Review Article Open Access

Big Data Analytics in Biology: A Systematic Review of Methods for Large-Scale Data Processing

Weipan Wang, Bing Zhang, Manman Li

Hainan Institute of Biotechnology, Haikou, 570206, Hainan, China

Corresponding author: manman li@hibio.org

Computational Molecular Biology, 2024, Vol.14, No.3 doi: 10.5376/cmb.2024.14.0012

Received: 29 Mar., 2024

Accepted: 22 May, 2024

Published: 02 Jun., 2024

Preferred citation for this article:

Wang W.P., Zhang B., and Li M.M., 2024, Big data analytics in biology: a systematic review of methods for large-scale data processing, Computational Molecular Biology, 14(3): 97-105 (doi: 10.5376/cmb.2024.14.0012)

Abstract This study explores various methods and tools developed for large-scale data processing in biological research. We studied comprehensive toolkits such as TBtools, which provide user-friendly interfaces for complex data analysis, as well as distributed computing frameworks such as MapReduce, which solve the problem of imbalance in large DNA datasets. In addition, we discussed the challenges posed by the heterogeneity and complexity of big biological data, emphasizing the need for powerful and scalable analytical frameworks, such as bigSCale for single-cell RNA sequencing, in order to gain a comprehensive understanding of the current status and future directions of big data analysis in the field of biology.

Keywords Big data analytics; Bioinformatics; High-throughput sequencing; Machine learning; Distributed computing

1 Introduction

The advent of high-throughput technologies has revolutionized the field of biology, ushering in the era of "big data." This transformation is characterized by the generation of vast amounts of data across various biological domains, including genomics, transcriptomics, proteomics, and metabolomics (Davis-Turak et al., 2017). The Human Genome Project, for instance, exemplifies the scale of data generation, having taken 13 years and over $3 billion to sequence the human genome, a task that can now be accomplished in a few days for a fraction of the cost (Li and Chen, 2014; Goh and Wong, 2020). The rapid accumulation of biological data has necessitated the development of sophisticated tools and techniques to manage, analyze, and interpret these large datasets (Greene et al., 2014; Chen et al., 2020).

The ability to process and analyze large-scale biological data is crucial for advancing our understanding of complex biological systems and translating this knowledge into practical applications. High-dimensional data spaces, such as those generated by genomic and proteomic technologies, present unique challenges in terms of data integration, analysis, and interpretation (Clarke et al., 2008). Effective data processing methods enable researchers to uncover hidden biological regularities, understand cellular processes, and develop predictive models for disease diagnosis and treatment (Ebrahim et al., 2016; Gutierrez et al., 2018). Moreover, the integration of multi-omic data provides a comprehensive view of biological systems, facilitating the discovery of novel insights that would be unattainable through single-omic approaches (Tariq et al., 2020; Juan and Huang, 2023).

This study provides a comprehensive overview of the methods and tools currently used for large-scale data processing in biology. By studying the challenges and opportunities related to big data in life sciences, we emphasize the advancements in data integration, quantitative analysis, and computing technologies that drive the field forward. In addition, this study will discuss the impact of these methods on future research and their potential applications in clinical and translational medicine, identify gaps in current methods, and propose directions for future research to improve the scalability and efficiency of biological big data analysis.

2 Overview of Big Data in Biological Research

2.1 Types of biological data

In the "Omics" era of life sciences, biological data is diverse and encompasses various levels of biological systems.

This includes genomic data, transcriptomic data, epigenomic data, proteomic data, metabolomic data, molecular imaging, molecular pathways, population data, and clinical/medical records (Li and Chen, 2014). The rapid development of high-throughput sequencing (HTS) techniques has significantly contributed to the generation of large-scale biological data, making it possible to profile biological systems in a cost-efficient manner (Greene et al., 2014). The data generated from these techniques are vast and complex, often requiring sophisticated tools and methodologies for effective analysis and interpretation.

2.2 Sources of big data in biology

The primary sources of big data in biology include next-generation sequencing (NGS) technologies, which have revolutionized the field by enabling the generation of massive datasets that can answer long-standing questions about human diseases and biological processes (Mardis, 2016). Additionally, observational networks and space-based data have facilitated the discovery of emergent mechanisms and phenomena on regional and global scales, further contributing to the pool of big biological data (Xia et al., 2020). The Human Genome Project is a notable example, which utilized extensive resources and collaboration to sequence the human genome, a task that can now be accomplished much more rapidly and cost-effectively due to advancements in sequencing technologies (Li and Chen, 2014).

2.3 Challenges in handling biological big data

Handling big biological data presents several challenges. One of the primary issues is the complexity and heterogeneity of the data, which requires integration from multiple autonomous sources (Wu et al., 2014). The volume, velocity, variety, and veracity of big data (the four V's) necessitate specialized theories and technologies for effective management and analysis (Li and Chen, 2014; Younas, 2019). Current data mining techniques often fall short in meeting the new space and time requirements posed by big data, highlighting the need for more robust and scalable solutions (Kamal et al., 2016). Moreover, the lack of standardized integration processes complicates the task of combining data from various sources into a unified format for analysis (Almasoud et al., 2020). The scientific community must also address issues related to data quality, security, and privacy to fully harness the potential of big data analytics in biological research (Wu et al., 2014; Chen et al., 2020).

3 Methods for Large-Scale Data Processing

3.1 Data storage and management

3.1.1 Distributed databases

Distributed databases play a crucial role in managing large-scale biological data. Technologies such as Apache Hadoop provide distributed and parallelized data processing capabilities, which are essential for handling petabyte-scale datasets in genomics and other biological fields (O'Driscoll et al., 2013). These systems enable efficient storage, retrieval, and processing of vast amounts of data by distributing the workload across multiple nodes, thus enhancing performance and scalability.

3.1.2 Cloud computing solutions

Cloud computing offers scalable and flexible solutions for storing and processing large biological datasets. Platforms like Sherlock leverage cloud technologies to provide a comprehensive data management system that supports data storage, conversion, querying, and sharing (Figure 1) (Bohár et al., 2022). Cloud-based solutions facilitate the handling of complex and large datasets by offering tools for distributed analytical queries and optimized storage formats, such as the Optimized Row Columnar (ORC) format, which enhances data processing efficiency.

3.1.3 Data security and privacy

As biological data often contain sensitive information, ensuring data security and privacy is paramount. The HACE theorem and associated data-driven models emphasize the importance of incorporating security and privacy considerations into big data processing frameworks (Wu et al., 2014). These models advocate for robust security measures to protect data integrity and confidentiality while enabling efficient data mining and analysis.

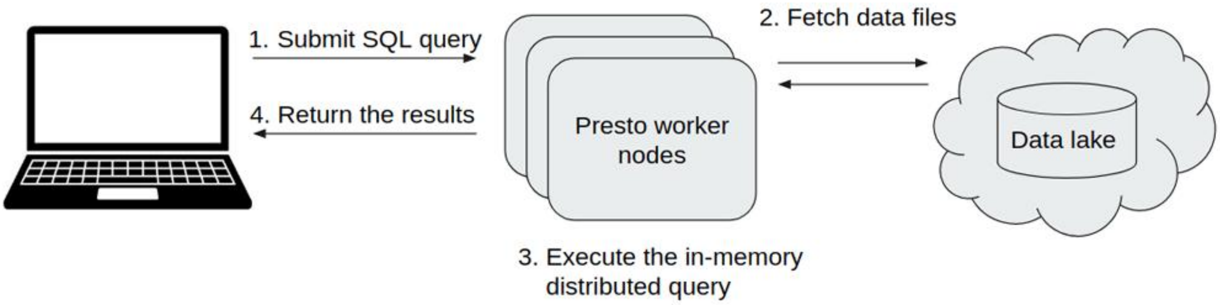

Figure 1 Overview of how the query engine and the Data Lake work together (Adopted from Bohár et al., 2022)

3.2 Data integration and interoperability

Integrating heterogeneous biological data from multiple sources is a significant challenge due to the diversity in data types and formats. Recent methods, such as non-negative matrix factorization-based approaches, have shown promise in effectively integrating various types of networked biological data, providing more holistic insights into biological systems (Gligorijević and Przulj, 2015). Additionally, frameworks that utilize domain ontology and distributed processing have been proposed to achieve seamless data integration, ensuring logical consistency and facilitating further research and analysis (Almasoud et al., 2020).

3.3 Data cleaning and preprocessing

Data cleaning and preprocessing are critical steps in preparing large-scale biological data for analysis. Tools like TBtools offer user-friendly interfaces and a wide range of functions for bulk sequence processing and interactive data visualization, making it easier for biologists to handle big data without extensive programming knowledge (Chen et al., 2020). Moreover, methodologies such as the MapReduce-based k-nearest neighbor (K-NN) classification approach have been developed to reduce data imbalance and enhance the efficiency of data classification and storage management (Kamal et al., 2016).

4 Analytical Techniques for Big Data

4.1 Machine learning algorithms

4.1.1 Supervised learning

Supervised learning algorithms are a cornerstone of big data analytics in biology, where labeled datasets are used to train models to make predictions or classify data. Common supervised learning techniques include linear regression, logistic regression, support vector machines (SVM), and random forests. These methods have been effectively applied to various biological datasets, such as protein-coding data for disease identification and treatment (Rahman, 2019). The use of supervised learning in bioinformatics allows for the development of predictive models that can provide insights into complex biological processes and disease mechanisms (Greene et al., 2014).

4.1.2 Unsupervised learning

Unsupervised learning algorithms are essential for analyzing large-scale biological data where labels are not available. Techniques such as clustering, principal component analysis (PCA), and hierarchical clustering help in identifying patterns and structures within the data. These methods are particularly useful in the initial stages of data exploration and for discovering hidden relationships in biological networks (Greene et al., 2014). Unsupervised learning has been applied to various biological datasets to uncover novel insights and generate hypotheses for further investigation (Jan et al., 2017).

4.1.3 Deep learning approaches

Deep learning, a subset of machine learning, has gained significant traction in the field of big data analytics due to its ability to handle large, complex, and heterogeneous datasets. Deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoders, have been successfully applied to biological data for tasks such as image classification, sequence analysis, and network prediction (Najafabadi et al., 2015; Tonidandel et al., 2018; Jin et al., 2020). These models can extract high-level features from raw data, enabling the discovery of intricate patterns and relationships that traditional methods might miss. Deep learning has shown promise in addressing challenges in big data analytics, including scalability, high-dimensional data, and the integration of diverse data types (Najafabadi et al., 2015; Shukla et al., 2021).

4.2 Statistical methods

Statistical methods play a crucial role in the preprocessing and analysis of big data in biology. Techniques such as data normalization, transformation, and noise reduction are essential for preparing data for further analysis. Methods like the Box-Cox transformation and linear transformation have been shown to improve the performance of machine learning algorithms by making the data more consistent and noise-free (Rahman, 2019). Additionally, statistical models such as the hidden Markov model (HMM) are used for sequence analysis and have demonstrated high accuracy and reliability in biological data analysis (Rahman, 2019).

4.3 Network analysis

Network analysis is a powerful tool for understanding the complex interactions within biological systems. By representing biological entities (e.g., genes, proteins) as nodes and their interactions as edges, network analysis can reveal the underlying structure and dynamics of biological networks. Techniques such as graph-based algorithms and network-based clustering are used to identify key components and modules within these networks (Kashyap et al., 2015; Jin et al., 2020). Deep learning approaches have also been integrated with network analysis to handle large and heterogeneous graph data structures, enabling the extraction of meaningful information from complex biological networks (JaseenaK and Kovoor, 2018; Jin et al., 2020). This integration has facilitated advancements in areas such as disease network analysis, drug discovery, and the identification of therapeutic targets (Kashyap et al., 2015; Jin et al., 2020).

5 Applications of Big Data Analytics in Biology

5.1 Genomics and transcriptomics

Big data analytics has significantly impacted the fields of genomics and transcriptomics, enabling researchers to handle and interpret vast amounts of data generated by high-throughput sequencing technologies. The integration of big data analytics in genomics has facilitated the rapid sequencing of genomes, which was exemplified by the Human Genome Project. This project, which initially took 13 years and over $3 billion, can now be accomplished in just a few days for a fraction of the cost (Li and Chen, 2014). The development of next-generation sequencing (NGS) technologies, such as whole-genome sequencing (WGS) and whole-exome sequencing (WES), has further accelerated the generation of genomic data, allowing for comprehensive studies of genetic variations and their implications in various biological processes and diseases (Hien et al., 2021).

Machine learning algorithms have been particularly useful in the analysis of genomic data, providing tools for the annotation of sequence elements and the integration of epigenetic, proteomic, and metabolomic data (Libbrecht and Noble, 2015). These algorithms help in identifying clinically actionable genetic variants, which are crucial for the development of personalized medicine (He et al., 2017). The integration of genomic data with electronic health records (EHRs) has also opened new avenues for individualized diagnostic and therapeutic strategies, although it presents challenges in data manipulation and management (He et al., 2017).

5.2 Proteomics and metabolomics

Proteomics and metabolomics are other critical areas where big data analytics have made substantial contributions. The advancements in mass spectrometry and other analytical methods have increased the intersection between proteomics and big data science, enabling the generation of large-scale proteomic and metabolomic datasets (Perez-Riverol and Moreno, 2019). The integration of these datasets with transcriptomic data provides a more comprehensive understanding of biological systems, as it allows for the analysis of gene expression, protein translation, and post-translational modifications in a unified manner (Kumar et al., 2016).

High-throughput strategies, such as the sample preparation for multi-omics technologies (SPOT), have been developed to enhance the efficiency of multiomic analyses. These strategies enable the simultaneous analysis of transcriptomic, proteomic, and metabolomic data from a common sample, thereby reducing the resources required and increasing the throughput of multiomic experiments (Gutierrez et al., 2018). Additionally, bioinformatics tools like Metabox facilitate the deep phenotyping analytics of metabolomic data, supporting its integration with proteomic and transcriptomic contexts (Wanichthanarak et al., 2017). The use of software containers and workflow environments, such as Galaxy and Nextflow, has further improved the scalability and reproducibility of proteomic and metabolomic data analysis. These tools allow for the distribution of analytics tasks across multiple computational resources, addressing the challenges of handling large and complex datasets (Perez-Riverol and Moreno, 2019). The integration of these high-throughput and scalable approaches is essential for addressing complex clinical and biological questions, ultimately leading to a better understanding of disease mechanisms and the identification of potential therapeutic targets (Gutierrez et al., 2018; Perez-Riverol and Moreno, 2019).

6 Tools and Platforms for Biological Big Data

6.1 Open-source tools

Open-source tools have become indispensable in the realm of biological big data due to their flexibility, cost-effectiveness, and community-driven development. One notable example is TBtools, a comprehensive toolkit designed for interactive analyses of big biological data. TBtools offers over 100 functions for tasks ranging from bulk sequence processing to interactive data visualization, all within a user-friendly interface. This platform-independent software is freely available and supports various operating systems with Java Runtime Environment 1.6 or newer (Figure 2) (Chen et al., 2020).

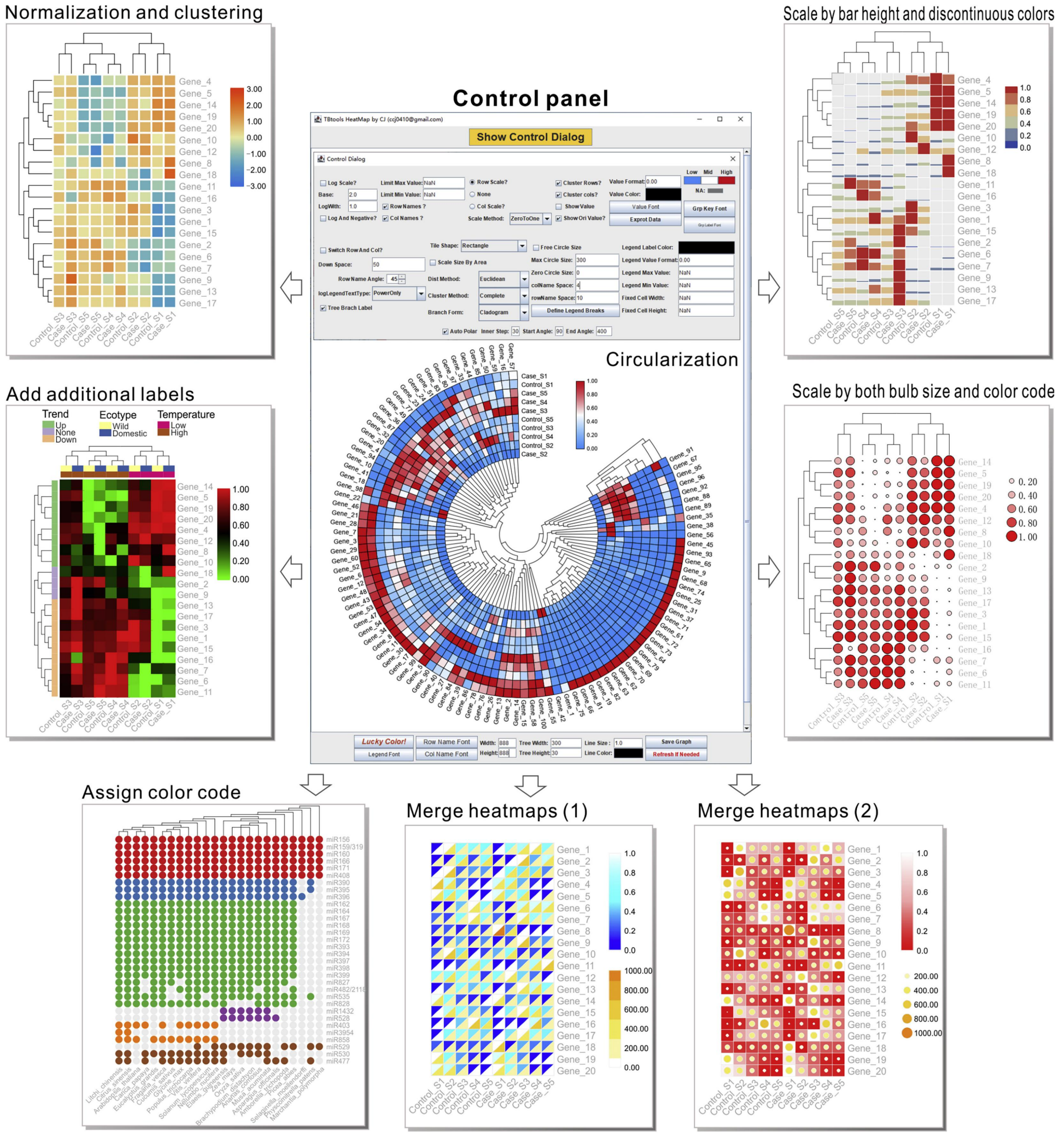

Figure 2 The Powerful Plotting Engine “JIGplot” in TBtools Displays Great Interactability (Adopted from Chen et al., 2020)

Another significant open-source platform is Sherlock, which addresses the challenges of data collection, storage, and analysis in computational biology. Sherlock leverages modern big data technologies like Docker and PrestoDB to enable users to manage, query, and share large and complex datasets efficiently. It supports various structured data types and converts them into optimized storage formats, facilitating quick and efficient distributed analytical queries (Bohár et al., 2022).

OpenBIS is another flexible open-source framework designed for managing and analyzing complex biological data. It allows users to collect, integrate, share, and publish data while connecting to data processing pipelines. openBIS is highly scalable and customizable, making it suitable for a wide range of biological data types and research domains (Bauch et al., 2011). PipeCraft is a flexible toolkit specifically designed for the bioinformatics analysis of high-throughput amplicon sequencing data. It provides a user-friendly graphical interface that links several public tools, allowing users to customize their analysis pipelines according to their specific needs. PipeCraft supports various sequencing platforms and ensures easy customization and traceability of analytical steps (Anslan et al., 2017).

6.2 Commercial software solutions

Commercial software solutions for biological big data often provide robust, enterprise-level support and advanced features that may not be available in open-source tools. These solutions are designed to handle the vast amounts of data generated by modern biological research and offer comprehensive support for data analysis, storage, and management. While the provided data does not include specific examples of commercial software solutions, it is important to note that these solutions typically offer enhanced performance, scalability, and integration capabilities. They often come with dedicated customer support, regular updates, and compliance with industry standards, making them suitable for large-scale and mission-critical applications in biological research.

6.3 Customized pipelines

Customized pipelines are essential for addressing the unique requirements of specific biological research projects. These pipelines often integrate multiple software tools and platforms to create tailored workflows that can handle the complexity and scale of big biological data. The use of application containers and workflows, such as those enabled by Docker, has revolutionized the deployment and reproducibility of computational experiments in genomics. By isolating applications and creating secure, scalable platforms, researchers can significantly reduce the time needed for data analysis and improve the reproducibility of their experiments (Schulz et al., 2016).

High-performance computing (HPC) platforms also play a crucial role in customized pipelines for big biological data analysis. These platforms provide the computational power needed to handle the complexity and volume of biological data, enabling researchers to gain deeper insights into biological functions. HPC platforms are particularly useful for tasks such as genomic sequencing data analysis and protein structure analysis, where traditional computing platforms may fall short (Yin et al., 2017; Yeh et al., 2023).

7 Challenges and Future Directions

7.1 Scalability and performance issues

The rapid growth of biological data, driven by advancements in high-throughput sequencing technologies, has outpaced the capabilities of traditional data analysis platforms. This has necessitated the development of high-performance computing (HPC) platforms and scalable algorithms to handle the massive computational demands of big biological data analytics (Yin et al., 2017). The scalability of bioinformatics software is a critical concern, as it must efficiently manage increasing workloads. Modern cloud computing frameworks like MapReduce and Spark have been employed to implement divide-and-conquer strategies in distributed computing environments, addressing these scalability challenges (Yang et al., 2017). However, ensuring the validity of computational outputs remains a significant issue, requiring robust software testing techniques such as metamorphic testing to ensure the accuracy and reliability of bioinformatics tools (Yang et al., 2017).

7.2 Integration of multimodal data

The integration of multimodal data, particularly in single-cell biology, presents a considerable challenge due to the complexity and heterogeneity of the data involved. Single-cell techniques now enable the simultaneous measurement of multiple data modalities, providing new insights into biological processes that cannot be inferred from a single mode of assay. However, integrating these complex datasets into coherent biological models requires sophisticated computational methods and data visualization approaches (Miao et al., 2021). Strategies for integrating matched data (measured on the same cell) include joint latent space inference and biological causal modeling, while unmatched data (measured on different cells) require methods like annotated group matching and aligning spaces (Miao et al., 2021). Despite these advancements, visualization methods for integrated multimodal single-cell data are still underdeveloped, and future challenges include accounting for modality-specific noise and improving computing efficiency (Miao et al., 2021).

7.3 Ethical and regulatory considerations

The use of big data in health research introduces novel ethical and regulatory challenges that must be carefully considered. The aggregation and analysis of large-scale, heterogeneous data sources can lead to significant preventive, diagnostic, and therapeutic benefits. However, the methodological novelty and computational complexity of big data health research raise unique challenges for Ethics Review Committees (ERCs) and institutional review boards (Ienca et al., 2018). These challenges include ensuring data privacy, managing sensitive personal health data, and addressing power dynamics in the doctor-patient relationship (Galetsi et al., 2019). ERCs must adapt their evaluation criteria to assess the methodological and ethical viability of health-related big data studies, ensuring that the benefits of big data analytics are realized without compromising ethical standards (Ienca et al., 2018). Future research should focus on developing standardized systems for securely extracting and processing private healthcare datasets to mitigate these ethical and regulatory concerns (Galetsi et al., 2019).

Acknowledgments

We sincerely thank the two anonymous reviewers for their valuable opinions and suggestions, and thank Ms. W. Huang from our research team for organizing the materials for this study.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Almasoud A., Al-Khalifa H., Al-Salman A.M., and Lytras M.L., 2020, A framework for enhancing big data integration in biological domain using distributed processing, Applied Sciences, 10(20): 7092.
https://doi.org/10.3390/app10207092

Anslan S., Bahram M., Hiiesalu I., and Tedersoo L., 2017, PipeCraft: flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data, Molecular Ecology Resources, 17: e234-e240.
https://doi.org/10.1111/1755-0998.12692

Bauch A., Adamczyk I., Buczek P., Elmer F., Enimanev K., Glyzewski P., Kohler M., Pylak T., Quandt A., Ramakrishnan C., Beisel C., Malmström L., Aebersold R., and Rinn B., 2011, openBIS: a flexible framework for managing and analyzing complex data in biology research, BMC Bioinformatics, 12: 468-468.
https://doi.org/10.1186/1471-2105-12-468

Bohár B., Fazekas D., Madgwick M., Csabai L., Olbei M., Korcsmáros T., and Szalay-Beko M., 2022, Sherlock: an open-source data platform to store, analyze and integrate big data for computational biologists, F1000Research, 10: 409.
https://doi.org/10.12688/f1000research.52791.2

Chen C.J., Chen H.R., Zhang Y., Thomas H., Frank M.H., He Y.H., and Xia R., 2020, TBtools-an integrative toolkit developed for interactive analyses of big biological data, Molecular Plant, 13(8): 1194-1202.
https://doi.org/10.1016/j.molp.2020.06.009

Clarke R., Ressom H.W., Wang A., Xuan J.H., Liu M.C., Gehan E.A., and Wang Y., 2008, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nature Reviews Cancer, 8(1): 37-49.
https://doi.org/10.1038/nrc2294

Davis-Turak J., Courtney S., Hazard E., Glen W., Silveira W., Wesselman T., Harbin L., Wolf B., Chung D., and Hardiman G., 2017, Genomics pipelines and data integration: challenges and opportunities in the research setting, Expert Review of Molecular Diagnostics, 17: 225-237.
https://doi.org/10.1080/14737159.2017.1282822

Ebrahim A., Brunk E., Tan J., O'Brien E.J., Kim D., Szubin R., Lerman J., Lechner A., Sastry A., Bordbar A., Feist A., and Palsson B., 2016, Multi-omic data integration enables discovery of hidden biological regularities, Nature Communications, 7(1): 13091.
https://doi.org/10.1038/ncomms13091

Galetsi P., Katsaliaki K., and Kumar S., 2019, Values, challenges and future directions of big data analytics in healthcare: a systematic review, Social Science & Medicine, 241: 112533.
https://doi.org/10.1016/j.socscimed.2019.112533

Gligorijević V., and Przulj N., 2015, Methods for biological data integration: perspectives and challenges, Journal of The Royal Society Interface, 12(112): 20150571.
https://doi.org/10.1098/rsif.2015.0571

Goh W.W.B., and Wong L., 2020, The birth of bio-data science: trends, expectations, and applications, Genomics, Proteomics & Bioinformatics, 18(1): 5-15.
https://doi.org/10.1016/j.gpb.2020.01.002

Greene C.S., Tan J.H., Ung M., Moore J., and Cheng C., 2014, Big Data bioinformatics, Journal of Cellular Physiology, 229(12): 1896-1900.
https://doi.org/10.1002/jcp.24662

Gutierrez D., Gant-Branum R., Romer C., Farrow M., Allen J., Dahal N., Nei Y., Codreanu S., Jordan A., Palmer L., Sherrod S., McLean J., Skaar E., Norris J., and Caprioli R., 2018, An integrated, high-throughput strategy for multiomic systems level analysis, Journal of Proteome Research, 17(10): 3396-3408.
https://doi.org/10.1021/acs.jproteome.8b00302

He K.Y., Ge D., and He M.M., 2017, Big data analytics for genomic medicine, International Journal of Molecular Sciences, 18(2): 412.
https://doi.org/10.3390/ijms18020412

Hien L., Van N., Oanh K.T.P., Ton N.D., Hue H.T.T., Duong N.T., Hằng P.L.B., and Ha N.H., 2021, Genomics and big data: research, development and applications, Vietnam Journal of Biotechnology. 19(3): 393-410.
https://doi.org/10.15625/1811-4989/16158

Ienca M., Ferretti A., Hurst S., Puhan M., Lovis C., and Vayena E., 2018, Considerations for ethics review of big data health research: a scoping review, PLoS ONE, 13(10): e0204937.
https://doi.org/10.1371/journal.pone.0204937

Jan B., Farman H., Khan M., Imran M., Islam I., Ahmad A., Ali S., and Jeon G., 2017, Deep learning in big data analytics: a comparative study, Comput. Electr. Eng., 75: 275-287.
https://doi.org/10.1016/J.COMPELECENG.2017.12.009

JaseenaK U., and Kovoor B., 2018, A survey on deep learning techniques for big data in biometrics, International Journal of Advanced Research in Computer Science, 9: 12-17.
https://doi.org/10.26483/IJARCS.V9I1.5136

Jin S.T., Zeng X.X., Xia F., Huang W., and Liu X.R., 2020, Application of deep learning methods in biological networks, Briefings in Bioinformatics, 22(2): 1902-1917.
https://doi.org/10.1093/bib/bbaa043

Juan H.F., and Huang H.C., 2023, Quantitative analysis of high-throughput biological data, Wiley Interdisciplinary Reviews: Computational Molecular Science, 13(4): e1658.
https://doi.org/10.1002/wcms.1658

Kamal M., Ripon S., Dey N., Ashour A., and Santhi V., 2016, A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset, Computer Methods and Programs in Biomedicine, 131: 191-206.
https://doi.org/10.1016/j.cmpb.2016.04.005

Kashyap H., Ahmed H.A., Hoque N., Roy S., and Bhattacharyya D.K., 2015, Big data analytics in bioinformatics: a machine learning perspective, Arxiv preprint arxiv, 2015: 1506.05101.

Kumar D., Bansal G., Narang A., Basak T., Abbas T., and Dash D., 2016, Integrating transcriptome and proteome profiling: strategies and applications, Proteomics, 16(19): 2533-2544.
https://doi.org/10.1002/pmic.201600140

Li Y., and Chen L., 2014, Big biological data: challenges and opportunities, Genomics, Proteomics & Bioinformatics, 12(5): 187-189.
https://doi.org/10.1016/j.gpb.2014.10.001

Libbrecht M., and Noble W., 2015, Machine learning applications in genetics and genomics, Nature Reviews Genetics, 16: 321-332.
https://doi.org/10.1038/nrg3920

Mardis E., 2016, The challenges of big data, Disease Models & Mechanisms, 1(2): 293-314.
https://doi.org/10.1242/dmm.025585

Miao Z., Humphreys B., McMahon A., and Kim J., 2021, Multi-omics integration in the age of million single-cell data, Nature Reviews Nephrology, 17: 710-724.
https://doi.org/10.1038/s41581-021-00463-x

Najafabadi M.M., Villanustre F., Khoshgoftaar T.M., Seliya N., Wald R., and Muharemagic E., 2015, Deep learning applications and challenges in big data analytics, Journal of Big Data, 2: 1-21.
https://doi.org/10.1186/s40537-014-0007-7

O'Driscoll A., Daugelaite J., and Sleator R., 2013, 'Big data', hadoop and cloud computing in genomics, Journal of Biomedical Informatics, 46(5): 774-781.
https://doi.org/10.1016/j.jbi.2013.07.001

Perez-Riverol Y., and Moreno P., 2019, Scalable data analysis in proteomics and metabolomics using biocontainers and workflows engines, Proteomics, 20(9): 1900147.
https://doi.org/10.1002/pmic.201900147

Rahman A., 2019, Statistics-based data preprocessing methods and machine learning algorithms for big data analysis, International Journal of Artificial Intelligence, 17: 44-65.

Schulz W.L., Durant T.J.S., Siddon A.J., and Torres R., 2016, Use of application containers and workflows for genomic data analysis, Journal of Pathology Informatics, 7(1): 53.
https://doi.org/10.4103/2153-3539.197197

Shukla R., Yadav A.K., and Singh T.R., 2021, Application of deep learning in biological big data analysis, Large-Scale Data Streaming, Processing, and Blockchain Security, 2024: 225-250.
https://doi.org/10.4018/978-1-7998-3444-1.ch006

Tariq M., Haseeb M., Aledhari M., Razzak R., Parizi R., and Saeed F., 2020, Methods for proteogenomics data analysis, challenges, and scalability bottlenecks: a survey, IEEE Access: Practical Innovations, Open Solutions, 9: 5497-5516.
https://doi.org/10.1109/ACCESS.2020.3047588

Tonidandel S., King E., and Cortina J., 2018, Big data methods, Organizational Research Methods, 21: 525-547.
https://doi.org/10.1177/1094428116677299

Wanichthanarak K., Fan S., Grapov D., Barupal D., and Fiehn O., 2017, Metabox: a toolbox for metabolomic data analysis, interpretation and integrative exploration, PLoS ONE, 12(1): e0171046.
https://doi.org/10.1371/journal.pone.0171046

Wu X., Zhu X., Wu G., and Ding W., 2014, Data mining with big data, IEEE Transactions on Knowledge and Data Engineering, 26: 97-107.
https://doi.org/10.1109/TKDE.2013.109

Xia J., Wang J., and Niu S., 2020, Research challenges and opportunities for using big data in global change biology, Global Change Biology, 26: 6040-6061.
https://doi.org/10.1111/gcb.15317

Yang A., Troup M., and Ho J., 2017, Scalability and validation of big data bioinformatics software, Computational and Structural Biotechnology Journal, 15: 379-386.
https://doi.org/10.1016/j.csbj.2017.07.002

Yeh C.W., Huang C.W., Yang C.L., and Wang Y.T., 2023, A high performance computing platform for big biological data analysis, 2023 9th International Conference on Applied System Innovation (ICASI), 2023: 68-70.
https://doi.org/10.1109/ICASI57738.2023.10179527

Yin Z., Lan H., Tan G., Lu M., Vasilakos A., and Liu W., 2017, Computing platforms for big biological data analytics: perspectives and challenges, Computational and Structural Biotechnology Journal, 15: 403-411.
https://doi.org/10.1016/j.csbj.2017.07.004

Younas M., 2019, Research challenges of big data, Service Oriented Computing and Applications, 13: 105-107.
https://doi.org/10.1007/s11761-019-00265-x

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.