Investigation of data mining technique and artificial intelligence algorithm in microflora bioinformatics

Huixing Li^{1,2,a}, Yan Xue^{1,2} and Xiancai Zeng^{1,2}

¹Henan Key Laboratory of Industrial Microbial Resources and Fermentation Technology, Nanyang Institute of Technology, Nanyang, Henan 473004, China

²School of Biological and Chemical Engineering, Nanyang Institute of Technology, Nanyang, Henan 473004, China

Abstract. Bioinformatics has gradually received widespread attention and has shown the characteristics of a large amount of calculation and high complexity. Therefore, it is required to adopt computer algorithms in bioinformatics to improve the efficiency of bioinformatics processing problems. Big data and artificial intelligence technologies have the characteristics of supporting bioinformatics and have achieved certain results in the field of bioinformatics. Introduced the application basis of big data and artificial intelligence in bioinformatics, analyzed data collection, preprocessing, data storage and management, data analysis, and mining technology. Furthermore, typical applications in bioinformatics are discussed in terms of gene expression data analysis, genome sequence information analysis, biological sequence difference and similarity analysis, genetic data analysis, and protein structure and function prediction. Finally, the bottlenecks and challenges in the application of big data and artificial intelligence in bioinformatics are discussed, and the application prospects of related technologies in bioinformatics have prospected.

1 Introduction

Bioinformatics, as a new cross-discipline, emerged with the rapid development of life sciences and computer science. It effectively clarifies and analyzes a large amount of biological data information by making full use of tools or methods such as biology, informatics, mathematics, physics, statistics, and computer networks, so that it becomes biological data information with corresponding biological significance. With the development of sequencing technology, biological data has increased exponentially, and traditional analysis methods have been unable to meet the requirements of massive data analysis. It covers the acquisition, processing, distribution, storage, and other aspects of genomic information. Through the comparison and analysis of biological information, it is possible to obtain information such as gene coding and nucleic acid, and protein structure and function which is one of the most dynamic and promising subjects. However, bioinformatics is facing application bottlenecks due to the large number of samples to be processed and analyzed, and a large amount of calculation. Based on this, the application of big data and artificial intelligence in bioinformatics is explored to provide theoretical support for effective data mining in bioinformatics.

2 Big data technology in bioinformatics

Bioinformatics technology is mainly based on genomics and proteomics, through obtaining genomic information

and related data, and systematic use to solve major problems in the fields of biology, medicine, and industry. Among them, the amount of research data on genomics and proteomics is huge and has been growing. How to make full use of these data, decipher the genetic code, and dig out more relevant information to promote the development of science is a problem that bioinformatics workers urgently hope to solve.

The era of big data has spawned a variety of new technologies. Data mining, cloud computing, etc. have been widely used on the Internet and have achieved better results, mainly in terms of faster processing speed and stronger knowledge discovery capabilities. Big data technology mainly refers to the technology that quickly obtains valuable information from various types of data. Its processing technology includes big data collection, big data preprocessing, big data storage and management, big data analysis and mining, big data display and application, etc.

Biological information and the network have a similar architecture. For example, the research object is composed of base A, T, G, and C sequences, which is similar to the computer metadata 0 and 1 sequence data; the network environment of genomics and the hierarchical structure of computer networks Very similar. In theory, the two have a certain isomorphic basis. The emergence of massive data has given birth to a new scientific research model, that is, in the face of massive data, researchers only need to directly find or mine the required information from the data, without direct contact with the object to be studied, which brings about the research of biological information.

^a*Corresponding author's e-mail: Lihuixing@nyist.edu.cn

New ideas. The Internet of Things collects data through sensors, and the Internet collects data through web crawlers. Bioinformatics is generally obtained by genetic sequencing.

Data preprocessing refers to data cleaning and redundancy elimination. Data cleaning refers to the process of finding inaccurate, incomplete, or irrational data in a data set, and repairing or removing these data to improve the quality of the data. Redundancy elimination refers to data. This is a common problem with many data sets. However, there are generally no redundant problems in the field of bioinformatics. Many seemingly useless data will often unearth new knowledge. This is also one of the characteristics of bioinformatics data mining that is different from existing data mining. Data cleaning in biological information mainly includes base substitution, insertion, or missing errors in the gene-sequencing process, and the main error correction of high-throughput sequencing data.

Data analysis and mining objects include structured data, text, Web data, multimedia data, social network data, etc. The mining methods are mainly classified and clustered through machine learning. By digging the relationships between the data, discover knowledge that has not been discovered before. At present, data mining technology has been applied to gene chip analysis, DNA sequence comparison, biological literature mining, and biological data visualization. For biological information mining, the data structure of the mining object is different from the gene sequence structure, and it is difficult to directly use the current mining technology. For example, the relationship between the sequence regulatory region and the expression profile has always been a problem to be solved in the field of bioinformatics, but this kind of problem cannot be directly realized with the existing data mining technology.

Big data technology has made certain developments in the basic field of biological information and has achieved good results. The domestic "Tianhe No. 1" supercomputer has deployed petabytes of core computing to accelerate, making the calculation that originally required 12 hours to complete in just 1 minute can be completed, and the rapid analysis and identification of microbial metagenomics can be completed within a few hours. The application of various big data technologies has brought new opportunities to bioinformatics technology.

3 Artificial intelligence algorithms commonly used in bioinformatics

As an important branch of computer science, algorithm occupies a core position in computer science. In the information age, algorithms are one of the important tools for problem-solving. It can quickly obtain the required output in a short time by inputting compliant information. Now it has been widely used in various fields. In bioinformatics, the application of computer algorithms also actively promotes the development of bioinformatics. Computer algorithms commonly used in bioinformatics mainly include the following:

3.1 Divide and conquer

The divide-and-conquer method [1] is to solve a large problem instance by decomposing the problem instance into several small problem instances with the same problem, and then using a recursive method to solve these small problem instances in turn, and then merge the resulting solutions. To get the solution to the big problem instance. The divide-and-conquer method is mainly used in the fields of merge sort, nearest pair, and convex hull problems. In bioinformatics, the divide-and-conquer method can be used to analyze and process sequence alignment and sequence alignment.

3.2 Graph algorithm

Graph algorithm [2] refers to a convenient algorithm for obtaining solution examples of problems through special line calculation graphs. As a nonlinear structure, graphs are extremely complex. Therefore, graph algorithms are widely used in the fields of engineering, artificial intelligence, and mathematics, or the fields of bioinformatics and computer science. Among them, in bioinformatics, the use of graph algorithms can solve many bioinformatics problems, such as DNA sequencing, protein sequencing, and so on.

3.3 Greedy algorithm

Greedy algorithm [3] refers to a solution to a problem instance constructed by formulating a series of steps under a certain standard and selecting the local optimal one from many solutions. The selection is not reversible, so it is selected accordingly until the overall situation is optimal. In bioinformatics, the greedy algorithm is mainly used to solve the problems of genome rearrangement and reverse order. The application of this algorithm in bioinformatics can not only make the problem get the best solution but also has a high speed of operation, it is an effective and feasible computer algorithm.

3.4 Dynamic programming algorithm

The dynamic programming algorithm [4] refers to the decomposition of a large problem instance into some small, similar, and interleaved sub-problem instances, the optimal value is obtained through the recursion from bottom to top, and the solutions of the sub-problem instances are effectively stored an algorithm strategy to prevent repeated calculation of sub-problems and obtain the optimal solution to the problem. Applying dynamic programming algorithms to bioinformatics can effectively analyze and process the overlap and correlation characteristics between data, so it is mainly used in DNA sequence comparison, local and global sequence alignment, multiple alignments, and gene prediction and fill in the missing data and other issues.

Figure 1. Artificial intelligence and its algorithms.

4 Typical applications in bioinformatics

4.1. Gene expression data analysis

Gene expression data analysis has always been a hot and difficult point in bioinformatics research. In today's work practice, cluster analysis in computer algorithms is often used to analyze and process gene expression data. By grouping genes with similar expression patterns into one category, they can identify genes that are related to each other and analyze gene functions. Computer algorithms can use the transcriptional regulatory network of genes to observe changes in gene expression patterns with environmental changes or corresponding changes under the action of drugs, clarify the regulatory effects of genes with each other, and study the promoters of genes to analyze the same expression patterns and composition characteristics of similar promoters. Cluster analysis in computer algorithms is one of the important methods for analyzing gene expression data. It can not only discover the linear relationship between genes, but also find the nonlinear relationship between genes, so it has gradually been recognized by the majority of researchers.

4.2. Genomic sequence information analysis

The genome sequence in bioinformatics is not a simple arrangement of genes, but a specific organization and information structure, and the result of long-term evolution. It is one of the basic conditions necessary for genes to fully perform their due functions. The use of computer algorithms to analyze genome sequence information and predict relevant functional sites is one of the main research directions in recent years. The analysis of gene set sequence information usually adopts two categories: ab initio algorithm and comparison of the same source sequence method. Among them, the ab initio algorithm is a statistical method, which refers to effectively distinguishing the regions between exons, introns, and genes by identifying the nature and characteristics of protein-coding genes; The information is compared with the gene information in the database to find new genes. The new DNA sequence, in addition to the gene, generally contains a lot of other information related to the structural characteristics of the nucleic acid. This information has a decisive influence on the interaction between the DNA and the protein or RNA. The regions with similar protein and expressed sequence tags and encoding them are one of the most ideal algorithms for

analyzing genome sequence information in bioinformatics.

4.3. Biological sequence difference and similarity analysis

In bioinformatics, analyzing the differences and similarities of biological sequences is one of the most basic and important operations. Through the analysis and comparison of the differences and similarities of biological sequences, the structure, function, and evolution of biological sequences can be obtained in time. information. Generally speaking, structure, function, and biological sequence show a mutually restrictive relationship. The structure is determined by the biological sequence, and the function is determined by the structure. The use of computer algorithms in the analysis of the differences and similarities of biological sequences can quickly achieve the research objectives. Among them, one of the goals is to find similar structures and functions through the similarities between biological sequences. Of course, there are special cases. For example, biological sequences that have almost no similarities have the same spatial shape and function of the molecules. The second purpose is to compare the similarities between biological sequences and compare the similarities between biological sequences. Judge the origin, and infer the evolutionary relationship between biological sequences based on this. In the process of analyzing the difference and similarity of biological sequences, the commonly used computer algorithms are mainly particle swarm algorithms and support vector machine algorithm.

4.4. Analysis of genetic data

In bioinformatics research, due to the complexity of genetic structure, group sequence information, and biological sequence, it is required to use computer algorithms in the analysis of genetic data information. Specifically, some visualization tools can be used to express genes in the form of diagrams, trees, chains, and cubes, to improve the understanding of gene information and gene patterns by relevant staff. Knowledge discovery, as one of the most powerful visualization tools for discovering genetic data, can fully mine genetic data and can also have a positive impact on the level of transcriptional regulation of the genome.

5 Application challenges and future research directions

Nowadays, the driving force for research in the field of big data and artificial intelligence is mainly the economic benefits of enterprises. The huge economic benefits drive enterprises to continuously expand the scale of data processing. The economic benefits in bioinformatics are not easily created directly by big data, so it is slightly insufficient in terms of power. At the same time, in bioinformatics, due to its discipline, the computer talent reserve is relatively insufficient, and some mature theoretical methods in information technology have failed to match the biological information. Combination of

problems: Talents who master big data and artificial intelligence technologies are often concentrated in the IT industry, but computer researchers are often unable to keenly discover the essence of biological problems, and thus cannot make breakthroughs in the intersection of biology and computer technology.

With the development of bioinformatics, Chinese experts and scholars pay more and more attention to bioinformatics, and the popularization and application of big data and artificial intelligence also provide new opportunities for the development of bioinformatics to a certain extent. However, from the perspective of the overall development of bioinformatics, there is still a big gap with the international level, and the following two issues need to be paid special attention to in future research:

(1) Training of professional talents. As an emerging discipline, bioinformatics requires relevant practitioners to have both solid biological knowledge and high-level computer science skills. However, looking at the current status of bioinformatics practitioners in my country, there are serious talents. Faults and lack of talents have restricted the application of computer algorithms in bioinformatics. Therefore, it is required that in the later research process, pay attention to the training of professional talents, effectively solve the current situation of talent shortage, and provide strong talent support for the application of computer algorithms in bioinformatics.

(2) The scope of application is expanded. With the launch of the Human Genome Project and the improvement of computer science, the application of computer algorithms in bioinformatics has made preliminary progress, in the analysis of gene expression data, genome sequence information, biological sequence differences and similarities, genetic data, and It has played an important role in predicting protein structure and function. However, the content of bioinformatics is extremely rich. Therefore, it is required that in the later research process, the application range of computers in bioinformatics should be expanded in a planned way, so that the value of intelligent technology can be maximized, and it is necessary for the research of bioinformatics. Provide strong technical support for effective development.

6 Conclusion and future work

Bioinformatics, as an emerging discipline that combines biology and computer science, has biology at its core and computer science as its basic tool. The new thinking mode of Internet+ continues to impact traditional thinking modes and bring about changes in thinking. Therefore, bioinformatics-related researchers are required to strengthen communication and cooperation between various disciplines in their work practice, and fully grasp the application of computer algorithms in bioinformatics, to solve the problem of a large amount of information and a large amount of calculation in bioinformatics. Problems, and promote the further development of bioinformatics. The development of modern intelligent computer networks will promote the resolution of hot issues such as

gene sequence, protein sequence, protein structure and function, molecular interaction, signal pathways, and regulatory elements. Big data and artificial intelligence technology will promote the scientific research of bioinformatics.

Acknowledgments

This research work is supported by the Natural Science Foundation of Henan Province of China (182300410151); Scientific and Technological Project of Henan Province (162102210114); Key Scientific Research Project of College and University in Henan Province (17A610003); Henan Key Laboratory of Industrial Microbial Resources and Fermentation Technology Open Project (HIMFT20200206); Interdisciplinary Sciences Project, Nanyang Institute of Technology.

References

1. Nguyen, C., Tay, Y., & Zhang, L. . (2005). Divide-and-conquer approach for the exemplar breakpoint distance. *Bioinformatics*, 21(10), 2171-2176.
2. Lytynoja, A., Vilella, A. J., & Goldman, N. . (2012). Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics*, 28(13), 1684-1691.
3. Wang, Y. I., & Kuo-Bin, L. I. . (2005). Multiple sequence alignment using an exhaustive and greedy algorithm. *Journal of Bioinformatics & Computational Biology*, 3(02), 243-255.
4. Robert, G. . (0). A systematic approach to dynamic programming in bioinformatics. *Bioinformatics*(8), 665-677.