Research on the Application of Computer Science in Bioinformatics

Qian Xu¹, Yi Zhang¹,*

¹Institution of Pharmacy, Chongqing Medical and Pharmaceutical College, Chongqing, China, 400113

*Corresponding author e-mail: zhangyi@cqyygz.com

Abstract. Bioinformatics is an emerging discipline that integrates many technologies such as computer application technology, network technology, biology, and statistics, etc. Without the rapid development of computer and network technology, bioinformatics would not be what it is today. Computer science has facilitated the exchange of information and the development of the discipline, which has brought great help to the collection and computational processing of biological data and information.

**Keywords**: Computer Technology, Data Mining, Bioinformatics, Genome

1. Introduction

With the increasingly widespread application of electronic computers, Internet technology, electronic communication technology, etc., people have now entered the information age. Computer science in the information age has been widely used and has developed significantly, and has provided the possibility of development and reinvention in various fields. In the study of bioinformatics, there are many techniques involved in the collection, storage, statistics, analysis and processing of data, such as a single cell, which contains an extremely large number of atoms in bioinformatics research. Bioinformatics is very convenient [1].

2. Fundamentals of combining computer technology and bioinformatics

Before the 1990s, bioinformatics was in the pre-genomic era, because of the discovery of protein electrophoresis, flow cytometry and the most important DNA helix structure, people focused on the sequence analysis of DNA and proteins. The application of computers was relatively rudimentary, mainly for the creation of databases and the design of various algorithms to improve the efficiency of sequence analysis. In 1990, the United States proposed the "Human Genome Project", and the development of bioinformatics entered the genome stage, with the aim of recording the complete human
gene sequence by using the principle of DNA complementarity. Now, with the completion of human genetic sequencing, people realize that the 23 human chromosomes contain 60,000-100,000 genes, and the essence of genes is composed of different base pairs, and these 100,000 genes are about 3 billion base pairs, which shows how huge the amount of information is. It is difficult for a person to understand one ten-thousandth of this information even if he or she has to calculate it manually. At this point, the help of a computer is very necessary [2].

3. Application of computer in bioinformatics

![Diagram showing the relationship between data mining, structural simulation, application of high performance computer in genomics research, development of subbiology tool software](image)

Figure 1. Application

3.1. Data mining

Traditional biology research mainly relies on experiments and observations, and draws conclusions from practice, such as the discovery of Mendel's laws of heredity. In the field of molecular biology, however, because of the introduction of computers, the study of biology has moved toward foresight, and theories can be derived first through logic and deduction, and then verified in experiments. Computers are able to control various types of equipment and can pre-program experiments, saving a lot of manpower for the implementation of experiments. Data mining is done by cleaning and integrating data and then analyzing and predicting trends. For example, protein structure prediction is traditionally performed experimentally, but unknown protein molecule structures cannot be determined experimentally. In modern biological research, computer data mining techniques, such as neural networks and support vector machines, are used to predict the secondary structure of the amino acids in the middle of the window by coding the 20 amino acids that make up the protein and then selecting a window of odd length and using the constructed neural network model or vector machine model. The data mining can also be used for microarray data analysis, protein homology study, etc.

3.2. Structural simulation

When the double helix structure of DNA was discovered, Watson and Crick used wire and iron to build the double helix structure of DNA to help them think and test their hypothesis. Nowadays, we can use
3D modeling software on the computer to save cost and time for building structural models, and we can directly input conditions and data for the computer to calculate the results, which facilitates the testing of structural hypotheses. The complex structure of proteins includes four levels of structure, which is not very realistic to imagine only by human imaginative thinking, and it is very difficult to restore the structure of proteins in reality and make models for large molecule proteins. Computer 3D modeling turns this impossibility into possibility, and makes the structure of genes and proteins more visual, which is very helpful for the study of biological information [3].

3.3. High performance computers in Genomics research

The first task of genomics is to perform large-scale gene sequencing to obtain information about genes and genomes, and it is now common to use a method consisting of a large number of random short sequences spliced by sequences. The determination of all short sequences is performed by a highly automated sequencing machine, which is responsible for converting the physicochemical signals about sequence information (i.e., the arrangement of base pairs A, T, C, and G in DNA) obtained in biological sequencing into digital information that can be processed by a computer, which is simply analyzed by a general computer and gives preliminary results, equivalent to raw experimental data. After obtaining the raw data for gene sequencing, the process of data processing and analysis involving huge amounts of data follows. All of this places high demands on the high-performance computing environment, and with the increasing maturity of gene sequencing technology, a country's advantage in genomics research is largely dependent on the high-speed computing power it can provide [4].

The high-performance computing environment includes three major components: computer hardware systems, operating system software, and networks. The current main way to improve computing power from computer architecture is massively parallel processing, and massively parallel processing (MPP) has long been a cutting-edge research in supercomputer applications. In the United States, the High Performance Computing and Communication (HPCC) program proposed in 1992 has announced the use of supercomputers for human genetic research as a major challenging application, and the goal of the HPCC program is to provide a 3T performance computer. At present, supercomputers designed for parallel processing based on the Von Neumann architecture are divided into two categories: shared common memory and non-shared distributed memory by physical structure, and SIMD and MIMD by data and instruction relationship. Multi-threaded and data stream computers are new architectural design ideas in the development of massive parallelism. Obviously, the development of supercomputers with MPP capability requires huge investment in human, material and financial resources, and the ability to develop and promote the application of supercomputers is very limited in the near future given our national conditions. Therefore, under the existing conditions, it is expedient to strengthen the application of PC clustering technology in bioinformatics and improve the processing capacity of computer systems based on the existing PCs. However, in the long run, it is important and far-reaching for China's bioinformatics and related industries to strengthen the research of high-performance computing system architecture with high parallelism, scalability and programmability.

3.4. Development of software for molecular biology tools

With a high-performance computing platform, it is necessary to develop software tools for biologists to use in molecular biology research, such as sequence splicing software, gene discovery software, gene structure analysis software, and gene function prediction software, which are all essential computational
and analytical tools for biologists. The following discussion focuses on the development of gene discovery software and gene function prediction software [5,6].

Gene finding refers to the identification of useful, relatively useful and useless sequence fragments from the distribution of nucleotides in DNA provided by large-scale gene sequencing in order to discover possible genes and structures. A number of related software have been developed abroad, and a large number of distinctive algorithms have been used, such as nucleotide discourse method, homology comparison algorithm, hidden Markov model, dynamic programming algorithm, linear discriminant analysis algorithm, Fourier analysis method and law-based gene prediction algorithm.

Gene function prediction refers to the prediction of the functions of discovered genes by certain methods or approaches to provide inspiration and guidance for further biological experimental validation. Currently, homology comparison is the main method for gene function prediction. The common algorithms for homology comparison are Smith-Waterman algorithm, BLAST algorithm and FASTA algorithm. The main difficulty of current homology comparison algorithms for gene function prediction is how to study and design algorithms that can provide both high sensitivity and high speed. To solve these problems, the authors believe that it is necessary to make full use of various methods and tools in computer science and to draw on the latest research and development results in computer science, especially in artificial intelligence, such as data mining and artificial neural network techniques in a timely manner [7,8].

Data mining technology is an extremely fast growing research field in computer science at present, which integrates machine learning, statistical analysis and database technology to provide services for decision-oriented use of data in databases. The essence of data mining is knowledge discovery, including rule generation, classification, clustering, sequence analysis, etc. Data mining techniques have great potential for gene finding and gene function prediction software design and development.

Artificial neural network technology is another branch of computer science that is developing fast. Artificial neural network is an information processing system that simulates the structure and characteristics of human brain neurons and the cognitive function of human brain with highly nonlinear dynamics. Through the interconnection of a large number of nonlinear neuronal components, artificial neural networks have the characteristics of parallel processing, associative memory, self-learning, self-referential organization and distributed storage, so ANN for gene and gene structure and function analysis and prediction also has the strength that cannot be underestimated. Taking the previous example of gene discovery software, if we were to design a new algorithm, it would be quite difficult and would not be able to take full advantage of the existing algorithmic ideas or results that have proven to be valuable, but if we consider using a neural network model to integrate the software or algorithms mentioned above with their respective strengths to obtain a more general gene discovery software, we can obtain The authors argue that ANN technology can be used for bioinformatics. The authors argue that the key to the use of ANN techniques for bioinformatics research lies in the selection of the appropriate model from among the many types of artificial neural networks currently available, and in making some necessary improvements to the chosen model, according to the characteristics of biology. Even the possibility of proposing new ANN models is not avoided. In addition, the combination of data mining techniques and artificial neural networks may have the effect of adding one to two.

4. Prospect of bioinformatics application
The development of bioinformatics is still in its initial stage, and the research task of bioinformatics now is mainly the collection of raw data and information. The collection of all human genetic information by the Human Genome Project is only the first step, and it is believed that in the future, the interpretation of the origin and role of each segment of the genetic code and the processing of the genetic code are the mountains that human beings need to climb over [9,10].

Because of the complexity of biological information, people do not know enough about the molecular level of biological information, lack global theoretical guidance, and lack a complete knowledge system construction, so they are easily limited by their thoughts and thinking when interpreting the gene functions. Nowadays, computer application technology is developing rapidly, and now the research in the field of artificial intelligence is advancing by leaps and bounds with remarkable results. In the future, it may allow bioinformatics to be combined with machine learning, so that computers can automatically summarize and analyze, reason and summarize, and derive the correct conclusions when analyzing biological data, which will make the research on the interpretation of biological data information twice as fast with half the effort, and the efficiency will be greatly improved.

Bioinformatics will be the basis for guiding the direction of treatment in the future. It can be assumed to a certain extent that most diseases have a genetic information basis that can be reflected at the molecular level such as genes and proteins, which brings a new direction to treatment. Using computer data models it is possible to design vaccines, disease diagnostics, and drug treatments for the structure of gene molecules or proteins.

5. Future work

The problem of algorithm optimization and modeling is present in bioinformatics in many places. Sequence splicing and assembly, mapping, sequence alignment, phylogenetic analysis, protein structure prediction, molecular simulation and drug design in gene sequencing require efficient algorithms and easy-to-use software support [11].
The public databases on the network are physically scattered all over the world, and the data resources all carry different degrees of redundancy or missing information. It is necessary to establish more than two levels of databases and corresponding processing and analysis tools on the basis of data processing in order to provide fast and efficient information services to scientists all over the world. Accordingly, research should be conducted and corresponding technologies should be developed in various aspects such as data quality control, standardization of data resources, network transmission, integration of multi-database systems, and data mining and knowledge discovery [12].

6. Conclusion

Marx once pointed out that "only when mathematics is successfully used in a discipline does it reach the point of perfection". The difference between bioinformatics and traditional biology lies in the use of mathematical ideas to guide the thinking, mathematical methods, modeling and other means to study biological information. The application of computers is indispensable to the processing of mathematics. In the use of computer tools bioinformatics has become the frontier of life sciences today, through the computer to decipher the genome information, for human understanding of the nature of life deepened a step, for the future diagnosis and treatment of disease has laid the foundation. At the same time, because of the massive data of bioinformatics, the computing speed of computer and the algorithm of more optimal solution have put forward higher requirements, which also promote the development of computer application technology. I believe that under the mutual promotion and influence of life information and computer science, the future of life sciences will definitely develop better.

References

[1] MYang Bingru, Hu Jian, Song Wei. A Typical Application of Biological Information Data Mining Technology [J]. Computer Engineering and Applications, 2007,(02):18-19.

[2] Wang Manlin. Application of Computer in Biomedical Research [J]. Advances in Biochemistry and Biophysics, 1980.

[3] Season, Xu Yunjian, etc. [M]. of the World Airborne Radar Manual Beijing: aviation Industry Press, 1989.

[4] Zheng Guoqing. Advances and Prospect of Bioinformatics [J]. Henan Agricultural Science, 2003,1:4-8.

[5] Sun Xiao, Lu Zuhong, Xie Jianming. Bioinformatics Basis [M]. Tsinghua University Press, 2005.

[6] Zhang Yangde. Bioinformatics. M]., Science Press 2004.

[7] DanE.Krane&MichaelL.Raymer, Sun Xiao, Lu Zuhong, Xie Jianming. [M]. of Bioinformatics 2004.

[8] Shi Xiaooiu, Kong Fansheng. Application of Computer in Bioinformatics [J]. 5 2001.

[9] MarchiE, Zhong Nan. Genomics in Genome Planning [J]. and Chemistry of Life, 1999(1):9-10.

[10] Gao Fengling, Application of Computer Science in Bioinformatics [J] Gao Fengling
[11] Application of Computer Science in Bioinformatics [J] by Zhu Wenylong; Tang Lin; Weng Quan.

[12] Application of computer in bioinformatics [J] Wang Fan; Liu Shuai.