mainland China published 290,600 SCI papers with first authors from China, but the vast majority of data could only be submitted to internationally renowned databases such as NCBI and EBI. It is estimated that over 25% of the data in NCBI databases originates from China.

(2) **Lack of a data sharing mechanism, creating data silos and low utilization efficiency.** Over the past decade, China has produced numerous database resources through project funding support rather than national special funds. According to the latest statistics from the Database Commons database, China' s total number of database resources ranks second globally. However, most databases lack long-term maintenance and serious manual curation, with marginalized content. These factors result in low-quality database resources, low utilization rates, and ineffective data sharing. The absence of national-level framework design and deployment has led to small, scattered database resources in China, making it difficult to cultivate large-scale, high-quality data centers with international leadership. Similarly, based on Database Commons statistics, databases from China with over 500 citations are extremely rare, and none have exceeded 1,000 citations.

(3) **Lack of integration between life big data and health big data.** Life big data (especially omics data) and health big data are typically produced by units under different administrative departments. Due to departmental segmentation, interest relationships, and the lack of national top-level coordination and constraints, these two major data types often remain disconnected and cannot form synergies to achieve maximum effectiveness.

Prospects for Life and Health Big Data

Life and health big data constitutes a crucial foundational resource for national population health and biosafety. Currently, the main factors restricting China' s life and health big data research development include the lack of a national-level framework and technology, top-level design for resource reuse, coordination and management mechanisms, data sharing systems, and long-term stable funding support. These deficiencies have resulted in severe data loss, loss of data sovereignty, inadequate security, and extremely low reuse efficiency. Therefore, accelerating the construction of a national-level life and health big data center to establish a national biological big data centralized management and sharing service platform is imperative. Specifically, this requires building bioinformatics infrastructure with petaflop-scale computing capabilities and EB-level biological big data storage capacity, forming the ability to effectively manage China' s biological resources, population health, environmental, and agricultural big data, and support the effective management of national human genetic resources. It also necessitates establishing a world-class bioinformatics platform based on the interdisciplinary integration of information science, life science, computational science, and clinical medicine, driven by advanced technologies such as cloud computing and artificial intelligence, to create an international center for bioinformatics research and application development.

References

1. Stephens Z D, Lee S Y, Faghri F, et al. Big Data: Astronomical or Genomical? *PLoS Biology*, 2015, 13(7): e1002195.
2. Chen R, Mias G I, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 2012, 148(6): 1293-1307.
3. Gao W, Emaminejad S, Nyein H Y Y, et al. Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. *Nature*, 2016, 529(7587): 509-514.
4. Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017, 542(7639): 115-118.
5. Nebbioso A, Tambaro F P, Dell' Aversana C, et al. Cancer epigenetics: Moving forward. *PLoS Genet*, 2018, 14(6): e1007362.
6. Vogel G. German law allows use of DNA to predict suspects' looks. *Science*, 2018, 360(6391): 841-842.
7. Genomes Project Consortium, Abecasis G R, Altshuler D, et al. A map of human genome variation from population-scale sequencing. *Nature*, 2010, 467(7319): 1061-1073.
8. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 2008, 455(7216): 1061-1068.
9. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 2011, 474(7353): 609-615.
10. Cancer Genome Atlas Research Network, Weinstein J N, Collisson E A, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 2013, 45(10): 1113-1120.
11. Turnbull C, Scott R H, Thomas E, et al. The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. *British Medical Journal*, 2018, 361: k1687.
12. [Reference appears to be about National Public Security DNA Database - citation format unclear in original]
13. BIG Data Center Members. Database Resources of the BIG Data Center in 2018. *Nucleic Acids Research*, 2018, 46(D1): D14–D20.
14. Wang Y, Song F, Zhu J, et al. GSA: Genome Sequence Archive. *Genomics Proteomics & Bioinformatics*, 2017, 15(1): 14-18.
15. [Additional BIG Data Center references]

16. Rigden D J, Fernandez X M. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 2018, 46(D1): D1-D7.

Author Information

BAO Yiming is Director and Professor of the BIG Data Center (BIGD) at the Beijing Institute of Genomics, Chinese Academy of Sciences. His research interests include biological databases, virus genome annotation, and virus evolution and classification. He received his B.S. in biochemistry from Peking University in 1987 and his Ph.D. in genetics from the John Innes Centre (through the University of East Anglia), UK, in 1994. Dr. Bao currently serves as Deputy Director of the National Institutes of Data Science in Health and Medicine at the University of Chinese Academy of Sciences and is a member of the Computational Biology and Bioinformatics Specialized Committee of the Chinese Society of Biotechnology. E-mail: baoym@big.ac.cn

The author thanks Dr. Ma Yingke for editorial assistance.

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