Computational Molecular Biology 2024, Vol.14, No.6, 266-275
http://bioscipublisher.com/index.php/cmb

Review and Progress Open Access

The Role of High-Performance Computing in Modern Biology: Tackling Big Data Challenges

Hongpeng Wang, Shiying Yu
Biotechnology Research Center, Cuixi Academy of Biotechnology, Zhuji, 311800, Zhejiang, China
Corresponding author: shiying.yu@cuixi.org
Computational Molecular Biology, 2024, Vol.14, No.6 doi: 10.5376/cmb.2024.14.0030
Received: 12 Nov., 2024
Accepted: 13 Dec., 2024
Published: 26 Dec., 2024

Preferred citation for this article:
Wang H.P., and Yu S.Y., 2024, The role of high-performance computing in modern biology: tackling big data challenges, Computational Molecular Biology, 14(6): 266-275 (doi: 10.5376/cmb.2024.14.0030)

Abstract With the rapid development of sequencing and imaging technologies, an increasing amount of biological data is being generated, making the storage, processing, and analysis of vast amounts of data a challenge nowadays. To address this issue, High-Performance Computing (HPC) has emerged, enabling scientists to swiftly process these big data through parallel computing and cloud platforms, thus becoming a crucial tool for handling biological big data. HPC finds applications in various fields, such as genome assembly, protein structure prediction, and multi-omics integration. HPC encompasses a range of tools, including Slurm, Hadoop, BLAST+, GROMACS, and others. HPC plays a significant role in cancer research, drug development, biodiversity monitoring, and many other aspects. Nowadays, the integration of deep learning, adaptive sampling, and HPC with cloud platforms has also opened up new opportunities. Every coin has two sides, and HPC has its drawbacks as well. Its usage cost is relatively high, operation is complex, and there are issues with data integration. However, on the whole, HPC is gradually transforming the way biological research is conducted and holds great potential for development.

Keywords High-performance computing (HPC); Biological big data; Parallel computing; Genomics and multi-omics; Cloud integration in biology

1 Introduction

Over the past decade, as scientists' research scope has expanded, the amount of data generated by biological research has become enormous and is growing at an unprecedented rate. This growth is primarily due to the rapid development of sequencing technologies. Technologies such as high-throughput sequencing can generate multiple terabytes of data in a single run. Biological data is crucial for both personalized medicine and genetic research (Schmidt and Hildebrandt, 2017). However, the current volume of accumulated biological sequence data is both enormous and complex. Traditional data analysis tools are unable to cope with this vast amount of biological data, posing a significant challenge for biologists and programmers (Yin et al., 2017). More efficient and accurate processing and analysis of this data, and the extraction of useful biological information from it, requires more powerful computational technologies (Cl, 2015).

In this environment, high performance computing (HPC) has emerged as an important tool to solve these problems. High performance computing can quickly receive, process, store and analyze a large number of miscellaneous data, and has made significant contributions to many fields such as genomics, proteomics and systems biology (Warris, 2019). HPC includes technologies such as cloud computing and GPU acceleration. These technologies can compare protein structures and do genome-wide screening, which are very useful in solving complex problems (Oehmen and Cannon, 2008; Cl, 2015). The combination of high-performance computing and bioinformatics can also achieve real-time analysis, customize personalized treatment schemes, and help researchers make breakthroughs in medical treatment, drug development, gene research and other fields (Zhou et al., 2018).

This seminar will give a comprehensive talk on the role of HPC in modern biology, focusing on how it can help us meet the challenge of biological big data. The article will analyze from several aspects. First of all, we will introduce what is special about biological data and what computational difficulties these data will bring. Then, we will see what advanced HPC platforms and technologies are available at present, and how they are used to

optimize data analysis to make analysis faster and more accurate. Then we will see the actual effect of HPC through some specific examples, such as gene comparison, protein structure prediction, etc. Finally, we will discuss the future development direction of HPC in biology, and what technical and social difficulties still exist. It is hoped that this study will not only be helpful to researchers, but also provide some ideas for the further promotion of HPC in biological research.

2 Core Foundations of High-Performance Computing

2.1 HPC system architecture: compute nodes, storage, and communication networks

High performance computing (HPC) system is a special tool for processing large-scale computing data tasks, which is mainly composed of three parts: computing nodes, data storage and communication network. Computing nodes are like small computers responsible for computing. Each "small computer" has multiple CPUs or GPUs, which work together to handle complex tasks. In order to read and write data quickly, HPC system generally uses high-speed hard disk (such as SSD) or distributed file system to store data. The main function of the communication network is to help these computing nodes "talk" and transmit data quickly. High speed communication networks (such as Infiniband) can reduce waiting time and improve data transmission speed (Pillardy, 2007; Zhou et al., 2018).

2.2 Fundamental principles and advantages of parallel computing

Parallel computing is a core technology in HPC. Its basic principle is simple: breaking down a large dataset into many smaller ones, then having multiple computing units process each of these smaller datasets simultaneously, thereby saving significant time. In the fields of bioinformatics and computational biology, parallel computing is a highly reliable approach for processing large amounts of complex data, such as gene sequencing, protein folding simulations, and large-scale biological data processing (Pillardy, 2007; Bukowski et al., 2010; Zhou et al., 2018).

2.3 Comparison and integration of high-performance computing and cloud computing

HPC and cloud computing are powerful computing platforms, but they have their own advantages in different aspects. HPC system is specially built for high performance, and scientists have made many optimizations on its hardware and software. HPC has high speed and low latency, which is especially suitable for large-scale computing tasks. In contrast, cloud computing is more flexible. It is just like ordering takeout. It uses resources on demand and can expand rapidly. It is suitable for tasks with large workload changes. Now, researchers often use HPC and cloud computing together, which can combine the powerful computing power of HPC with the flexibility of cloud computing, and greatly improve efficiency. In the research of Pillardy (2007) and Bukowski (2010), researchers can use HPC to do complex analysis work, and then use cloud services to save data or do some lightweight computing, so that more researchers can deal with the big data problem in Biology (Pillardy, 2007; Bukowski et al., 2010).

3 Sources of Big Data in Modern Biology

3.1 Genomics and multi-omics data

The emergence of high-throughput sequencing technology has completely changed genomics. Now, scientists can obtain a large amount of DNA and RNA sequence data in a very short time. The amount of these data is too large to be processed by ordinary computing tools. So high performance computing (HPC) is needed to help manage, analyze and interpret this information. Next generation sequencing (NGS) technology can quickly generate large data sets. To complete such tasks as genome assembly, variation recognition and functional annotation, HPC must be relied on for complex calculations (Merelli et al., 2014; Yin et al., 2017). And now research is not just about genetic data. Scientists will study transcriptome, proteome and metabolome at the same time. This method is called multiomics. Although this method can provide more comprehensive information, it also makes the amount of data larger and more complex. Therefore, using HPC for integration and analysis becomes indispensable (Maizel, 1998; Merelli et al., 2014).

3.2 Biological imaging and microscopy technologies

In recent years, microscope and biological imaging technology have also made great progress. Cryo electron microscopy, fluorescence microscopy and living cell imaging technologies can capture images with high

resolution, and the size of the data generated is terabytes. These big data also need HPC for image reconstruction, segmentation and quantitative analysis to help scientists extract useful information from them (Oehmen and cannon, 2008; Cl, 2015). With the computing power of HPC and these imaging technologies, researchers can more clearly see the structure and changes inside molecules or cells (Maizel, 1998; Wu et al., 2018).

3.3 Environmental and ecological data

In Environmental Science and ecology research, satellite remote sensing, climate model and ecological monitoring technology will also produce a large amount of data. These data are numerous and miscellaneous, including geographic information, climate parameters and ecological samples. Researchers need to use high-throughput computing (HPC) and other technologies to process these data. HPC assigns tasks to many machines for parallel operation, which can complete simulation or analysis faster (Bourne, 2003; Erickson et al., 2018). HPC can also integrate biological information such as environmental data and genome data to study ecosystem changes and species diversity (Warris, 2019).

4 Applications of High-Performance Computing in Large-Scale Data Analysis

4.1 Parallel computing in sequence alignment and genome assembly

High performance computing (HPC) enables us to process massive biological data at an unprecedented speed, which has completely changed the way of sequence alignment and genome assembly. Traditional comparison methods, such as Smith Waterman algorithm, require a lot of calculations, which are slow and inefficient. Later, parallel computing technology appeared, such as GPU and cloud computing, which greatly improved the processing speed. For example, there is a tool called pasvas, which uses NVIDIA's GPU, can quickly complete sequence alignment without compromising accuracy, and can provide detailed results (Warris, 2019). Some platforms, such as Hadoop, are also used to develop parallel protein structure alignment algorithms, which can process faster (Cl, 2015). These technological breakthroughs not only solve the problem of slow computing, but also help scientists analyze genome data more deeply.

4.2 Large-scale single-cell data processing and analysis

Analyzing single cell data is a very difficult task. Because each sequencing will generate a large number of complex data, sometimes even reaching the TB level. In the face of such data, ordinary computers are not enough, and HPC platform is needed to help. Such as grid computing or GPU based systems are very useful. For example, some laboratories will first preprocess some particularly time-consuming parameter calculations in the local grid system, which can speed up the subsequent genome-wide analysis (Warris, 2019). The combination of HPC and omics also enables some basic biomolecular data to be quickly transformed into medical applications, especially in personalized medicine (Merelli et al., 2014). In general, these technologies not only make the analysis faster and the results more accurate, but also adapt to the growing trend of biological data.

4.3 Case study: HPC in genome-wide association analysis

Genome wide association analysis (GWAS) is to analyze a very large data set in order to find out the genetic variation related to a certain trait or disease. This kind of work is almost difficult to complete without HPC support. On the HPC platform, researchers can disperse the algorithm to run on multiple machines to speed up the analysis, especially when calculating the distance matrix or doing multiple sequence alignment, which is very important (Merelli et al., 2014). Some teams will use cloud platforms such as Amazon EC2 to do these analyses. In this way, even small laboratories with few equipment can participate in large-scale research (Merelli et al., 2014). Through these HPC tools, GWAS becomes faster and more accurate, and allows more scientists to participate in complex genetic research.

5 Case Studies: Successful Applications of HPC in Biological Research

5.1 HPC in cancer research

High performance computing (HPC) has greatly promoted cancer research, especially in the processing of genetic data. Through HPC, scientists can quickly analyze a large amount of genome information and find out the key genes related to cancer. These information can help us understand the molecular mechanism of cancer, and can also be used to find new treatment methods (Bourne, 2003; Yin et al., 2017; Zhou et al., 2018). HPC is often used

for gene sequencing of cancer patients, which can quickly find the mutation point and provide the basis for personalized treatment. Moreover, researchers also use HPC to simulate the movement of cancer-related proteins (molecular dynamics simulation), so as to better understand their structure and function, which is particularly critical for the design of new drugs (Liu et al., 2016).

5.2 HPC in drug development

Drug development is also inseparable from HPC. It can help to simulate complex biological processes and quickly screen out potential drug molecules. With the support of HPC, researchers can process thousands of chemical molecules at the same time, greatly improving the efficiency of drug search. Scientists will conduct virtual screening on the HPC platform to find the best candidate drugs that bind to the target protein, and then use the simulation method to study how they interact with each other (Liu et al., 2016; castrignan ò et al., 2020; Sengupta et al., 2020) (Figure 1). The Chinese Academy of Sciences has also done very well in this regard. They have developed some new algorithms and software to help design new drugs, and have also received financial support from the government (Liu et al., 2016).

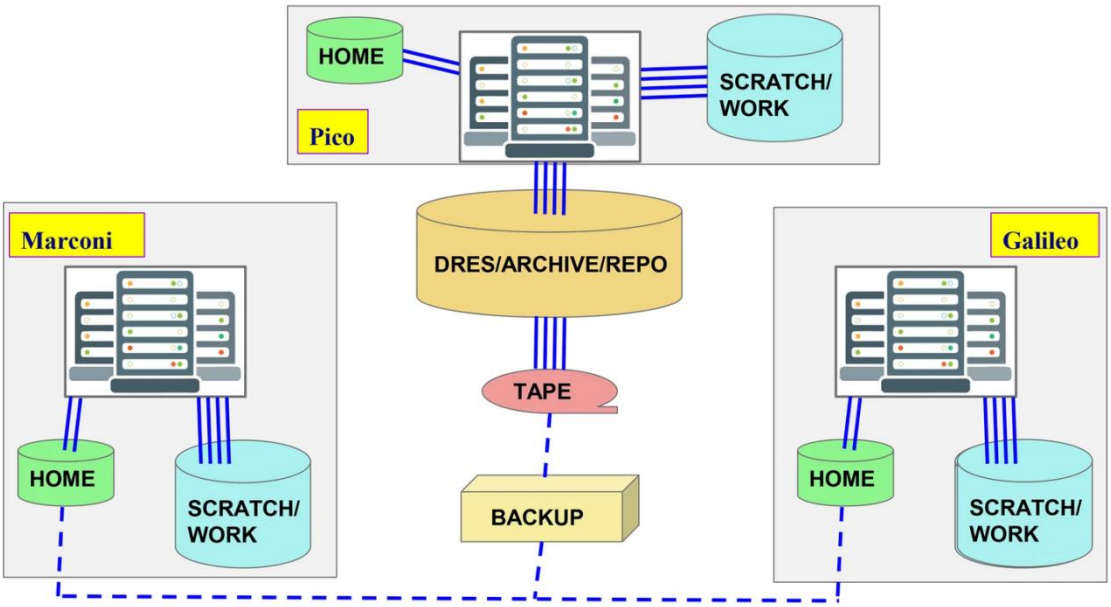


Figure 1 Schematic draw of CINECA system infrastructure . (i) HOME area: intended for source codes, executables, small data files; (ii) SCRATCH area: intended for the output of batch jobs; (iii) WORK area: output of batch jobs as well as for secure sharing within the project team; (iv) DRES: intended as a medium/long term repository and as a shared area within the project team and across HPC platforms; (v) tape area: i personal long term archive area - via Linear Tape File System (LTFS) (Adopted from Castrignanò et al., 2020)

5.3 HPC in environmental and ecological research

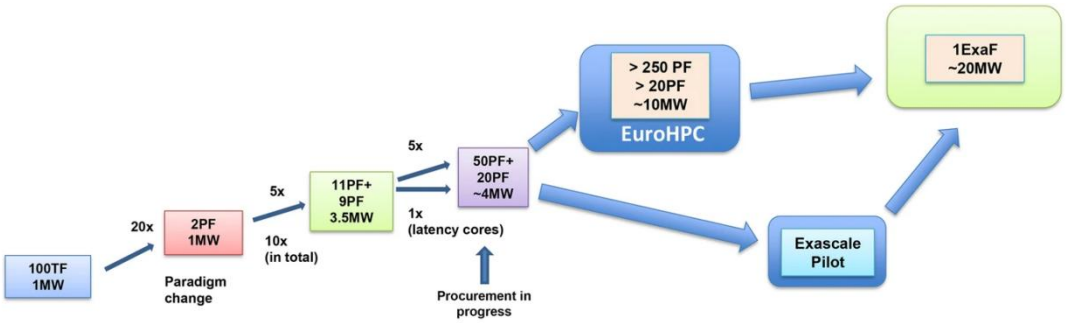
HPC also plays an important role in environmental and ecological research. Many studies need to deal with a large amount of data related to meteorology, geography and ecology. At this time, HPC is particularly useful. Researchers can use HPC to simulate weather changes, such as predicting hurricanes and analyzing the impact of climate change on ecosystems. These simulations involve many parameters and data, which can only be completed by powerful computing power (Schreen et al., 2019). HPC can also integrate environmental data from different places or sources, so as to have a more comprehensive understanding of ecological change trends and help formulate more scientific protection measures (Figure 2) (Yin et al., 2017; Castrignanò et al., 2020).

6 HPC Advancements in Structural Biology

6.1 Protein folding prediction and molecular dynamics simulations

High-performance computing (HPC) has greatly benefited structural biology, particularly in protein folding prediction and molecular dynamics (MD) simulations. Because protein structures are inherently complex, accurately predicting their folding processes is challenging and requires significant computing resources.

Traditional MD simulations are often too slow and time-consuming, making them impractical in many cases. However, new approaches, such as adaptive sampling, are now available. This approach is more efficient, can save significant time, and more quickly identify key points in protein dynamics. Tool frameworks like ExTASY automatically configure sampling strategies on HPC platforms, enabling predictions of protein folding without requiring extensive structural knowledge (Hruska et al., 2019). In addition to adaptive sampling, HPC is also being combined with deep learning to improve the accuracy of protein structure prediction. For example, some new systems combine AlphaFold2 with HPC platforms to perform large-scale alignments and predict protein tertiary structure. These systems utilize deep learning models trained on large data sets, resulting in faster and more accurate predictions (Gao et al., 2021).



2009	2012	2016	2019/2020	2021	2023-2025	2025-2027
IBM SP6 Power6	Fermi IBM BGQ PowerA2	Marconi Lenovo Xeon+KNL	Marconi PPI4HPC ICEI - PPIHBP	pre-Exascale with EuroHPC contribution	post-Marconi Exascale pilot technology (National Research Plan)	EuroHPC post Exascale/ National Full Exascale

Figure 2 CINECA roadmap towards systems of exaflop capabilities (Adopted from Castrignanò et al., 2020)

6.2 Accelerated mass spectrometry data processing and protein identification

HPC has also made mass spectrometry (MS) data processing more efficient, particularly for protein identification. Modern mass spectrometry techniques generate massive amounts of data, which are both voluminous and complex, making processing on conventional computers slow or even impossible. Consequently, HPC platforms have been developed specifically to analyze this large amount of data (Yin et al., 2017; Yeh et al., 2023). These platforms enable scientists to more quickly identify the types and quantities of proteins present. This is particularly important for drug discovery, as identifying specific protein targets is a crucial step in drug discovery. HPC platforms also support the development of new algorithms and tools, such as virtual drug screening and molecular dynamics simulations. These techniques help scientists identify potential new drugs and study how they bind to proteins (Liu et al., 2016; Stoilov and Yurukov, 2016). HPC is also being applied to high-resolution imaging techniques such as cryo-electron microscopy (cryoEM). CryoEM often requires processing massive amounts of image data, making it impossible to run without powerful computing power. Many commonly used cryoEM software packages now incorporate HPC capabilities, significantly improving imaging resolution and enabling researchers to see more clearly and understand the true structure of proteins more easily (Fernández, 2008).

7 High-Performance Computing in Multi-Omics Data Integration

7.1 Challenges of multi-omics data integration and HPC solutions

Biological research often integrates diverse data types, such as genomics, transcriptomics, proteomics, and metabolomics. Each of these data types can reach terabytes or even petabytes, representing enormous volumes and complexity (Misra et al., 2019). Different types of data vary in naming, cleaning, standardization, and biomolecule identification, making data integration more challenging (Misra et al., 2019) (Figure 3). High-performance computing (HPC) is crucial to addressing these challenges. HPC provides powerful computing power, enabling data processing and analysis within a reasonable timeframe (Yin et al., 2017; Koppad et al., 2021). HPC solutions, such as massively parallel clusters, grid computing, cloud computing, and on-chip supercomputing, have become

core tools for processing multi-omics data (Merelli et al., 2014). They can simultaneously support computationally intensive and data-intensive tasks, enabling more efficient analysis. Furthermore, these platforms have accelerated the process from raw data to clinical applications (Merelli et al., 2014). Cloud computing, in particular, has become a popular analytical tool due to its affordability and flexibility. It has also simplified the integration and interpretation of phenotypic data (Koppad et al., 2021).

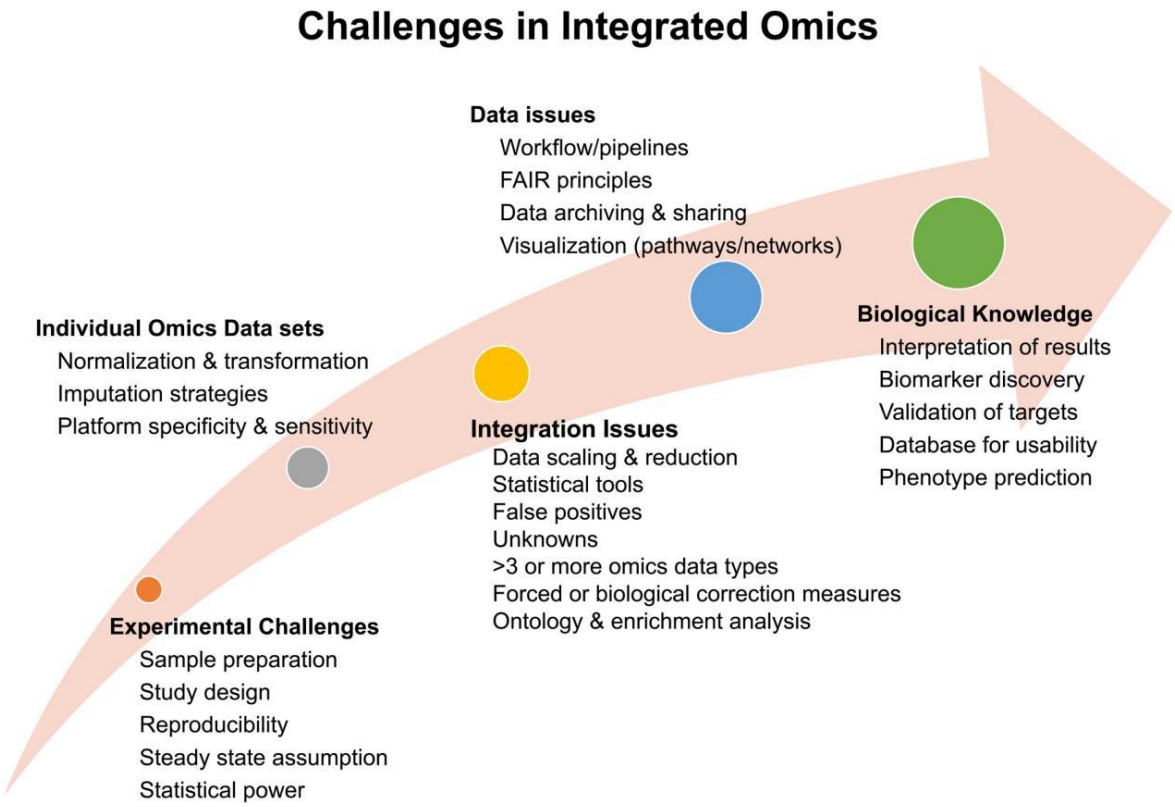


Figure 3 Five challenges associated with integrated omics which encompass (A) experimental challenges, (B) individual omics datasets, (C) integration issues, (D) data issues and (E) biological knowledge (Adopted from Misra et al., 2019)

7.2 Parallel computing methods for high-dimensional data reduction and feature extraction

Multi-omics data are generally high-dimensional. Extracting useful information from them requires dimensionality reduction and feature extraction. These two steps are crucial. However, conventional sequential computing methods are unable to process such complex and large data sets (Liu and Schmidt, 2006). Consequently, scientists have developed parallel computing methods, such as MPI (Message Passing Interface) and OpenMP (Open Multiprocessing). These methods can distribute the work across many processors, allowing simultaneous execution, significantly accelerating analysis (Liu and Schmidt, 2006; Merelli et al., 2014). Using MPI and OpenMP on multi-core clusters to perform distance matrix calculations is significantly faster than traditional methods. This has significantly improved the efficiency of multiple sequence alignment (Merelli et al., 2014). Furthermore, using GPUs for parallel computing has been very useful, achieving, for example, a 181-fold speedup in studying the coagulation system (Merelli et al., 2014). In addition to these, new approaches, such as "skeletal programming" and "pattern-based programming," have made complex computational processes more general and easier to build. The resulting programs are both flexible and easily scalable, making them ideal for dimensionality reduction and feature extraction in high-dimensional data (Liu and Schmidt, 2006).

8 HPC in Ecology and Evolutionary Biology

8.1 Parallel analysis of environmental DNA (eDNA) data and ecological monitoring

8.1.1 eDNA data collection and processing pipeline

Scientists can now directly extract DNA from environmental samples such as water, mud, and air and sequence it to monitor species. This technology, called environmental DNA (eDNA), has significantly transformed

biodiversity research (Deiner et al., 2017; Ruppert et al., 2019). High-throughput sequencing (HTS) can detect many species simultaneously, but this requires a large amount of data. eDNAFlow is a tool for processing this data. It automates the entire process from raw sequences to taxonomic units (ZOTUs) and outputs a table of species counts. Built using Nextflow and Singularity, this workflow can be run on a local computer, cloud platforms, or HPC systems, providing flexibility and reproducibility (Mousavi-Derazmahalleh et al., 2021).

8.1.2 HPC in eDNA data alignment and classification

Processing eDNA data, especially sequence alignment and classification, requires significant computing power. High-throughput sequencing generates massive amounts of data, and HPC can process these sequences in parallel, significantly improving speed and efficiency. eDNAFlow leverages HPC platforms to process large datasets, ensuring fast and accurate alignment and classification (Mousavi-Derazmahalleh et al., 2021). Some studies have also incorporated machine learning, combining it with HPC to better understand the relationship between environmental stress and species diversity. For example, one study used eDNA to monitor freshwater invertebrates (Keck et al., 2023).

8.1.3 Practical applications of eDNA analysis in biodiversity assessment

eDNA metabarcoding is particularly useful for monitoring rare or hard-to-find species, making it crucial for ecological conservation (Cl, 2015; Beng and Corlett, 2020). It can be used in a variety of environments, including lakes, oceans, and land, helping scientists understand changes in species abundance and ecosystems (Deiner et al., 2017). eDNA can also be combined with traditional survey methods to provide a more comprehensive understanding of biodiversity (Beng and Corlett, 2020). Furthermore, combining eDNA analysis with machine learning can also build predictive models for more accurate monitoring of environmental change (Cordier et al., 2017).

8.2 Phylogenetic tree reconstruction and molecular evolution analysis

When studying the evolutionary relationships between species, scientists often need to reconstruct phylogenetic trees. This type of analysis involves large amounts of genomic data and complex evolutionary computations, which places high demands on computing resources. HPC can use parallel computing to distribute these tasks across multiple processors, resulting in much faster analysis. This helps scientists better understand the origins and evolutionary history of species.

8.3 Case study: HPC in analyzing the impact of climate change on ecosystems

HPC can also be used to study the impact of climate change on ecosystems. One case study demonstrates its effectiveness: researchers used HPC to process large amounts of climate and biological data to simulate and predict how climate change will affect ecosystems.These analyses help us understand the potential consequences of climate change, such as species migration, changes in community structure, and alterations in ecological function. This also helps inform conservation strategies and policies.

9 HPC Platforms and Software Tools

9.1 Common HPC platforms: slurm, PBS, and hadoop

High-performance computing (HPC) platforms play a critical role in modern biological research. Modern biological experiments generate vast amounts of data, which require powerful platforms to manage and process. The most commonly used platforms are Slurm, PBS, and Hadoop. Slurm and PBS are scheduling systems primarily responsible for scheduling and allocating computing resources. They ensure that computational tasks are completed in an orderly and efficient manner. Their stability and scalability have led to their widespread use by universities and research institutions (Alnasir, 2021). Hadoop is a different beast. It is a big data processing framework that distributes tasks across many computers. It uses a data processing method called MapReduce, making it ideal for handling the extremely large datasets found in bioinformatics (Lago et al., 2012; O'Driscoll et al., 2015).

9.2 Popular parallel computing tools in biology: BLAST+, Bowtie2, and GROMACS

Processing speed is crucial in biological research. Many data analysis workloads are large, requiring the acceleration of parallel computing tools. BLAST+ is an upgraded version of BLAST that can handle multiple alignments simultaneously, making it ideal for large-scale gene sequence alignments (O'Driscoll et al., 2015). Bowtie2 is another common tool for aligning sequence reads to a reference genome. It optimizes computational speed and memory usage, enabling it to complete large amounts of work in a short period of time (Merelli et al., 2014; Hanussek et al., 2021). GROMACS is a software program used for molecular dynamics simulations, simulating the changes in biomolecules at different time points. It relies heavily on parallel computing and can handle very complex simulations (Hanussek et al., 2021).

9.3 Integration of cloud computing with hybrid HPC platform

Many research teams now combine HPC and cloud computing. This approach leverages the powerful computing power of HPC with the flexibility and scalability of cloud platforms. Cloud platforms like Amazon EC2 can provide large amounts of computing resources on demand. If workloads suddenly increase, cloud platforms can instantly scale up resources, which is very convenient (Lago et al., 2012; De Oliveira et al., 2013). This "hybrid model" is also very useful. Routine tasks can be completed on-premises on HPC systems, while particularly heavy workloads or short-term peaks can be handled by the cloud platform. Tools such as Cirrus are specifically designed to make frameworks like MapReduce more adaptable to cloud environments. This makes it possible to run bioinformatics tasks more efficiently on cloud platforms (Lago et al., 2012).

Acknowledgements

Sincerely thank the anonymous reviewers for their valuable opinions and suggestions. Their professional feedback has greatly improved the quality and rigor of 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

Alnasir J., 2021, Ten simple rules for success with HPC i.e. responsibly bashing that linux cluster, PLoS computational Biology, 17(8): e1009207. https://doi.org/10.1371/journal.pcbi.1009207

Beng K., and Corlett R., 2020, Applications of environmental DNA (eDNA) in ecology and conservation: opportunities challenges and prospects, Biodiversity and Conservation, 29: 2089-2121. https://doi.org/10.1007/s10531-020-01980-0

Bourne P., 2003, High performance computational biology-past progress and future promise, Computational Systems Bioinformatics, 2003: 13. https://doi.org/10.1109/CSB.2003.1227289

Bukowski R., Sun Q., Howard M., and Pillardy J., 2010, BioHPC: computational biology application suite for high performance computing, Journal of Biomolecular Techniques, 21(3 Suppl): S23.

Castrignanò T., Gioiosa S., Flati T., Cestari M., Picardi E., Chiara M., Fratelli M., Amente S., Cirilli M., Tangaro M., Chillemi G., Pesole G., and Zambelli F., 2020, ELIXIR-IT HPC@CINECA: high performance computing resources for the bioinformatics community, BMC Bioinformatics, 21(Suppl 10): 352. https://doi.org/10.1186/s12859-020-03565-8

Cl H., 2015, High-performance computing on very large-scale biological data, Journal of Computational Science, 11: 69-81. https://doi.org/10.4172/2332-0737.1000E117

Cordier T., Esling P., Lejzerowicz F., Visco J., Ouadahi A., Martins C., Cedhagen T., and Pawłowski J., 2017, Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning, Environmental Science and Technology, 51(16): 9118-9126. https://doi.org/10.1021/acs.est.7b01518

De Oliveira D., Ocaña K., Ogasawara E., Dias J., De A. R. Gonçalves J., Baião F., and Mattoso M., 2013, Performance evaluation of parallel strategies in public clouds: a study with phylogenomic workflows, Future Gener. Comput, Syst., 29: 1816-1825. https://doi.org/10.1016/j.future.2012.12.019

Deiner K., Bik H., Mächler E., Seymour M., Lacoursière-Roussel A., Altermatt F., Creer S., Bista I., Lodge D., De Vere N., Pfrender M., and Bernatchez L., 2017, Environmental DNA metabarcoding: transforming how we survey animal and plant communities, Molecular Ecology, 26: 5872-5895. https://doi.org/10.1111/mec.14350

Erickson R.A., Fienen M.N., McCalla S.G., Weiser E.L., Bower M.L., Knudson J.M., and Thain G., 2018, Wrangling distributed computing for high-throughput environmental science: an introduction to HTCondor, PLoS Computational Biology, 14(10): e1006468. https://doi.org/10.1371/journal.pcbi.1006468

Fernández J., 2008, High performance computing in structural determination by electron cryomicroscopy, Journal of Structural Biology, 164(1): 1-6.
https://doi.org/10.1016/j.jsb.2008.07.005

Gao M., Lund-Andersen P., Morehead A., Mahmud S., Chen C., Chen X., Giri N., Roy R., Quadir F., Effler T., Prout R., Abraham S., Elwasif W., Haas N., Skolnick J., Cheng J., and Sedova A., 2021, High-performance deep learning toolbox for genome-scale prediction of protein structure and function, 2021 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC), 2021: 46-57.
https://doi.org/10.1109/mlhpc54614.2021.00010

Hanussek M., Bartusch F., and Krüger J., 2021, Performance and scaling behavior of bioinformatic applications in virtualization environments to create awareness for the efficient use of compute resources, PLoS Computational Biology, 17(7): e1009244.
https://doi.org/10.1371/journal.pcbi.1009244

Hruska E., Balasubramanian V., Lee H., Jha S., and Clementi C., 2019, Extensible and scalable adaptive sampling on supercomputers, Journal of Chemical Theory and Computation, 16(12): 7915-7925.
https://doi.org/10.1021/acs.jctc.0c00991

Keck F., Brantschen J., and Altermatt F., 2023, A combination of machine-learning and eDNA reveals the genetic signature of environmental change at the landscape levels, Molecular Ecology, 32: 4791-4800.
https://doi.org/10.1111/mec.17073

Koppad S., Gkoutos G., and Acharjee A., 2021, Cloud computing enabled big multi-omics data analytics, Bioinformatics and Biology Insights, 15: 11779322211035921.
https://doi.org/10.1177/11779322211035921

Lago J., Ramet D., Torreno O., Karlsson J., Falgueras J., Chelbat N., Krieger M., and Trelles O., 2012, Mr.cirrus: a map-reduce approach for high level cloud computing, F1000 Research, 3.
https://doi.org/10.7490/F1000RESEARCH.1090046.1

Liu T., Lu D., Zhang H., Zheng M., Yang H., Xu Y., Luo C., Zhu W., Yu K., and Jiang H., 2016, Applying high-performance computing in drug discovery and molecular simulation, National Science Review, 3: 49-63.
https://doi.org/10.1093/nsr/nww003

Liu W., and Schmidt B., 2006, Parallel pattern-based systems for computational biology: a case study, IEEE Transactions on Parallel and Distributed Systems, 17: 750-763.
https://doi.org/10.1109/TPDS.2006.109

Maizel J.V., 1998, High performance computing in molecular biology, Proceedings, 1998 IEEE International Conference on Information Technology Applications in Biomedicine ITAB, 98: 40.
https://doi.org/10.1109/ITAB.1998.674669

Merelli I., Pérez-Sánchez H., Gesing S., and D'Agostino D., 2014, High-performance computing and big data in omics-based medicine, BioMed Research International, 2014: 825649.
https://doi.org/10.1155/2014/825649

Misra B.B., Langefeld C., Olivier M., and Cox L.A., 2019, Integrated omics: tools advances and future approaches, Journal of Molecular Endocrinology, 62(1): R21-R45.
https://doi.org/10.1530/JME-18-0055

Mousavi-Derazmahalleh M., Stott A., Lines R., Peverley G., Nester G., Simpson T., Zawierta M., De La Pierre M., Bunce M., and Christophersen C., 2021, eDNAFlow an automated reproducible and scalable workflow for analysis of environmental DNA sequences exploiting nextflow and singularity, Molecular Ecology Resources, 21(5): 1697-1704.
https://doi.org/10.1111/1755-0998.13356

O'Driscoll A., Belogrudov V., Carroll J., Kropp K., Walsh P., Ghazal P., and Sleator R., 2015, HBLAST: Parallelised sequence similarity - a hadoop mapreducable basic local alignment search tool, Journal of Biomedical Informatics, 54: 58-64.
https://doi.org/10.1016/j.jbi.2015.01.008

Oehmen C., and Cannon W., 2008, Bringing high-performance computing to the biologist's workbench: approaches applications and challenges, 125: 012052.
https://doi.org/10.1088/1742-6596/125/1/012052

Pillardy J., 2007, P43-S computational biology applications suite for high-performance computing (BioHPC. net), Journal of Biomolecular Techniques, 18: 16.

Ruppert K.M., Kline R.J., and Rahman M.S., 2019, Past present and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods monitoring and applications of global eDNA, Global Ecology and Conservation, 17: e00547.
https://doi.org/10.1016/J.GECCO.2019.E00547

Schmidt B., and Hildebrandt A., 2017, Next-generation sequencing: big data meets high performance computing, Drug Discovery Today, 22(4): 712-717.
https://doi.org/10.1016/j.drudis.2017.01.014

Schryen G., Kliewer N., and Fink A., 2019, High performance business computing, Business and Information Systems Engineering, 62: 1-3.
https://doi.org/10.1007/s12599-019-00622-2

Sengupta R., Perualila N., Shkedy Z., Biecek P., Molenberghs G., and Bijnens L., 2020, High dimensional surrogacy: computational aspects of an upscaled analysis, Journal of Biopharmaceutical Statistics, 30: 104-120.
https://doi.org/10.1080/10543406.2019.1657128

Stoilov A., and Yurukov B., 2016, Bioinformatics measurements with high performance computing, Biomath Communications, 3(1).
https://doi.org/10.11145/CB.V3I1.637

Warris S., 2019, Application of high performance compute technology in bioinformatics, Wageningen University and Research, 2019.
https://doi.org/10.18174/499180

Wu Y., Xiang Y., Ge J., and Mueller P., 2018, High-performance computing for big data processing, Future Gener. Comput. Syst., 88: 693-695.
https://doi.org/10.1016/j.future.2018.07.054

Yeh C., Huang C., Yang C., and Wang Y., 2023, A high performance computing platform for big biological data analysis, 2023 9th International Conference on Applied System Innovation (ICASI), 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

Zhou L., Rekik I., Yan C., and Wu G., 2018, Special issue on high performance computing in bio-medical informatics, Neuroinformatics, 16: 283.
https://doi.org/10.1007/s12021-018-9393-x