Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 9, MARCH 2022

DOI: 10.37943/AITU.2022.59.49.002

**V. Shevtsov**
Master of Technical Sciences, Researcher
shevtsovvladislav111@gmail.com, orcid.org/ 0000-0001-6202-2123
S.Seifullin Kazakh Agrarian Technical University, Kazakhstan

**A. Ismailova**
Doctor of Associate Professor of the Department of Information Systems
a.ismailova@mail.ru, orcid.org/0000-0002-8958-1846
S.Seifullin Kazakh Agrarian Technical University, Kazakhstan

**U. Aitimova**
Doctor of Information Technology, Senior-Lecturer in "Information Systems"
zada@mail.ru, orcid.org/ 0000-0002-0803-7137
S.Seifullin Kazakh Agrarian Technical University, Kazakhstan

**O. Khapilina**
Doctor of biological sciences, head of the "Plants genomics and bioinformatics" laboratory
oksfur@mail.ru, orcid.org/ 0000-0002-7256-568X
National Center for Biotechnology, Kazakhstan

# APPLICATION OF INFORMATION SYSTEMS AND TOOLS IN BIOINFORMATICS

**Abstract.** The pace at which scientific data is produced and disseminated has never been as high as it is currently. Modern sequencing technologies make it possible to obtain the genome of a specific organism in a few days, and the genome of a bacterial organism in less than a day, and therefore researchers from the field of life science are faced with a huge amount of data that needs to be analyzed. In this connection, various fields of science are converging with each other, giving rise to new disciplines. So, bioinformatics is one of these fields, it is a scientific discipline that has been actively developing over the past decades and uses IT tools and methods to solve problems related to the study of biological processes. In particular, a crucial role in the field of bioinformatics is played by the development of new algorithms, tools and the creation of new databases, as well as the integration of extremely large amounts of data. The rapid development of bioinformatics has made it possible to conduct modern biological research. Bioinformatics can help a biologist to extract valuable information from biological data by using tools to process them. Despite the fact that bioinformatics is a relatively new discipline, various web and computer tools already exist, most of which are freely available. This is a review article that provides an exhaustive overview of some of the tools for biological analysis available to a biologist, as well as describes the key role of information systems in this interdisciplinary field.

**Keywords:** Information systems, bioinformatics, databases.

## Introduction

The development and application of algorithms and databases in the field of molecular life sciences began in the field of structural biology in the early 1960s [1]. And also in the field of

DOI: 10.37943/AITU.2022.59.49.002



biochemistry and molecular biology with the determination of the first protein sequences in the period from 1970 to 1980 [2]. Data processing, such as comparing protein and nucleotide sequences, required precise calculations and certain capacities were impossible without the use of computers.

In order to invent a medical drug for a disease, it is important to identify the cause of this disease. Nowadays, scientists more often see the cause of the disease in disorders of the human genome, manifested in the violation of the structure of a particular gene and the appearance of proteins with altered properties, called biological markers of the disease. To prevent a disease or timely treatment, it is necessary to identify the underlying defects in the genome (biological markers of the disease) as early as possible using molecular diagnostics [3]. Such studies require large computational power. In this aspect, computing is given great importance because of its enormous parallelism and ease of computation. Recently, a fairly large number of biological algorithms have been invented, which are used to solve many complex problems in both applied science and medicine. For example, neural computing attempts to simulate the biological nervous systems of living beings in order to enable a significant amount of parallel and distributed processing in computing [4]. So, Bioinformatics has become an important bridge between science and the application of genomics in clinical practice.

There are three main tasks of bioinformatics:

- Development of algorithms and mathematical models for the study of relationships among a large set of biological data.
- Analysis and interpretation of heterogeneous data, including nucleotides, amino acid sequences, and protein structures.
- Implementation of tools for efficient storage, retrieval, and management of large biological databases.

Genetic algorithms are able to simulate the Darwinian evolutionary process through the crossing over and mutation of biological chromosomes. They are successfully used in many bioinformatics tasks that require intelligent approaches to search, optimization, and machine learning [5].

**Modeling as a task of computer science**

The huge amounts of data generated in microbiology can only be processed with intensive information processing. Bioinformatics should help to solve many issues that have arisen in the development of the necessary methods of informatics. Database technology helps store data, locate it, and link it to each other in a variety of ways. The process uses mathematical, but increasingly computer-based methods. The main idea is to be able to represent, model, analyze and predict biological systems and processes qualitatively and quantitatively more comprehensively and thus better understand them.

Before computers were invented, it was only possible to write power balance equations describing the balance of forces in a single triple junction and bulk stability equations for single cells. However, manually studying the interactions between a significant number of cells was impractical due to the large number of equations that had to be written and solved. To make matters worse, when the cells moved, their geometry changed, and the equations had to be redirected and solved for every small increment of movement.

When computers became available to university researchers in the early 1970s, that was a revolution in the science field. With the advent of computers, it was possible to write code to automatically build and solve these equations, and repeat this several times [6]. Then it would be possible to predict the course of cell movement over time and learn something new about how cells behave in model aggregates. Thus, computers have provided researchers with a new way to explore the interactions between different elements of the system. Knowing the

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 9, MARCH 2022

correct sequence of genes must come from a huge number of measurements and the creation of a sequence search system. The journal Nucleic Acids Research publishes a collection of molecular biology databases at the beginning of each year, and the list is often an important acquisition of knowledge. If errors are no longer found, the model becomes a scientific theory in the useful sense of a public database of the exact natural sciences [7].

Data by itself is not something that is useful. To draw a definite conclusion from a set of data, it is necessary to be able to represent, model, analyze and predict the biological structures and processes contained in the data from different points of view. Various applications are used for this. For example, as a result of genome sequencing, which involves the study of the nucleic sequence of genomes, methods for aligning [8], whose main function is to compare sequences for the highest match, search for given sequences in databases such as BLAST [9]. Many other applications are being used or are being developed, for example, to detect functional elements in DNA sequences, compare genomes, to perform phylogenetic analysis, which allows the representation of biological data in the form of a diagram, an image of the line of evolutionary lineage of different species, organisms or genes from a common ancestor [10].

In some areas of microbiology, it was possible to fix regularities in mathematical models. For example, there are models of metabolic processes in the form of systems of partial differential equations [11]. These models then allow predictions to be made about the behavior of the modeled area of nature that can be tested in experiments or falsified: if nature exhibits deviant behavior, then the model is wrong and must be corrected. In biology, besides mathematics, computer science also plays a crucial role in this process of model and theory formation. Calculations of mathematical models are usually so complex that computers are indispensable, and many of them require a complex set of algorithms and data structures. In addition, there are approaches to including original computer methods in the tools of mathematical modeling. The basis is discrete digital models of hardware and software.

**Databases in modern science**

Even at the initial stage of the development of bioinformatics, it became obvious that, in addition to evaluating the analysis of experimental data, the main need was to ensure the availability of biomolecular data around the world through the use of IT.

At the end of 1998, when for the first time, it was possible to obtain the complete genome of a multicellular organism [12]. Such data volumes can only be managed using IT tools. Obviously, database technology is needed, where the data must be stored in such a way that it can be searched for various values. In addition, an important criterion is the availability of a large number of users at the same time, remotely via the Internet, as well as the ability to be brought together from different sources over the network to have protection against damage and data loss.

There is no such a resource that keeps the information about the exact total number of databases, as well as the total amount of data stored in existing databases. However, immediately at the beginning of the use of biological databases, there were attempts to create a database to track the number of existing biological databases and the appearance of new ones, for example, DBcat [13].

Over the past 30 years, since the advent of resources like EMBL Data Library [14] and GenBank [15], databases have become an integral toolkit in modern biological process research. Due to the introduction of information technology and the increase in the number of research projects using databases, every year, more and more new databases appear and are distributed among the scientific community. Many modern databases, such as the European Institute of Bioinformatics, face the challenge of organizing the growing amount of information obtained from molecular biology and genome research.

DOI: 10.37943/AITU.2022.59.49.002



**Problems in the study of biological processes**

Systems biology research is becoming more sophisticated in terms of the capabilities expected from databases and web applications. Despite the inherent nature of the use of databases in science, the databases themselves are not without flaws and have a number of problems that currently exist when using them.

One of the basic problems is the sudden lack of support for the databases themselves. For example, some live databases contain notifications that they are no longer updated, the block of a database, or their database history shows that this is the case. The longevity of modern databases depends on finding a permanent source of funding. Databases, e.g., Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ) maintains one of the largest collections of microbial cultures in the world, including over 29,000 cultures representing a total of about 10,000 species and 2,000 genera [16]. Their free web directories will last as long as the DSMZ exists. They are funded and updated, which is a key part of the mission of their institution and is of great importance for science in general [17].

**Data visualization**

When conducting research, visualization of simulated data is an important tool for data analysis and interpretation. Visualization helps researchers not only understand the data, but also share their results using various visualization tools. The problem with visualization is that in systems research, by visualizing the data flow, the user can get a huge amount of information that will be very difficult to interpret.

Biologists need more effort to become familiar with the various implementations of IT standards. So, in 2009, a graphical representation standard for biology was proposed, which is called SBGN (Systems Biology Graphical Notation) [18]. However, at the moment, there are a small number of tools that include SBGN, such as Cell Designer [19]. SBGN is expected to be adopted by biological tools, databases, and web applications. One of the main problems with implementing SBGN is that none of the web browsers support rendering graphics written in SBGN [20].

**Calculation power**

Recent progress in high throughput sequencing or NGS (new generation sequencing) technology has created a number of new opportunities for biomedical research. At the same time, these advances created a number of challenges in bioinformatics, including quality control, data storage, and the analysis of complex data sets, such as next-generation sequencing data, metabolomic data, proteome data, and electron microscopic structural data. Power analysis is often one of the overlooked aspects of NGS data analysis. The power calculation is the first step in designing a successful study. Nowadays, for modern institutes and other life science-oriented organizations became normal to work with huge biological datasets, therefore, having a powerful computational machine is becoming a standard for such organizations. Unfortunately, due to financing question, not all scientific organizations are able to acquire such expensive equipment so many institutes are currently facing a problem of lack of computational power in the form of a powerful cluster with enough RAM and storage capacity. So collaboration and renting became the solutions to work under big projects.

**Bioinformatics tools and alignment algorithms**

The main instruments of bioinformatics are software tools and the internet. The primary task of bioinformatics is the data analysis including DNA and protein sequences with the use of various open-source programs and databases. Nowadays, for specialists from different fields such as healthcare workers or molecular biologists became possible to implement

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 9, MARCH 2022

research related to biological molecules such as nucleic acids and proteins using standard bioinformatics tools. This does not mean that processing and analysis of genomic data became a trivial task that can be done by everyone. Bioinformatics is a developing discipline, and experienced bioinformaticians now use complex programs to analyze, predict, and store DNA and protein sequence data.

Large commercial enterprises such as pharmaceutical companies employ bioinformaticians to fulfill and service the large-scale and complex bioinformatics needs of these industries [21]. Moreover, an individual researcher will certainly need a bioinformatics conclusion for any complex analysis.

Sequence analysis involves analysis and understanding of the various features of a biomolecule, such as a nucleic acid or protein, that represents a unique function. Retrieving sequences of the corresponding molecules retrieved from public databases became a routine task. If necessary, various tools are used to predict their features related to their function, and structure. Which tool to use to a greater extent depends on the nature of the analysis being carried out (Table 1).

Table 1. Tools for primary sequence analysis

| | Tool | Description | Reference |
|---|---|---|---|
| 1 | PubMed | Central Repository of Life Sciences Literature | [20] |
| 2 | BioEdit | Biological sequence alignment editor | [21] |
| 3 | Mega | Bolecular evolutionary genetics analysis software for microcomputers | [22] |
| 4 | CLUSTAL | Tool for aligning multiple protein or nucleotide sequences | [23] |
| 5 | BLAST | An algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins | [8] |

For example, data mining tools such as PubMed [22], allow you to search and extract data from a wide range of subject areas.

Sequence alignment is a basic bioinformatics tool that is used to establish structural, functional, and evolutionary relationships between sequences, in programs called sequence editors such as BioEdit [23] or Mega (Figure 1) [24].


Figure 1. Visualization of the nucleotide alignment process in the Mega program

New multiple reference alignment databases include PREFAB, SABMARK, OXBENCH, and IRMBASE. However, CLUSTAL [25] is still the most popular alignment tool today, the latest methods provide significantly higher alignment quality and, in some cases, reduced computational costs.

DOI: 10.37943/AITU.2022.59.49.002



An example is SAGA, an alignment program that uses a genetic algorithm that evaluates alignments using OF, which is simply a measure of the quality of a multiple alignment. The principle is to indicate the cost of each pair of equalized balances in each equalization column (replacement cost) and the other gap cost (gap cost). They are added to indicate the global leveling cost. In addition, each pair of sequences is assigned a weight related to their similarity to other pairs. Variations include: (i) using different sets of sequence weights; (ii) different sets of replacement costs [e.g. PAM matrices [26] or the table BLOSUM [27]; (iii) various gap scoring schemes [28]. Then the cost of multiple alignment (A) by the formula:

$$Aligment\ cost\ (A) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} W_{ij}\, COST(A_i, A_j)$$

where COST is the alignment score between two aligned sequences ($A_i$ and $A_j$), while  is their weight.

SAGA is a well-known model for data processing of a single time series and extracting useful information out of it are well-known research topics. A simple example is creating a DNA sequence by analyzing the time series provided by sequencers [29], detecting changes in protein abundances in samples [30], or mapping a set of DNA sequences to a reference [31] are the examples of SAGA model applications.

Generally, the model has found wide application in analyzing a group of time series in order to learn the variations or common patterns across individual signals which allows aligning a pair of signals called a “pairwise alignment”. Aligning more than two signals is called a “multiple alignment” and it can be achieved by doing a series of pairwise alignments. One of the signals can be chosen as a “reference”, then the other signals are pairwise aligned to the reference one at a time [32]. After that, the average of the signals can be chosen as a reference.

The SAGA model has found application not only in molecular biology but in areas like linguistics or psychology. Thus, SAGA was used in comparison of the linguistic characteristics of unimodal (speech only) and multimodal (gesture-accompanied) forms of language use [33]. In terms of linguistic characteristics research, the investigation of SAGA structures and their various implementations is mainly concentrated on dealing gesturally with objects, especially landmarks[34].

Data visualization tools like Jalview [35], IGV [36], and TreeView [37], allow researchers to view data in a graphical representation. These tools use advanced mathematical modeling and statistical inference such as dynamic programming, regression analysis, artificial intelligence for clustering, and sequence analysis.

**Conclusion**

With the rapid growth of annotated genomic sequences in an accessible form, bioinformatics has become a complex and interesting field of science. This is the perfect harmony of statistics, biology, and machine learning methods for analyzing and processing biological information in the form of genes, DNA, RNA, and proteins. The development of this field has become a global enterprise, creating computer networks that provide easy access to biological data and allow the development of programs for the analysis of these data. Numerous international projects aimed at providing databases of genes and proteins are freely available to the entire scientific community via the Internet.

At the moment, bioinformaticians do not have time to process the generated array of biological data. It is reasonable to assume that in the short term, probably the main breakthrough in the field of molecular biology will not be the development of new equipment for sequencing, but the creation of new methods and web-based platforms for biological data processing.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 9, MARCH 2022

This article describes the main tools and methods of bioinformatics, as well as the important role of information technology in this field.

**Acknowledgment**

This work was supported by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan in the framework of program funding for research 2021–2022, program number OR11465422 ("Creation of a biobank of rare and endangered species of flora and fauna of Kazakhstan for the conservation of biodiversity"). The authors wish to thank Eskendir Satekov and Natalia Premina (Altai Botanical Garden, Ridder) for providing pictures of the endangered relict *Allium ledebourianum* species in the wild.

## References

1. Hagen, J. (2000). The origins of bioinformatics. *1*(3), 231-236.
2. Fox, G., Stackebrandt, E., Hespell, R., Gibson, J., Maniloff, J., Dyer, T., . . . Magrum, L. J. S. (1980). The phylogeny of prokaryotes. *209*(4455), 457-463.
3. Borodin, E. (2017). Personified medicine-medicine of the 21st century. *3* (19), 13-15.
4. Poznyak, A. S., Yu, W., Sanchez, E. N., & Perez, J. (1999). Nonlinear adaptive trajectory tracking using dynamic neural networks. *10*(6), 1402-1411.
5. Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.
6. Koza, J. (1994). Genetic programming as a means for programming computers by natural selection. *4*(2), 87-112.
7. Atlamazoglou, V., Thireou, T., Alexandridou, A., & Spyrou, G. (2008). *A high throughput approach to keep alive a web-based database system for multiple search among published bioinformatics tools and databases*. 2008 8th IEEE International Conference on BioInformatics and BioEngineering,
8. Edgar, R. C., & Batzoglou, S. J. C. (2006). Multiple sequence alignment. *16*(3), 368-373.
9. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. (2008). NCBI BLAST: a better web interface. *36*(2), 5-9.
10. Avise, J. C. (2006). *Evolutionary pathways in nature: a phylogenetic approach*. Cambridge University Press.
11. atCREARE, U. A MODEL OF METABOLIC PROCESSES IN A HETEROGENEOUS MILIEU: FUNCTIONAL AND NUMERICAL SOLUTIONS.
12. Hodgkin, J., Paulini, M., & Tuli, M. A. (2012). Genome Mapping and Genomics of Caenorhabditis elegans. In *Genome Mapping and Genomics in Laboratory Animals*. Springer. 17-30.
13. Discala, C., Ninnin, M., Achard, F., Barillot, E., & Vaysseix, G. J. (1999). DBcat: a catalog of biological databases. *27*(1), 10-11.
14. Hamm, G. H., & Cameron, G. N. (1986). The EMBL data library. *14*(1), 5-9.
15. Burks, C., Fickett, J. W., Goad, W. B., Kanehisa, M., Lewitter, F. I., Rindone, W. P., . . . Bilofsky, H. (1985). CABIOS REVIEW: The GenBank nucleic acid sequence database. *1*(4), 225-233.
16. Uphoff, C., & Drexler, H. (1992). Die Deutsche Sammlung von Mikroorganismen und Zellkulturen: Abteilung" Menschliche und tierische Zellkulturen". *9*(1), 39-44.
17. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., . . . Thomas, D. (2003). The UCSC genome browser database. *31*(1), 51-54.
18. Novère, N. L., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., . . . Wimalaratne, S. (2009). The systems biology graphical notation. *27*(8), 735-741.
19. Matsuoka, Y., Funahashi, A., Ghosh, S., & Kitano, H. (2014). Modeling and simulation using CellDesigner. In *Transcription Factor Regulatory Networks*. Springer. 121-145.
20. Sreenivasaiah, K., & Kim, D. (2010). Current trends and new challenges of databases and web applications for systems driven biological research. *1*, 147.
21. Menon, S. (2021). Bioinformatics approaches to understand gene looping in human genome. *6*(7), 170-173.

DOI: 10.37943/AITU.2022.59.49.002



22. Canese, K., & Weis, S. (2013). PubMed: the bibliographic database. *2*(1).
23. Hall, T., Biosciences, I., & Carlsbad, C. (2011). BioEdit: an important software for molecular biology. *2*(1), 60-61.
24. Tamura, K., Dudley, J., Nei, M., Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *24*(8), 1596-1599.
25. Hung, L., Lin, S., Lin, C., Chung, C., Chung, Y. (2015). CUDA ClustalW: An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs. *58*, 62-68.
26. Dayhoff, M. O. (1972). *Atlas of protein sequence and structure*. National Biomedical Research Foundation.
27. Henikoff, S., & Henikoff, J.(1992). Amino acid substitution matrices from protein blocks. *89*(22), 10915-10919.
28. Lipman, D. J., Altschul, S. F., & Kececioglu, J. (1989). A tool for multiple sequence alignment. *86*(12), 4412-4415.
29. Akella, L. M., Rejtar, T., Orazine, C., Hincapie, M., & Hancock, W. S. (2009). CLUE-TIPS, Clustering Methods for Pattern Analysis of LC– MS Data. *8*(10), 4732-4742.
30. Powell, D. W., Weaver, C. M., Jennings, J. L., McAfee, K. J., He, Y., Weil, P. A. (2004). Cluster analysis of mass spectrometry data reveals a novel component of SAGA. *24*(16), 7249-7259.
31. Krebs, A. R., Karmodiya, K., Lindahl-Allen, M., Struhl, K., & Tora, L.(2011). SAGA and ATAC histone acetyl transferase complexes regulate distinct sets of genes and ATAC defines a class of p300-independent enhancers. *44*(3), 410-423.
32. Kaya, H., & Gündüz-Öğüdücü, Ş. (2013). SAGA: A novel signal alignment method based on genetic algorithm. *228*, 113-130.
33. Hahn, F., & Rieser, H. (2010). Explaining speech gesture alignment in mm dialogue using gesture typology.
34. Lücking, A., Bergman, K., Hahn, F., Kopp, S., & Rieser, H. (2013). Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. *7*(1), 5-18.
35. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *25*(9), 1189-1191.
36. Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *14*(2), 178-192.
37. Page, R. (1996). Tree View: An application to display phylogenetic trees on personal computers. *12*(4), 357-358.