BOHR

BOHR Journal of Biocomputing and Nano Technology
2024, Vol. 2, No. 1, pp. 6–10
DOI: 10.54646/bjbnt.2024.13
www.bohrpub.com

REVIEW

# Is cloud computing the future of bioinformatics?

**Khadeeja Hanna and Hemalatha N**[*†]

St Aloysius (Deemed to be University) Institute of Management and Information Technology (AIMIT), Mangaluru, India

[*]**Correspondence:**
Hemalatha N,
hemalatha@staloysius.ac.in

[†]**ORCID:**
Hemalatha N
0000-0002-7923-4293

**Received:** 06 November 2023; **Accepted:** 08 April 2024; **Published:** 03 May 2024

Bioinformatics is defined as the application of computational and analytical tools to collect and interpret biological data. With the advancement in technology and biological research, a vast amount of data is being produced, necessitating large-scale storage and processing capabilities. Sequencing instruments lack sufficient storage capacity. Cloud computing offers an efficient solution to these challenges, with low management costs. Cloud computing involves storing and accessing data "over the Internet," providing a scalable and flexible system for addressing storage and processing problems. It allows anytime, anywhere access to applications, data sharing, and resources, removing the need for researchers to manage computing clusters. In the field of bioinformatics, cloud computing can provide services such as probe sequencing, quality control, and reporting for next-generation sequencing (NGS) data and other bioinformatics tools. Over the years, cloud computing has become integral to many bioinformatics solutions. Major companies such as Microsoft, Google, and Amazon offer cloud-based platforms, and some cloud-based bioinformatics systems include Apache Hadoop, Dryad, Azure, and BioNimbus. However, concerns have been raised regarding data management, infrastructure, and environmental performance. Overall, cloud computing has the potential to revolutionize bioinformatics, making it more accessible to researchers and professionals. In this research paper, we explore how the introduction of cloud computing has impacted the field of bioinformatics.

**Keywords:** bioinformatics, cloud computing, storage, AWS, Hadoop, MapReduce, BioCloud, BioNimbus, Galaxy

## 1. Introduction

As many of you are aware, bioinformatics involves the computational analysis of biological data, including sequencing, function prediction, and analysis. It has widespread applications in forensic science, genetics, evolutionary science, medicine, agriculture, and the like. The development of sequencing techniques such as next-generation sequencing (NGS), whole-genome sequencing (WGS), whole-exome sequencing (WES), and RNA sequencing has facilitated the study of entire genomes, as opposed to a single gene, along with multiple sequence alignment methods (1) and led to the growth of experimental and clinical data.

Due to the large volume of data, managing it requires substantial storage and processing capacity. Cloud computing, which involves storing and accessing data "over the Internet," plays a crucial role in bioinformatics (2) by enabling efficient data sharing (3) and software access. This significantly enhances the efficiency of bioinformatics analysis and facilitates data distribution and pipeline analysis. Cloud computing offers powerful computing capabilities and the convenience of pay-as-you-go, especially beneficial for small research labs working with large-scale quantitative data (4). In recent years, enormous development of bioinformatics tools was observed.

While cloud computing offers numerous advantages, including data storage, easy access, cost-effective management, and fast processing, it also has limitations such as security vulnerabilities, Internet dependency, difficult transitions, security degradation, and privacy concerns. In bioinformatics, cloud-based tools are categorized as data as a service (DaaS), software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS). One such computer is Apache Hadoop, which processes data using MapReduce and stores an ample amount of data in a system called Hadoop Distributed File System (HDFS) (5). Additionally, major companies such as Amazon, Google, and Microsoft offer their cloud-based platforms such as Dryad, Azure, MPI, PhyloD, SolarGenomics (6), and BioNimbuZ. An open-source distributed file system called Gluster File System (GFS) collects data from multiple servers and serves as global web-based software (7, 8).

## 2. Literature review

The advent of cloud computing has brought significant changes to the field of bioinformatics. Researchers around the world are continuously exploring the use of cloud computing.

In their article, Calabrese B et al. discussed cloud-based microarray data analysis solutions, emphasizing the key issues and challenges associated with using such platforms to store and analyze patient data (2).

Bijetha Seth et al. discussed how cloud computing has affected pharmaceutical companies. They discussed various cloud platforms that can be utilized in genomics and proteomics, providing solutions to security and privacy issues. They also highlighted how Fog is a reliable cloud-based tool that is easily accessible (9).

In the following sections, we will delve into the applications of cloud computing and cloud-based platforms in bioinformatics.

### 2.1. Bioinformatics-timeline

Bioinformatics has a rich history in science. The first analysis of amino acid sequences in the 1950s laid the foundation for this field. At that time, DNA and proteins were not well understood. In 1951, Pauling and Corey published two papers describing alternative structures (α-helices and β-sheets) for proteins. In 1958, the first protein with a three-dimensional structure was created using the X-ray diffraction technique.

The first bioinformatics software was developed in the 1960s to address the challenge of sequencing large proteins. Margaret Dayhoff, recognized as the "father, and mother of bioinformatics" by David J. Lipman, pioneered the use of personal computers for research.

In 1962, Dayhoff and Robert S. Ledley developed COMPROTEIN, the first software for determining protein structure for the IBM 7090 (1). Dayhoff also created an amino acid code in 1965, which was used in the Atlas of Protein Sequence and Structure (the first protein sequence database).

In the 1970s and mid-1980s, Needleman and Wunsch made significant progress in sequence alignments, while advancements in multiple sequence alignment (MSA) were simultaneously taking place. The development of the polymerase chain reaction (PCR) in 1971 made DNA manipulation more accessible. In 1976, the Maxam–Gilbert method became the first widely accepted method for DNA sequencing. Rodger Staden created the first program for Sanger sequencing in 1979.

In 1985, Richard Stallman's free programming contributed to bioinformatics, leading to the development of EMBOSS. The International Nucleotide Sequence Database Collaboration (INSDC) was established at this time. In 1987, Larry Wall created PERL (Practical Extraction And Reporting Language), which by the late 2000s had become the "common language" of bioinformatics. Clustal, a MSA Tool, was created in 1988. Guido Van Rossum developed Python in 1989, and, like PERL, it was used in bioinformatics in the early 2000s. Meanwhile, some mechanical tools were developed after Felsenstein's work on subatomic phylogenetics (1).

Bayesian statistics, widely used in bioinformatics today, was developed in 1990 and was inspired by Felsenstein. In addition, Tim Berners-Lee created World Wide Web (www), an intervention for bioinformatics communication. EMBL's Nucleotide Sequence Database, the release of GenBank in 1992, NCBI's publication in 1994, Genomes publication in 1995, PubMed publication in 1997, and Human Genome publication in 1998 are some examples of developments in the field.

The Human Genome Project, which began in 1991, led to the completion of the large genome assembly by J. Craig Venter in 1995. Several trailblazers of PERL-based programming were created in the mid- to late 1990s to gather entire genome sequencing peruses, such as MIRA, EULER, PHRAP, Celera Assembler, TIGR Assembler, and more. Second-generation sequencing, also known as next-generation sequencing or NGS, made DNA sequencing accessible to a wide audience. Advances in NGS technology and bioinformatics have opened up new possibilities for studying global gene expression, methylation patterns, epigenetic markers, and more. After the advent of technology, the so-called big data grew exponentially, and Moore's law became obsolete (1).

### 2.2. Cloud computing: an overview

Cloud computing originated in the 1960s when J.C.R. Licklider developed the concept while working with ARPANET. It was adopted in 2006 when Amazon created the Elastic Compute Cloud (EC2). ESDS was the first company in India to adopt cloud computing. The latest definition is provided by NIST (10), describing it as a pay-as-you-go computing model (2).

Cloud computing offers computing services "over the Internet," including networking, storage, software, and analytics. This encompasses both the applications delivered over the Internet as a service and the hardware and software components in the data centers that provide the service. Major cloud services are offered by companies such as Google, Microsoft, and Amazon (2).

Cloud computing can be divided into two categories: delivery and service. The distribution model refers to the level of availability in cloud distribution and consists of four types: public, private, hybrid, and collaborative data centers. However, all cloud service models can operate under the "XaaS" paradigm, which stands for "anything as a service" or "everything as a service." The most commonly used types are IaaS, SaaS, DaaS, and PaaS (10). The exponential growth in cloud usage has resulted in many services such as storage and security.

One of the key advantages of the cloud is its ability to provide easy delivery services over the Internet. It offers easy installation, flexible programs, scalable storage, escrow storage, operating system, software, monitoring, and device management. Moreover, it is also cost-effective, time-saving, and reduces uncertainty. Different service models provide different cloud resources. For example, SaaS provides software availability with security and safety functions.

In addition to these advantages, cloud computing has its limitations. Legal, technological, managerial, and security concerns exist. Some of these include integration, security, data locking, suspicious activity, poor data transfer, lack of control, policy for insiders and wrongdoers, operating system mitigation, patches, physical data sources, Internet vulnerabilities, auditing, security traceability, encryption, authentication and authorization, privacy, and security (2).

## 2.3. Cloud computing applications in bioinformatics

In addition to general applications, cloud computing has found applications in bioinformatics. With the increasing data production in bioinformatics, often referred to as big data, cloud computing has emerged as a solution. The cloud-provided services for bioinformatics studies include storage, processing power, communication, data integration, and analysis. Cloud computing has the potential to merge the fundamental principles of bioinformatics, including interpretation, services, and knowledge management, with the unparalleled flexibility and accessibility offered by the cloud (2).

Traditional bioinformatics research, which involves retrieving publicly available data from sources such as NCBI and Ensembl, installing local software, and performing internal analysis, can be enhanced through the transmission of data and software to the cloud and providing them as a service to improve the analysis and storage of bioinformatics data (2).

First, we talked about the significance of the service model offered by the cloud in bioinformatics. SaaS provides diverse tools for sequence analysis, application mapping, genetic analysis, and more. Similarly, PaaS and IaaS offer platforms for high-throughput data analysis, processing genomic and phenotypic data, sequence analysis, and more (11).

### *2.3.1. Cloud platforms in bioinformatics*

Cloud computing has been a major innovation for repositing and exploring colossal data in bioinformatics. One of the main cloud applications used in bioinformatics is Apache Hadoop, which features two key factors: HDFS and MapReduce (4, 12).

Open Data Registry is a cloud service provided by Amazon Web Services (AWS). Another advantageous tool is BioLinux is another public virtual machine that provides a basic bioinformatics research framework. A private cloud-based platform called the BioNimbus Protected Data Cloud (BPDC) (2, 3) is used to manage, secure, analyze, and distribute ample amounts of genomic and phenotypic data (13). miCloud is a platform that integrates with genome sequencers over a local network, facilitating the analysis of genome data from all cloud data. Menon et al. mapped the reads in the cloud using the Burrows–Wheeler transform in MapReduce to scale the entire genome for reads (14). Similarly, SEAL, CloudBurst (15), CloudAligner (16), and Crossbow are prominent MapReduce-based applications that are used for mapping millions of short reads to the reference genome. Eoulsan and Galaxy Cloudman (2) are two other examples. VAT (Variance Annotation Tool) (17) was developed to account for differences in the genomes of different individuals. Another cloud platform is Microsoft Azure, which comes with storage and communication processes and separate, scalable components called "roles." These cloud platforms and additional services support bioinformatics analysis and are categorized into the four main cloud service models (9).

***2.3.1.1. Apache Hadoop.*** Apache Hadoop (12) is a software framework based on the MapReduce paradigm that allows large datasets to be deployed on the cluster farm. Hadoop's ability to instantly process data stored in millions of files—a process that can take weeks to complete—is one of its greatest strengths. Hadoop has three components: HDFS, Hadoop MapReduce, and Hadoop YARN. In HDFS, data is organized in blocks by file size and stored on multiple servers. These blocks are then arbitrarily split and placed on slave machines. Hadoop YARN (One More Asset Moderator) is a group feature of the Hadoop execution layer. YARN is the middleware between HDFS and MapReduce in the Hadoop architecture introduced in Hadoop 2.0. It has two main components: the scheduler, which allocates resources to various processes and schedules resources according to the application's requirements, and the application manager accepts tasks submitted by clients or administrators, and fails to restart the application host when an event occurs.

Hadoop provides a new schema for bioinformatics to analyze the significance of protein statistics, which is important for further research such as protein–ligand docking, clustering, and assembly of protein–ligand complexes (18). A similar version of BLAST is coming soon with the help of the Hadoop MapReduce system (12). When performing a series of queries, CloudBLAST (12) splits them and sends them to different nodes, iterating the entire data string to each node. To accelerate and improve the accuracy of MSAs, Sadasivam et al. proposed a new concept that combines the good nature of the algorithm with the computational modeling of the Hadoop data grid (19). Similar to BLAST, BlastReduce (12) uses a gene-finding association algorithm but uses the LandauVishkin algorithm to rapidly expand genes and find associations with more variables.

Myrna is a distributed tool for identifying large-scale expression in large RNAseq datasets. Crossbow specializes in SNP detection and WGS. Crossbow and Myrna use the Hadoop Stream format with PERL scripts as overlays (20). SAMQA is a similar tool for finding errors in demographic data. It works scalably and reliably using the Hadoop MapReduce framework. Lin and Schatz proposed three designs for the MapReduce graph algorithms, which became the technology for processing large graphs (21).

***2.3.1.2. BioNimbus.*** The BioNimbus (3) platform utilizes a simple, greedy storage concept with large clouds that provide enough space for data storage. BioNimbus is part of a larger project called the Open Science Information Cloud (OSDC) and has variants including Bionimbus People Group Cloud for open data and production and Bionimbus PDC for data management. BioNimbus is a stratified software with three distinct layers: application, core, and framework. The framework layer contains modules that helps in planning communication between federated clouds, while the core layer provides management services such as computational assets, e.g., virtual machines, information capacity, and organizations. The core layer administration includes project management, monitoring, security, discovery, storage, scheduling, and crime prevention services. The application layer is the bridge between BioNimbus and the client (3).

In the experiment created using the original BioNimbus model, it was found that the bioinformatics application performs well when the necessary data is available in the cloud at runtime. The application can also use data stored in another cloud for its operations, allowing sufficient time to modify the data files before starting the actual work. This affects real time, as files containing genomic data are often gigabytes in size. The recommendation is to consider BioClouZ, a potential concept with four domains (availability, cost, downtime, and uptime) weighted by their impacts on information capacity and recuperation. The BioNimbus design was iterated, leading to a rescue called BioNimbuZ, a more powerful and effective platform using ZooKeeper and Avro. Common bioinformatics tools such as BLAST, HMMER/Pfam, and others can be seamlessly integrated with BioNimbuZ if cloud access is provided as a service (3).

***2.3.1.3. Cloud Biolinux.*** The Cloud BioLinux (13) project provides a cloud environment tailored for the bioinformatics community, usable on private or public commercial cloud computing platforms. It uses virtual machines (VMs) for creating business snapshots, which capture all changes made to the server, including data sent by clients, configuration settings, and bioinformatic pipeline analysis results. This allows professionals to collaborate and share data transmission, analysis, and bioinformatics tools in image-based computing.

NEBC BioLinux was among the early pioneers in providing such a model for bioinformatics work. It includes NCBI applications such as blastall and blast + , Staden toolset, EMBOSS, HMMer, and phylip, as well as tools for sequence analysis, clustering, alignment, imaging, editing, phylogeny and NGS. Moreover, itt provides an overview of superior, pre-engineered computing solutions for bioinformatics. The system also includes Fastx utilities, SAM and BAM tools, Genome Analysis Toolkit (GATK), BWA, Novoalign and Necktie aligners, Mummer toolkit and Velvet, SSAKE, Mira, Newbler and Cap3 genome sequence assemblers, and other NGS data analysis tools. Integration with bioinformatics code libraries such as BioPython, BioPerl, BioRuby, Bio Java, R, and the R-Bioconductor programming language is also supported. CloVR, Bioconductor, Qiime, and GMOD are examples of bioinformatics projects that leverage cloud technology for programming (13).

***2.3.1.4. Amazon Web Service.*** AWS is a public cloud platform offering IaaS, PaaS, and SaaS options. Customers can choose the answer that best fits their needs. AWS is a major IaaS provider and is generally considered to provide copper-bottomed, flexible, cost-effective, and intuitive web hosting solutions. AWS-provided services include Amazon Elastic Compute Cloud (EC2), Amazon Simple Storage Service (S3), Amazon SimpleDB, Amazon Simple Queue Service (SQS), Amazon Simple Notification Service (SNS), Amazon CloudFront, and Amazon Elastic MapReduce (EMR). AWS S3 also stores Human Microbiome Project (HMP) data funded by the National Institutes of Health (NIH) and supports bioinformatics operations such as mapping, genotyping, alignment, NGS, and BLAST, which is one of the consistently used tools in bioinformatics research. Images from the BLAST server can be easily deployed on AWS.

For NGS analysis, CloudMan (2) software is available in the AWS cloud framework. Its tools were utilized as a platform for showcasing data and analysis results. AWS S3 stores petabytes of PCAWG-generated data for quick access. The KnowEnG framework in AWS enables biomedical researchers to perform data mining, web mining, and artificial intelligence computations to extract information from genomic data.

***2.3.1.5. Galaxy.*** Galaxy (22, 23) is a web-based bioinformatics analytics platform that renders a public website for data analysis. To address storage, network bandwidth, and data security limitations, Galaxy provides a machine view of tools that can be easily used on AWS called Galaxy Cloudman with a web-based interface. Managing systems and computing in the cloud when using Galaxy Cloudman consists of only two web interfaces: the AWS Management Console and the Galaxy Cloudman web interface. Users can upload data to the server and perform bioinformatics analyses, similar to those available in the Yinhe public server, and access the results after the analysis is completed. Galaxy Cloudman can be seen as a software that works as a PaaS.

## 3. Conclusion

Cloud computing in scientific research, especially in bioinformatics, is still evolving and also handles with ease big data not only from NGS but also from many different data sources. Bioinformatics uses cloud platforms to fathom the shortcomings of big data. Because cloud computing can offer a cost-effective way of storing and processing personal genome data, its application in bioinformatics and lowering sequencing costs will support the development of personalized medicine. This combination of computing power and cloud computing in the field of biological science is set to rapidly transform biology in the near future.

## References

1. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. *Brief Bioinf.* (2019) 20:1981–96.
2. Calabrese B, Cannataro M. Bioinformatics and microarray data analysis on the cloud. *Microarray Data Anal.* (2016) 25:39.
3. Lima D, Moura B, Oliveira G, Ribeiro E, Araujo A, Holanda M, et al. A storage policy for a hybrid federated cloud platform: a case study for bioinformatics. *Proceedings of the 2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing.* Piscataway, NJ: (2014). p. 738–47.
4. Zhou S, Liao R, Guan J. When cloud computing meets bioinformatics: a review. *J Bioinf Comput Biol.* (2013) 11:1330002.
5. Qiu X, Ekanayake J, Beason S, Gunarathne T, Fox G, Barga R, et al. Cloud technologies for bioinformatics applications. *Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers.* Piscataway, NJ: (2009).
6. Yang J. Genomics analysis by pipelined bioinformatics software in the cloud. *Proceedings of the 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA).* Piscataway, NJ: (2017). p. 1–6.
7. Kirby A, Henson B, Thomas J, Armstrong M, Galloway M. Storage and file structure of a bioinformatics cloud architecture. *2019 IEEE Cloud Summit.* Piscataway, NJ: (2019).
8. Galloway J, Thomas J, Henson B, Kirby A, Armstrong M. On the Challenges of Executing Bioinformatic Tools Using a Distributed Cloud Platform. *2019 IEEE Cloud Summit.* Piscataway, NJ: (2019). p. 31–6.
9. Seth B, Dalal S, Kumar R. Securing bioinformatics cloud for big data: Budding buzzword or a glance of the future. *Recent Adv Comput Intell.* (2019) 121:147.
10. Al-Ahmad AS, Kahtan H. Cloud computing review: features and issues. *In 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE).* Piscataway, NJ: (2018).
11. Navale V, Bourne PE. Cloud computing applications for biomedical science: A perspective. *PLoS Comput Biol.* (2018) 14:e1006144.
12. Li X, Jiang W, Jiang Y, Zou Q. Hadoop applications in bioinformatics. *2012 7th Open Cirrus Summit.* Piscataway, NJ: (2012). p. 48–52.
13. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, et al. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC Bioinf.* (2012) 13:1–8.
14. Menon RK, Bhat GP, Schatz MC. Rapid parallel genome indexing with MapReduce. In: Proceedings of the Second International Workshop on MapReduce and Its Applications. (2011). p. 51–8.
15. Schatz M. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics.* (2009) 25:1363–9.
16. Nguyen T, Shi W, Ruden D. CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping. *BMC Res Notes.* (2011) 4:1–7.
17. Habegger L, Balasubramanian S, Chen D, Khurana E, Sboner A, Harmanci A, et al. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics.* (2012) 28:2267–9.
18. Saldanha H, Ribeiro E, Holanda M, Araujo A, Rodrigues G, Walter ME, et al. A cloud architecture for bioinformatics workflows. *In International Conference on Cloud Computing and Services Science.* Piscataway, NJ: (2011).
19. Sadasivam GS, Baktavatchalam G. A novel approach to multiple sequence alignment using hadoop data grids. In: Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud. (2010). p. 1–7.
20. Daugelaite J, O'Driscoll A, Sleator RD. An overview of multiple sequence alignments and cloud computing in bioinformatics. *Int Sch Res Not.* (2013):2013.
21. Lin J, Schatz M. Design patterns for efficient graph algorithms in mapreduce. In: Proceedings of the Eighth Workshop on Mining and Learning with Graphs. (2010). p. 78–85.
22. Wang R, Brewer D, Shastri S, Swayampakula S, Miller J, Kraemer E, et al. Adapting the galaxy bioinformatics tool to support semantic web service composition. *Proceedings of the 2009 Congress on Services-I.* Piscataway, NJ: (2009).
23. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* (2020) 48:W395–402.