Computational Molecular Biology 2024, Vol.14, No.6, 248-255
http://bioscipublisher.com/index.php/cmb

Research Perspective Open Access

Emerging Techniques in Biological Big Data Processing: From Algorithms to Applications

Shudan Yan
Institute of Life Science, Jiyang College of Zhejiang A&F University, Zhuji, 311800, Zhejiang, China
Corresponding emil: shudan.yan@jicat.org

Computational Molecular Biology, 2024, Vol.14, No.6 doi: 10.5376/cmb.2024.14.0028
Received: 07 Nov., 2024
Accepted: 08 Dec., 2024
Published: 20 Dec., 2024

Preferred citation for this article:
Yan S.D., 2024, Emerging techniques in biological big data processing: from algorithms to applications, Computational Molecular Biology, 14(6): 248-255 (doi: 10.5376/cmb.2024.14.0028)

Abstract With the rapid advancement of biological research, the growth of biological big data has reached unprecedented scale and complexity. This diversity and sheer volume of data present significant challenges in storage, management, and analysis, while simultaneously driving the rapid development of emerging data processing technologies. This study provides an overview of the latest progress in biological big data processing, covering topics from data preprocessing and cleaning techniques to efficient algorithms and computational frameworks, as well as the applications of artificial intelligence and machine learning in disease prediction, genomic analysis, and other fields. It further explores strategies and methods for multi-omics data integration and the implementation of scalable data visualization techniques in the analysis of biological networks and genomic data. Additionally, the article examines the potential applications of cutting-edge technologies such as quantum computing and edge computing in biological big data, along with the future development of automated data processing pipelines. The goal is to contribute to sustained innovation and progress in the field of biological data analysis.

Keywords Biological big data processing; Data preprocessing; Artificial intelligence; Multi-omics data Integration; Quantum computing

1 Introduction

In recent years, biotechnology has advanced rapidly, and high-throughput experiments have become increasingly common. As a result, the amount of data is snowballing. Fields such as genomics, proteomics, biological imaging, and medical imaging are rapidly generating vast amounts of information (Mahmud et al., 2017; Muzio et al., 2020). However, this data is much more than simple text or numbers. Besides its sheer volume, its structure is also complex, placing considerable strain on storage, processing, and analysis.

In fact, traditional data mining methods are now struggling to cope with this massive and complex biological big data (Kamal et al., 2016; Yang et al., 2020). The challenges are particularly acute when the data is imbalanced, high-dimensional, or requires real-time processing. Clearly, more powerful algorithms and tools are needed.

The good news is that technological advances are opening up new possibilities. Machine learning and deep learning are playing an increasingly important role, and more and more researchers are using them to extract useful information from this complex data. These methods are finding applications in everything from DNA sequence alignment to protein function prediction and even disease diagnosis. Simultaneously, algorithms inspired by biological systems, combined with systems engineering approaches, are attempting to address long-standing problems, such as overfitting and dynamic data analysis (He and Wang, 2020; Dou et al., 2023; Pham and Raahemi, 2023).

Of course, efficient big data processing isn't just about accelerating research; more importantly, it's about generating more accurate and reliable results. This article isn't simply a technical introduction; rather, it aims to systematically review the main approaches currently used to process biological big data, introducing some new algorithms and their application scenarios. We'll also focus on tools currently under development to see if they can address data processing challenges that have long plagued researchers.

2 Characteristics and Processing Requirements of Biological Big Data

2.1 Extensive data types and numerous problems

Genomes, protein structures, expression profiles… these are just some of the types of biological big data. With this proliferation of data comes a host of challenges. Different biological systems and processes operate independently, generating a diverse array of data, ranging from the smallest molecules to entire ecosystems (Dall'Alba et al., 2022). While seemingly comprehensive, processing is far from straightforward. Each type of data has its own unique characteristics, forcing us to use different algorithms and tools, making a "one key opens all locks" nearly impossible (Gill and Buyya, 2019).

2.2 Storage is more than just "storing data"

Faced with massive data volumes, traditional storage methods simply cannot cope. Especially as data continues to grow, storage pressures continue to increase. New technologies like cloud computing and NoSQL have alleviated the situation somewhat. However, things are not that simple. Data must be transmitted quickly, accessible, and secure—all challenges. Especially when real-time data processing and analysis are required, many systems clearly struggle (Pal et al., 2020).

2.3 To effectively process this data, a solid foundation is essential

Successful biological data processing relies not on a single technology but on a comprehensive set of infrastructure. Speed and scale require parallel or distributed computing (Almasoud et al., 2019). Data comes from a variety of sources and formats, making it more complex than a jigsaw puzzle, requiring robust data integration methods. Furthermore, useful information in data is often deeply hidden, making it impossible to discover it with the human eye. Machine learning and data mining tools are essential for uncovering it. However, AI alone doesn't guarantee a complete solution—problems like data imbalance and feature extraction still require practical solutions. Another often-overlooked aspect is that the tools must be user-friendly. After all, not every biologist can code.

3 Data Preprocessing and Cleaning Techniques

3.1 Noise reduction and standardization methods

If biological big data is of poor quality, subsequent complex analyses will struggle to produce reliable results. Therefore, removing noise from the data and performing standardization are crucial. While many methods are currently available, not all are suitable for all situations.

Techniques such as robust principal component analysis (PCA) are commonly used in industrial data processing. Their advantage is their ability to handle significant amounts of noise and data inconsistencies (Zhu et al., 2018). However, biological data is particularly complex, and the situation presents a different challenge. For example, the pguIMP tool is particularly well-suited for biological data due to its simple graphical interface, making it suitable for users without programming experience. It also incorporates state-of-the-art machine learning imputation techniques, such as predictive mean matching. While much of the data cleaning process is automated, it also allows for manual adjustments based on specific data (Figure 1) (Malkusch et al., 2021).

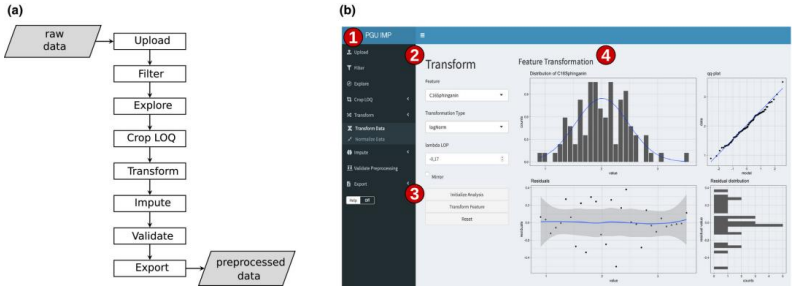

Figure 1 (a) Flowchart of the data engineering pipeline as it is used in the *pguIMP* package; (b) Screenshot of the graphical user

Traditional statistical methods, such as linear transformations, Box-Cox transforms, and hidden Markov models, remain widely used. These methods are particularly useful when we need to quickly standardize data to lay the foundation for subsequent machine learning (Rahman, 2019)

3.2 Missing value handling and outlier detection

Missing values or outliers in data have always been a challenge in data cleaning. This is particularly common in clinical and bioanalysis data. In fact, many solutions to these problems have long been available.

For example, k-nearest neighbor imputation is often used to impute missing values. Compared to simple mean imputation, this method better preserves the original data structure, such as clustering characteristics, and thus produces more accurate analytical results (Malkusch et al., 2021). Of course, not all methods are suitable for every scenario. Clinical data is often disorganized and even unbalanced, necessitating a systematic cleaning process. A review of medical data preprocessing suggests that combining data dimensionality reduction and outlier detection is key to improving clinical data quality (Idri et al., 2018).

In industrial process modeling, the situation is somewhat different. Here, we prefer a robust data mining strategy, aiming to ensure that the model remains robust and reliable even in the face of complex and changing processes. In general, no matter the field, the most important factor in choosing a processing method is the characteristics of the data itself.

4 Efficient Algorithms and Computing Tools

4.1 Mapreduce: handling data in parallel

MapReduce has been widely used in bioinformatics because it splits big jobs into smaller parts. These parts can be processed at the same time on different machines, which works great for huge datasets like DNA sequencing. One example is mrPNN, which mixes probabilistic neural networks with MapReduce. This approach helps classify microarray data faster and more precisely than older methods (Baliarsingh et al., 2020). Another study used MapReduce with the k-nearest neighbor method to handle large, uneven DNA datasets. This not only sped up classification but also saved storage (Kamal et al., 2016). Though MapReduce isn't perfect, these cases show it's still handy for certain tasks.

4.2 Spark vs. hadoop: popular choices for biological data

Spark and Hadoop are both go-to tools for biological big data. While both handle large datasets, Spark is faster because it keeps data in memory, making it better for jobs that need lots of read-write cycles (Guo et al., 2018). For instance, FastKmer uses Spark to pull k-mer details from huge sequences quickly while staying scalable. Hadoop isn't falling behind—tools like HBlast boost sequence alignment speed by running tasks in parallel. There's also BioSpark, which blends Spark and Hadoop to tackle massive, complex datasets like those in simulations (Klein et al., 2017). Though they work differently, both aim to make bioinformatics analysis quicker and more reliable.

4.3 Real-world use: speeding up genome analysis

What really matters is how well these tools perform in real cases. For genome data, Hadoop and Spark have proven useful. Take FastKmer—it uses Spark and has a special feature to evenly spread work across nodes. This fixes common data imbalance issues and boosts speed (Petrillo et al., 2018). MapReduce also plays a role in live genome analysis, from gathering data to crunching numbers later. No tool is perfect, but these examples show one thing clearly: with genomic data growing bigger and more complex, old-school methods aren't enough. We need smart, fast computing solutions.

5 Applications of Artificial Intelligence and Machine Learning in Biological Big Data

5.1 Use of deep learning in gene sequence analysis

Traditional methods often struggle to capture complex patterns in high-dimensional data such as genomes. Deep learning, particularly convolutional neural networks (CNNs), performs well in this area. Already in several studies (Koumakis, 2020; Liu et al., 2020), CNNs have been used to predict the structure of functional gene regions, such as promoters and enhancers, and to analyze changes in gene expression. The "associations" they uncover often escape previous machine learning methods.

5.2 Disease prediction and classification: the strengths of machine learning

If you ask whether modern machines can help predict disease, the answer is more optimistic than you might think. Deep learning models, in particular, can process large amounts of genomic, proteomic, and metabolomic data, identify biomarkers, and even infer potential disease progression pathways. These models can now accomplish tasks such as medical image recognition and gene classification (Cao et al., 2018; Schmidt and Hildebrandt, 2020). This has also driven the development of personalized medicine. However, this isn't applicable in all cases, as data quality and computing resources are crucial.

5.3 Case study: application of convolutional neural networks in protein structure prediction

5.3.1 Infrastructure design

When it comes to using CNNs for protein structure prediction, don't assume their design is simple. Typically, multiple convolutional layers are stacked one on top of another, interspersed with pooling layers, and finally connected to a fully connected layer. This design aims to capture spatial features in protein data, from local to global scales (Jin et al., 2020; Mahmud et al., 2020). Some groups are experimenting with more complex network structures, such as residual networks (ResNet) and recurrent neural networks (RNN), to see if they can improve accuracy, but the effectiveness depends on the specific task and data.

5.3.2 Model training is detailed

There are no shortcuts. Before predicting protein structure, high-quality sequences and corresponding structures must be collected from databases like the PDB. Next, the data must be organized and grouped before training the CNN model. This process is often computationally intensive, especially when backpropagation and gradient descent algorithms are used (Angermueller et al., 2016; Wang and Fang, 2024). Data augmentation and regularization methods are often used to prevent model overfitting, but these methods are not always effective.

5.3.3 Is the model accurate

The results are only known through evaluation. There are a few commonly used evaluation metrics, primarily accuracy, precision, recall, and F1 score. Cross-validation is a common testing method used to assess the stability and reproducibility of a model across different datasets. Numerous studies have demonstrated that CNNs outperform traditional methods in protein structure prediction (Libbrecht and Noble, 2015). However, models still require continuous improvement, especially when dealing with unknown structures.

6 Comprehensive Analysis of Multi-omics Data

6.1 Challenges and solutions in multi-omics data integration

Integrating data from different omics groups sounds appealing, but in practice, it's not that simple. The genomic, transcriptomic, proteomic, and metabolomic levels are like different "languages," with large amounts of data and often mixed with noise. These differences compound and complicate analysis. Especially when the number of features far exceeds the number of samples, models can easily become overloaded, and seemingly reliable results may not be trustworthy (Mirza et al., 2019). High data dimensionality not only easily leads to overfitting but also makes it more difficult to interpret the results.

However, don't assume that these problems are beyond your control. An increasing number of researchers are using tools like deep learning, which are not afraid of high data dimensionality and can detect complex nonlinear relationships. However, deep learning suffers from the "black box" problem, raising questions about the reliability of its results. Another network-based approach uses graphs to represent the connections between different omics models, identifying key nodes or small networks. This can sometimes reveal underlying biological mechanisms (Demirel et al., 2021; Agamah et al., 2022). Often, known biological knowledge is incorporated to help the model better interpret the data.

6.2 Commonly used multi-omics integration methods: bayesian networks and multi-layer models

When it comes to integration methods, Bayesian networks and multi-layer modeling are relatively common. Bayesian networks, like a mental map, can express dependencies between variables and integrate existing biological knowledge. Many studies have used them to predict disease, identify biomarkers, and analyze the

mechanisms of complex pathologies (Li et al., 2016; Cominetti et al., 2023). This type of model can simultaneously process multiple omics data, providing a more comprehensive view of biological systems.

However, Bayesian networks are not the only option. In recent years, multi-layer models have also gained popularity, such as heterogeneous multi-layer networks (HMLNs), which attempt to unify different types of data and present the structure of biological systems through the relationships between layers. HMLNs have been used to reveal the causal relationship between genes and phenotypes, and to study the impact of environmental changes on organisms (Lee et al., 2020). In addition, deep learning has also been introduced into multi-layer models, leveraging its sensitivity to nonlinear relationships and showing great potential in disease classification, marker discovery, and drug response prediction.

7 Ways to Show Big Data Visually

7.1 Drawing biological connections

Reading about molecular interactions in text or tables can be hard to follow. Turning them into pictures makes it easier to spot relationships. For instance, showing how proteins interact as network diagrams works great for complicated data with lots of details (Cruz et al., 2019). In network biology, researchers often combine different genetic data to find important connection points or groups that work together (Charitou et al., 2016). But regular pictures don't show everything. That's why scientists use graph neural networks (GNNs) - special tools that find hidden patterns and help with things like guessing protein jobs or testing medicines. FYI, GNNs are smart computer programs made specifically for connection-type data.

7.2 Viewing genetic info at different scales

Genetic researchers constantly switch between close-up and big-picture views. One moment they're studying single genes, the next they're looking at whole genomes. Multi-level visualization tools help toggle between these views easily. They let scientists sort and rearrange data by different features, helping spot new gene-trait links. While making accurate predictions used to be tough, new computer vision techniques (CNNs) have made faster, better guesses possible (Wang et al., 2020; Huang, 2024). But good software isn't enough - when dealing with huge amounts of data, powerful computers and special chips really help speed things up.

7.3 Real example: working with complicated gene networks

A concrete example explains better than theory. In one study, scientists mixed patient records and lab data to create a detailed network model. The process wasn't simple, but paid off. They used a "brush-and-link" technique to combine smaller networks, letting them compare different parts easily (Vehlow et al., 2015). Smart computer programs didn't just crunch numbers - they highlighted which genes mattered most for sorting data. Tools like "importance maps" show how much single genes affect results, while "response finders" pinpoint genes that trigger strong reactions in the model (Figure 2) (Müller and Gat-Viks, 2020). These aren't just fancy tricks - they genuinely help us understand how genes work together and lead to better disease testing and treatment options.

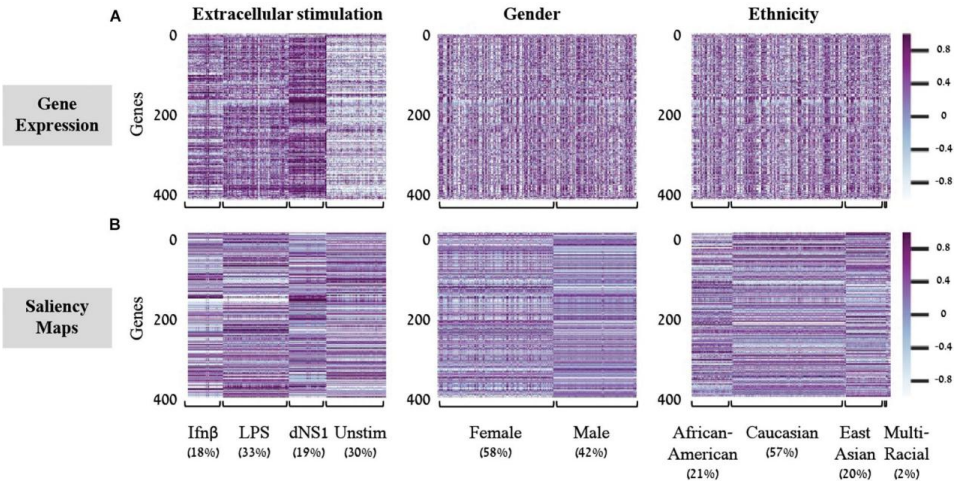

Figure 2 Gene expression versus saliency maps patterns (Adopted from Müller and Gat-Viks, 2020)

8 Future Trends and New Methods

8.1 Processing biological data on edge devices

Before, most biological data had to be sent to faraway servers for analysis. This wasn't just slow—it could also delay urgent tasks. For things like brain-computer interfaces or medical imaging, even a small wait can be a problem. Now, more researchers are trying "edge computing," where data gets processed right where it's collected instead of being sent to a central server. This cuts down on network traffic and speeds things up. As Goh and Wong (2020) pointed out, putting computing power directly on devices makes the whole system work better. Basically, it helps data move faster and more reliably.

8.2 Could quantum computing boost big data analysis in biology

Some might say "quantum computing" still feels like science fiction, but it could actually help a lot with biological data. Unlike regular computers, quantum systems are better at handling tricky problems, like studying genes or finding new drugs (Outeiral et al., 2020). There's even work being done to mix quantum tech with AI to spot diseases like cancer (Emani et al., 2019). But we're not there yet—today's quantum machines still struggle with issues like having too few qubits and being easily disrupted. Until these problems are fixed, we won't see them used everywhere.

8.3 More automation—but it's not the whole answer

With experiments generating more data than ever, labs are using automated tools for jobs like cleaning up data, putting it together, and running models—work that used to take hours by hand (Muzio et al., 2020). The idea is to let scientists spend less time on boring tasks and more on big-picture thinking. Still, automation isn't perfect. Sometimes the system misses odd data points or makes mistakes, so human checks are still needed. Some teams are now testing hybrid approaches, mixing automation with designs inspired by nature to make things more flexible (Pham and Raahemi, 2023).

Acknowledgements

Thanks Jessi Zhang for her assistance in references collection and discussion for this work completion.

Conflict of Interest Disclosure

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

References

Agamah F.E., Bayjanov J.R., Niehues A., Njoku K.F., Skelton M., Mazandu G., Ederveen T., Mulder N., Chimusa E., and Hoen P., 2022, Computational approaches for network-based integrative multi-omics analysis, Frontiers in Molecular Biosciences, 9: 967205. https://doi.org/10.3389/fmolb.2022.967205

Almasoud A.M., Al-Khalifa H.S., and Al-Salman A.S., 2019, Handling big data scalability in biological domain using parallel and distributed processing: a case of three biological semantic similarity measures, BioMed Research International, 2019: 6750296. https://doi.org/10.1155/2019/6750296

Angermueller C., Pärnamaa T., Parts L., and Stegle O., 2016, Deep learning for computational biology, Molecular Systems Biology, 12(7): 878. https://doi.org/10.15252/msb.20156651

Baliarsingh S., Vipsita S., Gandomi A., Panda A., Bakshi S., and Ramasubbareddy S., 2020, Analysis of high-dimensional genomic data using MapReduce based probabilistic neural network, Computer Methods and Programs in Biomedicine, 195: 105625. https://doi.org/10.1016/j.cmpb.2020.105625

Cao C., Liu F., Tan H., Song D., Shu W., Li W., Zhou Y., Bo X., and Xie Z., 2018, Deep learning and its applications in biomedicine, Genomics Proteomics and Bioinformatics, 16: 17-32. https://doi.org/10.1016/j.gpb.2017.07.003

Charitou T., Bryan K., and Lynn D.J., 2016, Using biological networks to integrate visualize and analyze genomics data, Genetics Selection Evolution : GSE, 48(1): 27. https://doi.org/10.1186/s12711-016-0205-1

Cominetti O., Agarwal S., and Oller-Moreno S., 2023, Editorial: advances in methods and tools for multi-omics data analysis, Frontiers in Molecular Biosciences, 10: 1186822. https://doi.org/10.3389/fmolb.2023.1186822

Cruz A., Arrais J.P., and Machado P., 2019, Interactive and coordinated visualization approaches for biological data analysis, Briefings in Bioinformatics, 20(4): 1513-1523.
https://doi.org/10.1093/bib/bby019

Dall'Alba G., Casa P.L., Abreu F.P., Notari D., and Silva S., 2022, A survey of biological data in a big data perspective, Big Data, 10(4): 279-297.
https://doi.org/10.1089/big.2020.0383

Demirel H.C., Arici M.K., and Tuncbag N., 2021, Computational approaches leveraging integrated connections of multi-omic data toward clinical applications, Molecular Omics, 18(1): 7-18.
https://doi.org/10.1039/d1mo00158b

Dou B., Zhu Z., Merkurjev E., Ke L., Chen L., Jiang J., Zhu Y., Liu J., Zhang B., and Wei G.W., 2023, Machine learning methods for small data challenges in molecular science, Chemical Reviews, 123(13): 8736-8780.
https://doi.org/10.1021/acs.chemrev.3c00189

Emani P., Warrell J., Anticevic A., Bekiranov S., Gandal M., McConnell M., Sapiro G., Aspuru-Guzik A., Baker J., Bastiani M., McClure P., Murray J., Sotiropoulos S., Taylor J., Senthil G., Lehner T., Gerstein M., and Harrow A., 2019, Quantum computing at the frontiers of biological sciences, Nature Methods, 18: 701-709.
https://doi.org/10.1038/s41592-020-01004-3

Gill S., and Buyya R., 2019, Bio-inspired algorithms for big data analytics: a survey taxonomy and open challenges, Big Data Analytics for Intelligent Healthcare Management, 2019: 1-17.
https://doi.org/10.1016/B978-0-12-818146-1.00001-5

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

Guo R., Zhao Y., Zou Q., Fang X., and Peng S., 2018, Bioinformatics applications on Apache Spark, GigaScience, 7(8): giy098.
https://doi.org/10.1093/gigascience/giy098

He Q.P., and Wang J., 2020, Application of systems engineering principles and techniques in biological big data analytics: a review, Processes, 8(8): 951.
https://doi.org/10.3390/pr8080951

Huang W.Z., 2024, Application of synthetic biology in directed evolution to enhance enzyme catalytic efficiency, Bioscience Evidence, 14(3): 131-142.
https://doi.org/10.5376/be.2024.14.0015

Idri A., Benhar H., Alemán J., and Kadi I., 2018, A systematic map of medical data preprocessing in knowledge discovery, Computer Methods and Programs in Biomedicine, 162: 69-85.
https://doi.org/10.1016/j.cmpb.2018.05.007

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

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

Klein M., Sharma R., Bohrer C., Avelis C., and Roberts E., 2017, Biospark: scalable analysis of large numerical datasets from biological simulations and experiments using Hadoop and Spark, Bioinformatics, 33: 303-305.
https://doi.org/10.1093/bioinformatics/btw614

Koumakis L., 2020, Deep learning models in genomics; are we there yet, Computational and Structural Biotechnology Journal, 18: 1466-1473.
https://doi.org/10.1016/j.csbj.2020.06.017

Lee B., Zhang S., Poleksic A., and Xie L., 2020, Heterogeneous multi-layered network model for omics data integration and analysis, Frontiers in Genetics, 10.
https://doi.org/10.3389/fgene.2019.01381

Li Y., Wu F., and Ngom A., 2016, A review on machine learning principles for multi-view biological data integration, Briefings in Bioinformatics, 19: 325-340.
https://doi.org/10.1093/bib/bbw113

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

Liu J., Li J., Wang H., and Yan J., 2020, Application of deep learning in genomics, Science China Life Sciences, 63: 1860-1878.
https://doi.org/10.1007/s11427-020-1804-5

Mahmud M., Kaiser M., Hussain A., and Vassanelli S., 2017, Applications of deep learning and reinforcement learning to biological data, IEEE Transactions on Neural Networks and Learning Systems, 29: 2063-2079.
https://doi.org/10.1109/TNNLS.2018.2790388

Mahmud M., Kaiser M., Mcginnity T., and Hussain A., 2020, Deep learning in mining biological data, Cognitive Computation, 13: 1-33.
https://doi.org/10.1007/s12559-020-09773-x

Malkusch S., Hahnefeld L., Gurke R., and Lötsch J., 2021, Visually guided preprocessing of bioanalytical laboratory data using an interactive R notebook (pguIMP), CPT: Pharmacometrics and Systems Pharmacology, 10: 1371-1381.
https://doi.org/10.1002/psp4.12704

Mirza B., Wang W., Wang J., Choi H., Chung N.C., and Ping P., 2019, Machine learning and integrative analysis of biomedical big data, Genes, 10(2): 87.
https://doi.org/10.3390/genes10020087

Müller R., and Gat-Viks I., 2020, Exploring neural networks and related visualization techniques in gene expression data, Frontiers in Genetics, 11: 402.
https://doi.org/10.3389/fgene.2020.00402

Muzio G., O'Bray L., and Borgwardt K., 2020, Biological network analysis with deep learning, Briefings in Bioinformatics, 22: 1515-1530.
https://doi.org/10.1093/bib/bbaa257

Outeiral C., Strahm M., Shi J., Morris G.M., Benjamin S.C., and Deane C.M., 2020, The prospects of quantum computing in computational molecular biology, Wiley Interdisciplinary Reviews: Computational Molecular Science, 11(1): e1481.
https://doi.org/10.1002/wcms.1481

Pal S., Mondal S., Das G., Khatua S., and Ghosh Z., 2020, Big data in biology: the hope and present-day challenges in it, Gene Reports, 21: 100869.
https://doi.org/10.1016/j.genrep.2020.100869

Petrillo U., Sorella M., Cattaneo G., Giancarlo R., and Rombo S.E., 2018, Analyzing big datasets of genomic sequences: fast and scalable collection of k-mer statistics, BMC Bioinformatics, 20(Suppl 4): 138.
https://doi.org/10.1186/s12859-019-2694-8

Pham T., and Raahemi B., 2023, Bio-inspired feature selection algorithms with their applications: a systematic literature review, IEEE Access, 11: 43733-43758.
https://doi.org/10.1109/ACCESS.2023.3272556

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

Schmidt B., and Hildebrandt A., 2020, Deep learning in next-generation sequencing, Drug Discovery Today, 26: 173-180.
https://doi.org/10.1016/j.drudis.2020.10.002

Vehlow C., Kao D.P., Bristow M.R., Hunter L.E., Weiskopf D., and Görg C., 2015, Visual analysis of biological data-knowledge networks, BMC Bioinformatics, 16(1): 135.
https://doi.org/10.1186/s12859-015-0550-z

Wang M., and Fang J., 2024, A new chapter in sugarcane genomics: constructing the R570 reference genome and the future of agricultural biotechnology, Genomics and Applied Biology, 15(1): 8-11.
https://doi.org/10.5376/gab.2024.15.0002

Wang G., Pu P., and Shen T., 2020, An efficient gene bigdata analysis using machine learning algorithms, Multimedia Tools and Applications, 79: 9847-9870.
https://doi.org/10.1007/s11042-019-08358-7

Yang A., Zhang W., Wang J., Yang K., Han Y., and Zhang L., 2020, Review on the application of machine learning algorithms in the sequence data mining of DNA, Frontiers in Bioengineering and Biotechnology, 8: 1032.
https://doi.org/10.3389/fbioe.2020.01032

Zhu J., Ge Z., Song Z., and Gao F., 2018, Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annu, Rev, Control., 46: 107-133.
https://doi.org/10.1016/j.arcontrol.2018.09.003