Efficient algorithm implementation for effective alignment of gene sequences

Yuren Liu

In the fulfilment of the requirement for the degree of

Master of Philosophy

Centre for excellence in Signal and Image Processing

Department of Electronic and Electrical Engineering

University of Strathclyde, Glasgow

Supervised by

Doctor Jinchang Ren

Professor Stephen Marshall

June 25, 2022

Declaration of Authorship

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Yuren Liu

June 25, 2022

Acknowledgements

As time goes by, my master's study life is coming to an end. In the beautiful city of Glasgow, I will never forget my study experience at the famous University of Strathclyde. Walking in the beautiful campus, watching the teachers and students of different colors walking by, I feel infinite emotion in my heart. In retrospect, I felt very confused in the face of the pressure of language, the discomfort in life, and the loneliness in a foreign country. Fortunately, I was very lucky to have my supervisors and colleagues, who helped me successfully carry out my work and finish my studies.

First of all, I would like to thank my First supervisor, Dr. Jinchang Ren, who is knowledgeable, gentle and considerate like my father, for providing me with valuable learning opportunities, providing me careful guidance and help in my studies, and providing me meticulous care in my life. I was touched by his rigorous academic attitude and integrity. I would like to express my heartfelt thanks and sincere respect. I'd also like to thank my second Supervisor, Prof. Stephen Marshall, for his help in my studies.

Thank you, Dr. Yijun Yan, who was in the same laboratory with me, for discussing with me in the process of research and experiment, which gives me a lot of inspiration and great help for the smooth completion of the thesis. Miss Xiaoquan Li, Mr. Guoliang Xie, Miss Ping Ma and other colleagues' care in daily study and life made me deeply feel the warmth of the big family. It's nice to meet you!

Thank my parents in particular for providing me with such a good learning opportunity. Your full love and selfless dedication are the eternal driving force and strong backing on my learning path.

Finally, I would like to express my heartfelt thanks and sincere blessing to my teachers, classmates, family and friends who have helped and cared for me on the way of growing up!

Abstract

With the completion of the Human Genome Project (HGP) and the vigorous development of the Model Organism Genome Project, more and more molecular sequence data have been generated. Scientific analysis, processing and research of these sequence data not only promote the development of bioinformatics research methods and technologies, but also have broad application background in the fields of prevention, diagnosis, treatment and new drug development of human diseases and major epidemic situations. How to give an effective graphic expression of gene sequence and analyze the similarity and evolutionary relationship of genes on this basis has become a hot topic in bioinformatics.

This dissertation mainly studies the graphical representation of DNA sequence, the similarity analysis of biological sequences and the algorithm for constructing the phylogenetic tree. The main achievements are summarized as below:

Firstly, the JZ-curve, a new graphical expression of the gene sequence, is introduced. By defining three mathematical mapping, a gene sequence can be transformed into three curves. It proves that the JZ-curve not only avoids the limitations associated with crossing and overlapping, but also retains the biological information of gene sequences.

Secondly, we construct a new characteristic matrix, named J/J matrix. When we study the sequence comparability based on graphical representation of DNA sequence, the J/J characteristic matrix based on JZ-curve can describe the chemical characteristic and the biological significance of gene sequences. The examination of similarities/dissimilarities among the coding sequences of the first exon of β-globin gene of different species illustrates the utility of the approach.

Thirdly, based on the JZ-curve, a fuzzy clustering algorithm on the basis of spectral graph theory for constructing phylogenetic tree is proposed. With the cluster analysis method, we build phylogenetic trees and determine the evolutionary relationship between the sequences. Meanwhile, the algorithm not only considers the divergence between classes, but also considers the similarity between classes, increase the accuracy of the results. The phylogenetic relationships for the coding sequences of the first exon of β-globin gene of 11 different species and the NA(H1N1) sequences of avian

influenza virus illustrate that algorithm is credible.

Keywords: DNA sequence; Graphic representation; Characteristic matrix; DNA sequence; Graphic representation; Characteristic matrix; Evolutionary tree construction algorithm

Content

List of Figure

List of Table

Chapter 1. Introduction

1.1 Research Background, Purpose and Significance

With the continuous development of various genome projects and the vigorous development of model organism genome projects, biological data are increasing exponentially. Scientific analysis and research on these biological data not only have broad application background in the fields of prevention, diagnosis, treatment and new drug development of human diseases and major epidemic situations, but also promote the rise and development of bioinformatics, an interdisciplinary discipline. Bioinformatics is a new discipline formed by the intersection of biology, computer science, information science and applied mathematics. With the advent of the post-genome era, the extraction and data analysis of biological information have become an important research direction in bioinformatics[1].

Data mining is a new scientific computing technology and data analysis method in recent years. It is a process of extracting hidden information and knowledge that people do not know in advance but are potentially useful from a large number of incomplete, noisy, fuzzy and random data.

It is difficult to say how valuable data mining of biological information in commerce is, but it has benefited mankind greatly. This not only helps us understand the essence and evolution of life, but also is of great significance to the discovery of new therapies and drugs. Therefore, how to apply many data mining technologies to biological data information analysis is currently a research hotspot, including the design of biological data mining architecture, the research of various analysis algorithms and the functional research on biological data mining analysis research[2,3].

With the growth of gene data, the traditional letter representation method of biological data can no longer meet people's requirements of simple, image and overall grasp of long sequences. Because graphics are simple and intuitive, more and more scholars adopt graphic representation to describe biological data. One of the advantages of graphical representation is its strong intuition. Compared with the alphabetic expression, graphic representation can better mobilize the participation of the human brain. In the fields of gene sequence similarity analysis, it promotes the generation of innovative research methods and opens up a new field of bioinformatics research in the

process of understanding and analysing gene sequences. However, graphic representation has some shortcomings such as loop and loss of biological information. Therefore, how to give an effective graphic representation method is a research hotspot in biology at present.

Sequence similarity refers to the degree to which one DNA or protein sequence is similar to another sequence. There are two purposes in studying sequence similarity. One is to obtain similar structures or similar functions through similar sequences. Another purpose is to judge the evolutionary relationship between sequences through the similarity of sequences. In the post-genome era, the structure and function information of the new sequence can be inferred by detecting the similarity relationship between the newly determined sequence and the sequence with known structure and function in the annotated database, By comparing the similarities and differences of the same conserved domains of different species, we can reveal the evolutionary relationship of species. By comparing the differences between normal sample sequences and pathogenic sample sequences, pathogenic genes can be found, and new drugs can be developed. Therefore, sequence similarity analysis has become a necessary step for gene function confirmation, and is an important means to study the longitudinal evolution direction of species and gene lateral transfer (LGT) . At the same time, it has a wide application background in the fields of understanding, treatment and new drug development of human diseases and major infectious viruses. Thus, sequence similarity analysis is a basic and very important research content in the field of bioinformatics. The accuracy of its analysis results plays a key role in the study of species classification, biological evolution, structure and function prediction, discovery of new drugs and new therapeutic schemes, etc. Which has a wide range of applications in bioinformatics, so sequence similarity is a topic of great research significance[5,6].

Phylogeny studies the evolutionary relationship between species. A phylogenetic inference is often expressed in the form of an evolutionary tree. Its purpose is to understand the laws of species metabolism, development, differentiation, and evolution. A reliable phylogenetic inference is important for understanding natural development, It is helpful to reveal the sequence of biological evolution by analysing the process of constructing phylogenetic trees. It is helpful to reveal the essence of evolutionary dynamics through the hidden phylogenetic relationship between species, to understand the history and evolution mechanism of biological evolution, and to study biological evolution tree is also very critical to solve many problems in modern molecular biology,

such as multi-sequence alignment, protein structure and function prediction, drug design, etc[3]. Therefore, the reconstruction of phylogenetic tree of species is an important issue in modern molecular evolution research.

1.2 Research Status and Development Trend

21st century is an era of life science and information. With the implementation of the Human Genome Project, the sequence and structure data of nucleic acids and proteins are increasing exponentially. Facing huge and complicated data, it is imperative to use computers to manage data, control errors and accelerate the analysis process. Since the end of the 1980s, bioinformatics has developed vigorously. Its research results not only play a great role in promoting related basic disciplines, but also have a great impact on industries such as medicine, hygiene, food and agriculture. Therefore, governments of various countries attach great importance to this and invest a lot of money to set up corresponding research institutions. European countries, the United States and Japan have successively set up biological information data centers, such as the US National Center of Biotechnology Information (NCBI), the National Center of Genome Resource (NCGR), the European Bioinformatics Institute (EBI), and the Japan's Center of Information Biology (CIB) etc. Among them, the European Molecular Biology Network (EMBNet), which is dominated by European countries, is currently the largest international research, development, and service organization for molecular biology information. Meanwhile, due to the development of electronic information technology, such as the information superhighway and the development of the Internet, the United States, Japan and European countries have successively established numerous bioinformatics network nodes on the Internet to provide large-scale database services. EMBNet (European Molecular Biology Network) has connected 22 national nodes and 8 large-scale biological computing centres, becoming the largest bioinformatics network[3,7]. More and more scholars pay attention to describing gene sequences by graphic expression because of its unique intuition [12]. Many models have emerged in the research of gene sequence graphical expression: E.Hamori and J.Ruskin first proposed G-curve[13] in 1983, which realized the graphical expression of gene sequence for the first time. Because G-curve is a 5D space representation, it does not have the advantage of graphical visualization. Gates, Nandy, Leong and Mogenthaler put forward 2D graph representation of DNA sequence based

on orthogonal coordinate system [14 ~ 16], which is the earliest 2D graph representation of gene and has the advantage of visualization, but these graphs may appear loops, so they cannot describe biological information comprehensively. Jeffrey proposed CGR graph representation method based on chaos theory in 1990 [17], which corresponds the sequence to a graph revealing its inherent fractal structure and has achieved good results in the fields of genome analysis. Academician Zhang Chunting put forward the Z-curve theory [3, 9, 17, 18], which initiated a new field of analyzing and studying DNA sequences by geometric methods. However, the defect of Z-curve is that there is a loop. In order to avoid loops, Liao Bo and others put forward a series of graphic representations [19 ~ 23]. Based on the 3D graphic representation of parameters introduced by Liao Bo and the Z-curve proposed by Academician Zhang Chunting, Z-curve proposed by Zhang Xizhen [24] solved the degradation phenomenon well and avoided the information loss caused by overlap and intersection in graphic representations. It can be predicted that the research on graphic representation of gene sequences would have a broad development space.

Sequence similarity analysis is one of the hot spots in the application of computer in biology. Gibbs in 1970 proposed the dot matrix method [25] is the earliest sequence similarity analysis algorithm. The basic idea is that when the same letters appear in two sequences at the same time, they are at the intersection point. Maizel then used the colored dot matrix method to compare the sequences of amino acids and nucleic acids. Needleman and Wunsch proposed a globally optimized sequence alignment algorithm in 1970 [26], and Smith and Waterman proposed a locally optimal sequence alignment algorithm in 1981 [27]. The idea of these two algorithms is that dynamic programming algorithm allows matching, mismatching and missing. However, the time and space complexity required are very high, which is not suitable for database search. A variety of heuristic algorithms have emerged, among which the most famous are FASTA algorithm proposed by Lipman and Pearson in 1981 [28] and BLAST algorithm proposed by Altschul in 1988 [29]. There are dozens of improved algorithms for these algorithms [30]. In 2000, Randi and others put forward the method of comparing biological sequences by matrix for the first time, which simplified the complex problems. The basic idea is to construct an appropriate matrix to represent a sequence, and then consider the invariants of these biological sequences, so that the comparison between sequences is transformed into the comparison between matrix invariants. The existing matrices include [31] E/E matrix, M/M matrix, L/L matrix (also called D/D matrix), high-

order matrix and compressed matrix, which have their own characteristics: the elements of E/E matrix represent the straight-line distance between points, but their disadvantages are that the element values increase with the increase of sequence length. When the sequence is too long, the matrix element values vary greatly, which is not conducive to later data processing; M/M Matrix is a symmetric matrix, M/M Matrix can effectively correct the M/M Matrix element difference is too large problem; L/L matrix is based on M/M matrix to normalize the element values, all element values less than 1; The high-order matrix is a 0,1 matrix, which requires a lot of calculation. The existing matrix invariants include the maximum eigenvalue λ, the sum of the largest (smallest) rows, the average of the sum of all elements on the secondary diagonal, the trace of the matrix and so on. In 2005, Li Chun provided ALE-index invariant [32], and proved that ALE-index invariant is close to the maximum eigenvalue λ of matrix numerically, but the calculation amount is obviously less than λ. In 2006, Zhang Yusen considered the Inv invariant factor proposed by row average [33], and proved that the calculation amount of Inv invariant is better than ALE-index of Li Chun. Z_lnv invariant [24] proposed by Zhang Xizhen is simple in calculation and closer to the maximum eigenvalue invariant of matrix than other invariants. Wang Tingsong directly used the curvature of graphics [34] as a new invariant to compare the similarity of creatures, which greatly reduced the computational complexity. Zheng Wenxin [35] calculated the center points (X, Y, Z) of the graph on the Z-curve to describe the curve distribution of the sequence. However, the existing methods of invariant characterization and comparison of biological sequences would be accompanied by the loss of some structural information, which realizes the similarity comparison of biological sequences. In 2006, Nandy et al. found that different matrix methods can get different results by comparing the above methods [36], and analyzed that the main reason is that these methods only consider the position of the constituent bases of the sequence, but do not consider the correlation between the model and the sequence and the biological significance such as base mutation or degradation. In addition, some people consider the evolutionary distance of organisms, semantic characteristics of codons, codon usage preference and other indicators to compare the similarity between organisms [37].

In 2006, Liao Bo et al. [23] used transitive closure method of fuzzy clustering to construct Evolutionary Trees in their 2D and 3D graphics. Liu Jingjun et al. [51] fused graph theory into fuzzy clustering, transformed fuzzy similarity matrix into a connected

graph with weights, and applied classical Kruskal algorithm to construct the minimum spanning tree, that is, evolutionary trees. Based on the 4D representation of DNA sequences, Li Gangcheng et al. [52] proposed a method that regards the similarity matrix of sequences as the fuzzy matrix of fuzzy clustering, and then uses the maximum tree method to construct biological Evolutionary Trees. This method is not accurate, but it avoids the disadvantage of high closure complexity in traditional fuzzy clustering calculation and has the advantages of simplicity and rapidity. Su Zhizhong [53] introduced a fuzzy clustering algorithm based on information dissimilarity to build a system tree. This algorithm calculates the frequency vectors of each molecular sequence as sample vectors, and classifies similar sample vectors by fuzzy clustering, which obtains the system tree of molecular sequences.

1.3 Contribution of the Thesis

Based on the background of bioinformatics, this thesis studies the 3D expression method of DNA sequence and the construction of phylogenetic tree. The contribution is as follows:

(1) Typical graphic expression methods of biological gene sequences, such as two-dimensional, three-dimensional and high-dimensional expression methods, are briefly introduced, and the advantages and disadvantages of different graphic expression methods are pointed out. In order to overcome degeneration and loop, a new 3D expression of DNA sequence is proposed in Chapter 3.

(2) Based on the similarity analysis of DNA sequences proposed in Chapter 3, four eigenvalues of the matrix are used to characterize the sequence characteristics and verified by the gene sequence experiment of the first exon of 11 common species in Chapter 4, which is a relatively new method of define and distinguish different DNA sequences. Also, a new algorithm description for constructing evolutionary tree is introduced based on the fuzzy similarity matrix. Through the clustering experiment of two virus databases, the advantage of the new cluster algorithm can be seen by the evolutionary tree and the labels of the database.

(3) The similarity analysis method and the construction method of phylogenetic tree are introduced in detail, and various methods of constructing phylogenetic tree are compared with those generated from the new 3D expression of DNA sequence in Chapter 5.

(1) A new graphical representation of DNA sequence- JZ curve group is proposed. By defining mathematical mapping, a gene sequence is transformed into three spatial curves, which proves that the curve not only overcomes the degeneration of patterns, but also retains the biological characteristics of the original DNA sequence.

(2) The feature matrix of similarity measurement between DNA sequences- J/J matrix is constructed. The J/J matrix combined with the JZ curve group not only describes the chemical properties of sequence bases, but also extracts the biological significance of gene sequences. Through the similarity analysis of the coding sequence of the first exon of the β globulin gene of 11 organisms, the experimental results show that the similarity of the DNA sequence can be simply and effectively analysed on the basis of the JZ graphic representation combined with the J/J matrix.

(3) A fuzzy clustering algorithm based on atlas theory is proposed to construct evolutionary tree. Clustering analysis of sequences is carried out to establish evolutionary trees and determine the evolutionary relationship between sequences. By constructing the phylogenetic tree of the coding sequence of the first exon of the b globulin gene of 11 organism and the NA gene sequence of H1N1 virus, the experimental results show the effectiveness of the algorithm.

1.4 Structure Arrangement of the Thesis

Based on the background of bioinformatics, this paper studies the graphic representation method of DNA sequences and the construction of evolutionary trees. The full text mainly includes 5 chapters, and the contents of each chapter are arranged as follows:

Chapter 1 Introduction. Chapter 1 mainly introduces the source, research background and significance of this thesis, the current situation and development trend at home and abroad, the research content and organizational structure of this thesis.

Chapter 2 provides the relevant literature review focusing on recent developments in the field of bioinformatics and different kinds of sequence alignment methods.

Chapter 3 is the similarity analysis based on graphical representation of DNA sequences. In this chapter, a new graphical representation method of DNA sequence, JZ curve set, is proposed, and the characteristics of the new graph are proved. Finally, the graphical curves of 11 species gene sequence are drawn with Matlab. Then, a new characteristic matrix- J/J matrix, for measuring similarity between DNA sequences is

constructed and applied to the coding sequence of the first exon of 11 organism for similarity analysis. Finally, experiments prove that the J/J matrix can simply and effectively analyse the similarity of DNA sequences.

Chapter 4 introduces graphic-based evolutionary tree construction method. Firstly, an algorithm of constructing evolutionary tree by fuzzy clustering based on graph theory is proposed. Then, by clustering the coding sequence of the first exon of the β globulin gene of 11 organism and the sequence of the NA gene of H1N1 virus, an evolutionary tree is constructed and the evolutionary relationship between the sequences is determined. Experimental results show that the algorithm is effective.

Chapter 5 gives some concluding remarks about the work in this thesis and detailed plans to further improve the introduced results in this thesis.

Chapter 2. Theoretical background and Literature Review

2.1 Brief of bioinformatics

The main task of bioinformatics is to analyse, process and study various biological information contained in DNA sequence data. Bioinformatics includes sequence comparison, protein structure comparison and prediction, gene recognition, molecular evolution and comparative genomics, sequence overlap group assembly, structure-based drug design and so on [8].

2.1.1 Nucleic acid and Protein

Nucleic acid is a one-dimensional polymer chain, which contains four monomers, each of which is called nucleotide. Nucleic acids carry genetic information, which is mainly expressed in the sequence of nucleotides. According to the different types of nucleotides, nucleic acids are divided DNA and ribonucleic acid (RNA). Nucleotides consist of phosphoric acid, deoxyribose or ribose and bases. The bases that make up nucleotides are divided into purine and pyrimidine. The former mainly refers to adenine (A) and guanine (G), both of which are contained in DNA and RNA. The latter mainly refers to cytosine (C), thymine (T) and uracil (U) [2]. Cytosine exists in DNA and RNA, thymine only exists in DNA, and uracil only exists in RNA. Among them, DNA is the main material basis for storing, replicating and transmitting genetic information, and RNA plays an important role in protein synthesis.

2.1.2 Variation

Variation refers to the alteration of some bases of DNA sequence in the course of biological evolution. Variations can be classified into three categories:

(1) Substitution: Substitution of one base in a sequence by another in the course of biological evolution.

(2) Insert or delete: Adding or deleting one or more bases in the course of biological evolution.

(3) Rearrangement: Some segments of a DNA or protein sequence undergo a change in the sequence of links during synthesis. For example, if the normal sequence

is ‘ATCGATCG’, then the sequence ‘ATGACTCG’ stands for a simple rearrangement to the original sequence.

Variation plays a very important role in the actual research process. Variation not only causes genetic variation and disease, but also species diversity.

2.2 Sequence Alignment

2.2.1 Overview of Sequence Alignment

In scientific research, comparison is one of the most common methods. In order to find the similarities and differences between objects or to discover the possible characteristics of objects, we usually use the method of comparison. In bioinformatics, comparisons are the alignment of multiple and similar sequences. Sequence alignment originated from the theory of evolution. If the two sequences are very similar, it can be inferred that the two sequences may have the same ancestors, which evolved from the compilation process of their ancestors through different gene substitution, addition, deletion and rearrangement. In addition, the structure and function of a given protein sequence can also be defined during sequence alignment, which is because when proteins are transcribed, each protein will be determined by its corresponding coder. Therefore, sequence alignment can be applied to secondary structure prediction, functional domain recognition of proteins and gene recognition.

```
A G C T T C G A C C A        A G C T T C G A C C A
| | | | | | | | | |        | | | | | | | | | |
A G C T T C G C C A        A G C T T C G - C C A
        (a)                          (b)
```

Figure 2.1 Matching before the insertion of space ‘-’ (a) and after the insertion of space ‘-’(b).

Sequence alignment is to use a specific mathematical model or algorithm to find out the maximum matching base number between sequences, that is, insert a space ‘-’ in two or more string sequences to achieve the maximum number of matched characters. For example, Figure 2.1 shows the sequence alignment of two sequences ‘AGCTTCGACCA’ and ‘AGCTTCGCCA’. Figure 2.1 (a) contains 8 same bases and Figure 2.1 (b) contains 10 same bases.

Compared with the method of not inserting spaces, it increases the number of matches. It can be seen from this that inserting vacancy is very necessary, and the process also reflects the process of biological evolution. The realization of sequence alignment generally depends on a mathematical model. Different mathematical models may reflect different characteristics of sequence structure, function, and evolutionary

relationship. It is difficult to judge whether a mathematical model is good or bad, or whether a mathematical model is right or wrong. It only reflects the biological characteristics of a sequence from a certain point of view.

2.2.2 Multi-sequence Alignment

While in multi-sequence alignment, we can use a quintuple to describe it. See Eq 2.1:

$$\mathrm{MSA} = (\Sigma, \mathrm{S}, \mathrm{A}, \mathrm{O}, \mathrm{F}) \tag{2.1}$$

where Σ represents a set of symbols for multiple sequence alignments with a value of $\Sigma 1 \cup \{-\}$; $\Sigma 1$ represents finite set of symbols, while in protein sequence alignment $\Sigma 1 = \{\mathrm{A}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}, \mathrm{G}, \mathrm{H}, \mathrm{I}, \mathrm{K}, \mathrm{L}, \mathrm{M}, \mathrm{N}, \mathrm{P}, \mathrm{Q}, \mathrm{R}, \mathrm{S}, \mathrm{T}, \mathrm{V}, \mathrm{W}, \mathrm{Y}\}$ (each element represents a different base forming the protein), and while in DNA sequence alignment, $\Sigma 1 = \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$ (each element represents a different base which forms a DNA sequence), - means a space which will be inserted during the process of alignment.

S means the sequence set to be aligned. The unmodulated sequence (which is the original and untranscribed sequence) of protein sequence alignment is composed of amino acids. DNA sequence alignment is that each sequence is composed of bases and the sequence length is different. $S = \{S_i | \mathrm{i} = 1,2, \ldots, \mathrm{m}\}$,

$S_i = (C_{ij} | j = 1,2, \ldots, l_i)$, where m equals to the number of sequences, C_{ij} is the j^{th} base in sequence S_i, l_i means the length of the i^{th} sequence.

A represents the result matrix, $A = (a_{ij})_{m \times n}, a_{ij} \in \Sigma$. In the result matrix, line I represents the i^{th} sequence, and each j lists of the matrix represents the result of the comparison of the j^{th} base. The base sequence in the sequence cannot be changed before and after alignment.

O is a set of comparison operations, $O = \{\mathrm{insert\_space},\ \mathrm{delete\_space}\}$, which is the operation of insertion and deletion of the gap '-'.

F is the algorithm of alignment in order to figure out the specific position of the insertion and deletion.

2.3 Vacancy Penalty

In the process of sequence alignment, in order to make the results of sequence alignment more in line with certain expectations, the insertion or deletion of sequences is compensated by introducing vacancies. However, we should not introduce vacancies

indefinitely, otherwise the results will lack biological significance. In order to limit the insertion of spaces, the usual method is to deduct the total score by inserting spaces. The deduction score is a penalty score, which restricts the insertion of vacancies into the penalty score of vacancies. Therefore, when obtaining sequence alignment scores, the total score of matching between two sequences should take space penalty scores into consideration [9]. And this kind of alignment method can obtain better alignment results between two sequences taking biological significance into consideration.

Suppose S_1 and S_2 are used to represent the sequence to be aligned, S_{10} and S_{20} are used to represent the result of alignment, and L is used to represent the length of alignment. Generally, there are three kinds of space penalty rules.

2.3.1 Vacancy Penalty

The simplest penalty rule is the vacancy penalty score. When a vacancy is inserted into a sequence, a fixed penalty score Wg is given. For the whole sequence, the total blank penalty score is equal to the inserted blank Rg multiplied by the penalty score Wg for each vacancy [10].

Vacancy penalty points do not add extra running time, so it is the simplest penalty rules. However, in biology, for practical reasons, scores need to be calculated differently. On the one hand, the partial similarity between two sequences is a common phenomenon; On the other hand, a single gene mutation event may lead to the insertion of a single long gap. Therefore, the connection gap with long gap is usually more favoured than multiple scattered short gaps. In order to take this difference into account, the concepts of gap opening, and gap expansion are added to the scoring system. The gap opening score is usually higher than the gap expansion score.

2.3.2 Constant Vacancy Penalty

This penalty rule is not for every space, but for every vacancy. Here the connected space is called a vacancy. The penalty score is based on the vacancy inserted in the sequence, and the penalty score of every inserting vacancy is Wg [10]. The specific operation is as follows:

Assuming whether matched or not, the score value is expressed by σ, we have $\forall \mathrm{x}, \sigma(\mathrm{x}, -) = \sigma(-, x) = 0$;

The representation of alignment score becomes:

$$\sum_{i=1}^{L} \sigma(S_{10}[i], S_{20}[i]) + W_g \times gaps \quad (2.2)$$

where *gaps* represents the number of vacancy and i stands for the position of the base on this DNA sequence.

Constant space penalty can avoid the defect of space length penalty, but when too many connected spaces are inserted, this penalty rule cannot be limited, which may lead to the splitting of sequence segments by connected spaces. This requires a penalty score rule which is closely related to the length of the space. Its penalty score does not only depend on the length of the space, but also does not ignore the length of the space.

2.3.3 Affine Vacancy Penalty

The rule divides the penalty of vacancy into two parts [10]: open vacancy penalty and extended vacancy penalty. If q is the length of the vacancy, W_g is the penalty score of the open vacancy, W_s is the penalty score of the extended vacancy and W is the total penalty score, then: $\mathrm{W} = W_g + \mathrm{q} \times W_s$, so we can have the formula calculating the alignment score:

$$\sum_{i=1}^{L} \sigma(S_{10}[i], S_{20}[i]) + W_g \times gaps + W_s \times spaces \quad (2.3)$$

where *gaps* is the number of space and spaces represents the number of vacancy.

In practical biology research, the probability of inserting and deleting multiple vacancies connected by a vacancy coin is small, so the affine vacancy penalty score has more biological significance.

2.3.4 Scoring Matrix

In sequence alignment, a matrix is usually used to record the score of each variation, which becomes the score matrix. The results of sequence alignment will be different if different scoring matrices are chosen. The simplest scoring matrix is a single matrix, also known as a sparse matrix. Using this matrix, we only need to detect whether the bases of corresponding sites between sequences are identical, and the scores of the same bases are 1, and the differences are 0. This matrix, which only considers the identity of bases, has great limitations.

In order to better reflect the biological characteristics, we need to design a more optimized scoring matrix. PAM (point accepted mutation) [11] is the first widely used optimal matrix. A PAM indicates that 1% of the amino acids have changed, that is, the evolutionary unit of variation. Generally, sequences with high similarity use lower PAM

matrix, while sequences with low similarity use higher PAM matrix. For this matrix, we first need to observe the replacement event and get an observable or acceptable point mutation matrix A. A_{ij}means the number of amino acid i being replaced by amino acid j. The mutation probability matrix M can be further obtained from A. M_{ij} means the empirical frequency of amino acid i being replaced by amino acid j. Taking the percentage of observable mutations, PAM, as a unit of time measurement, 1PAM represents the time required for an expected replacement in 100 amino acid polypeptide chains. The PAM250 matrix obtained by processing the mutation probability matrix M to the power of 250 is a suitable time unit for studying the evolutionary relationship between distant proteins. PAM_n means that a polypeptide sequence containing 100 amino acids has undergone a period of evolution, during which n amino acid substitution events have occurred.

Besides PAM matrix, BLOSUM (blocks substitution matrix) [11] matrix is also widely used. BLOSUM matrices also use numbering to distinguish different BLOSUM matrices, where numbering is mainly used to distinguish the similarity of sequences. For example, BLOSUM62 matrix is generally used to align at least 62% of the same proportion of sequences. So, the use of BLOSUM matrix is exactly the opposite to that of PAM matrix. For BLOSUM matrix, we have f_{ij} which is the number of amino acid i and j pairs in sequence alignment. Then we have proportion or frequency of certain amino acid pairs $q_{ij} = f_{ij}/\sum_{i,j} f_{ij}$, and the expected frequency of occurrence of each amino acid $p_i = p_{ii} + \frac{1}{2}\sum_{i \neq j} q_{ij}$, the reason for the emergence of $\frac{1}{2}$ is that when two protein sequences have ij pairing with each other, for a specific sequence, the probability of exactly getting i is only half. Then calculate the expected frequency of amino acid pairs $e_{ij} = \begin{cases} p_i^2, i = j \\ 2p_i p_j, i \neq j \end{cases}$, 2 appears because for any ij pair, there are two cases, i is divided into a sequence or i is divided into b sequence. Now the BLOSUM matrix element is defined as $s_{ij} = 2\log_2 \frac{q_{ij}}{e_{ij}}$, which is the ratio between the frequency of occurrence and the frequency of expectation.

2.4 Classic Alignment Algorithms

At present, many sequence alignment algorithms are based on dynamic programming algorithm, considering different improvements in computing speed and

storage space. Sequence alignment algorithms have several different classification methods. According to the number of alignment sequences, sequence alignment can be divided into double sequence alignment and multiple sequence alignment. According to the range of sequence alignment, sequence alignment can be divided into global sequence alignment and local sequence alignment.

2.4.1 Global Sequence Alignment and Local Sequence Alignment

Local sequence alignment considers the local similarity of sequences, which is a forehead method to find partial similarity regions of sequences. Local sequence alignment is mainly applied to protein sequence alignment, which is more sensitive and biologically significant than complete sequence alignment. Global sequence alignment is the whole sequence, which considers the similarity of sequences from the global scope. Global sequence alignment is mainly used to predict the homology between sequences and the structure and function of proteins.

2.4.2 Double-sequence Alignment

Double sequence alignment is to find the maximum similarity match between two DNA or protein sequences. The search process is based on some algorithm or model. Multiple sequence alignment and sequence database search are based on double sequence alignment. At present, the most classical double sequence alignment algorithms are lattice graph method and dynamic programming algorithm.

2.4.2.1 Lattice Graph Method

The simplest double sequence alignment algorithm is the lattice graph method. In this method, the sequence to be aligned is placed on a two-dimensional plane, a sequence is placed horizontally on the top of the plane, and a sequence is placed vertically on the left of the plane. A point is marked at the intersection of any two identical bases of the two sequences. Finally, linking the points parallel to the diagonal line constitutes the result of two sequence alignments [12].

The lattice graph method can visually display the insertion and deletion of sequences, and all matched base sequences between two sequences can be visually reflected by lattice graph. However, since the sequence length is counted in thousands, it is unrealistic to use all the lattice computing programs to calculate the real alignment

sequence because when calculating the similarity of a large dataset of gene sequences in which each sequence is several Gb at least, every two sequences are aligned there will be a huge matrix generated, considering more sequences to be aligned together we will face infinite need for calculation capacity and the internal memory of the computer, so other alignment methods will be more used to achieve.

2.4.2.2 Dynamic Programming Algorithm

Dynamic programming algorithm was first proposed by Needleman and Wunsch and has been widely used and improved. It has gradually become one of the most important theoretical foundations in computational biology. The most classical dynamic programming algorithms are Needleman-Wunsch algorithm (NW algorithm) [13] and Smith-Waterman algorithm(SW algorithm) [14]. All global alignment algorithms are based on NW algorithm, while SW algorithm is improved based on NW algorithm, mainly applied to local sequence alignment. The following is a brief introduction to the dynamic programming algorithm.

Given sequence s_1 and s_2, the length of which are m and n correspondingly. $s_1[1 \dots i]$ and $s_2[1 \dots j](1 \leq i \leq m, 1 \leq j \leq n)$ represent prefix subsequences separately of s_1 and s_2. And the alignment result of s_1 and s_2 contains that of $s_1[1 \dots i]$ and $s_2[1 \dots j]$, which is a recursive relationship.

From this relationship, it can be seen that the optimal solution to its subsequence is the premise of solving the global alignment of two sequences. Through this recursive relation, the optimal value of the whole sequence can be obtained. Then, the optimal alignment result of the sequence is obtained by backtracking the path of the obtained optimal value.

The basic step of dynamic programming algorithm is to use a binary matrix to store the similar scores of two sequences, and then retrieve the optimal alignment of the sequences according to the scores in the matrix. Assume that the sequence s_1 and s_2 are compared by using dynamic programming algorithm, their lengths are m and n, respectively. Firstly, we need to construct a two-dimensional matrix with the size of $(m+1) \times (n+1)$. The element $M[i,j](0 \leq i \leq m, 0 \leq j \leq n)$ in the matrix represents the highest alignment score of its prefix subsequence $s_1[1 \dots i]$ and $s_2[1 \dots j]$. The ratio of prefix subsequence a to vacancy '-' is expressed in both row 0 and column 0 of the matrix. Therefore, the initial values of the elements in the matrix are:

$$M[0,0] = 0 \tag{2.4}$$

$$M[i,0] = \sum_{k=1}^{i} \sigma(s_1[i], -) \ (1 \le i \le m) \tag{2.5}$$

$$M[0,j] = \sum_{k=1}^{j} \sigma(-, s_2[j]) \ (1 \le i \le n) \tag{2.6}$$

By analyzing the prefix subsequence $s_1[1 \dots i]$ and $s_2[1 \dots j]$, there may be three cases to get the optimal score $M[i,j]$:

(1) The sum of the alignment score between $s_1[i]$ and $s_2[j]$ and the score between the subsequence $s_1[1 \dots i-1]$ and $s_2[1 \dots j-1]$ which is $M[i-1, j-1]$;

(2) The sum of the alignment score between $s_1[i]$ and a space '-' and the score between the subsequence $s_1[1 \dots i-1]$ and $s_2[1 \dots j]$ which is $M[i-1, j]$;

(3) The sum of the alignment score between $s_2[j]$ and a space '-' and the score between the subsequence $s_1[1 \dots i]$ and $s_2[1 \dots j-1]$ which is $M[i, j-1]$. And it's shown below (Figure 2.1):

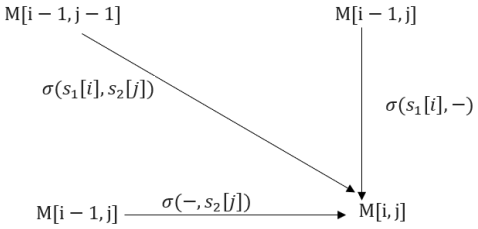

Figure 2.2 Source of Matrix Elements

Thus, the recursive relation is obtained as:

$$M[i,j] = max \begin{cases} M[i-1, j-1] + \sigma(s_1[i], s_2[j]) \\ M[i-1, j] + \sigma(s_1[i], -) \\ M[i, j-1] + \sigma(-, s_2[j]) \end{cases} \quad (1 \le i \le m, 1 \le j \le n) \tag{2.7}$$

Using the upper formula, the values of each element in the matrix are calculated in order from left to right and from top to bottom. When the position of the last column in the last row, element $\mathrm{M[m, n]}$, is calculated, the best alignment score of two sequences s1 and s2 is obtained.

After obtaining the optimal score, we use the backtracking method to construct the comparison record. That is, starting from the position of the optimal alignment score, i.e. the position of the last column in the last row of the matrix, we retrospect along the path to get the value until reaching column 0 of row 0. At this point, the sequence corresponding to the intersection points in the backtracking process is the alignment result.

When using dynamic programming algorithm for sequence alignment, it is necessary to calculate each element in a two-dimensional array of $(m+1) \times (n+1)$ size. Its time complexity is *O(mn),* and space complexity is *O(mn)*. The backtracking process is from the lower right of the array to the upper left of the array, passing through *(m+n)* elements, so its time complexity is *O(m+n)*, and there is no additional storage overhead. Therefore, the basic dynamic programming algorithm takes a lot of time and space to solve the problem of double sequence alignment, so people put forward various improved algorithms.

2.4.2.3 Multi-sequence Alignment

The generalization of double sequence alignment in multi-sequence alignment is to extend the alignment problem from two sequences to multiple sequences. Therefore, the method to solve the double sequence problem is also applicable to multiple sequence alignment, but the number of alignments has increased, which makes the problem more complex. Murata has successfully applied dynamic programming algorithm to the alignment of three sequences [15], but because the alignment time is long and the space required is large, it is almost impossible to extend it to more than three sequences alignment. Generally, the dynamic programming algorithm is seldom used in multi-sequence alignment. Basically, heuristic algorithm, random algorithm and partition algorithm are used. Among them, there are many kinds of heuristic algorithms, such as star alignment algorithm, progressive alignment algorithm, iterative thinning method and so on.

1. Star Alignment Algorithm

Star alignment algorithm is a fast heuristic method for solving multiple sequence alignment problems. It needs to find a central sequence, and the result of alignment is established by comparing the central sequence with other sequences. Star alignment algorithm follows a rule, that is, in the alignment process, the space must be added to the central sequence continuously so that the central sequence and alignment sequence can reach the maximum number of matches. Spaces added to the central sequence cannot be removed, and always remain in the central sequence, knowing that the central sequence and the ordered sequence are aligned.

2. Progressive Alignment Algorithm

Another simple and effective heuristic algorithm is the progressive alignment

algorithm. The basic idea of progressive alignment algorithm is to use dynamic programming algorithm to iteratively align two sequences. That is to say, the two sequences are aligned first, and then the new sequences are added, until all the sequences are added. However, different addition order may lead to different comparison results. Therefore, the key of progressive alignment algorithm is how to determine the sequence of alignment. Generally, the alignment begins with the two most similar sequences, and then proceeds to the far alignment to complete all the sequences.

Progressive alignment algorithm mainly consists of three steps:

(1) Computation of the distance matrix;

(2) Construction of guidance tree;

(3) Alignment of the sequences according to the constructed guide tree.

Nowadays, the most widely used progressive comparison procedure is ClustalW. It gives a set of schemes of dynamic selection of comparison parameters, which mainly solves the problem of parameter selection in the process of comparison. Usually, the scoring matrix and reflective blank penalty score are used to solve the selection problem of comparison parameters, and it is hoped that the effective parameters can be set to achieve the desired results.

2.4.3 Graphic representation of gene sequence

Because the original sequence of organisms is a string form represented by four bases (A, G, T, C), it is relatively difficult to find information directly from the original sequence itself, so we can observe the biological sequence more intuitively by using graphics to represent the original sequence of organisms. Their basic idea of using graphic representation is: first, the sequence is transformed into a graphical representation, and then a matrix is constructed according to the graphical representation, and the similarity of biological sequences is analyzed by using the invariants related to the matrix (for example, maximum eigenvalue, row sum, trace, average value of elements, etc.). Here we will introduce several typical graphical representations of DNA sequences.

2.4.3.1 G-curve and H-curve

In 1983, E. Hamori and J. Ruskin first expressed DNA sequences as spatial curves -- g-curves and H-curves. G-curve is a kind of 5-Dimensional space representation, in

which four coordinate directions are four nucleotides respectively, and the remaining one represents the position of nucleotide in DNA sequence, but this method cannot realize visualization[85]. The four directions of the two axes are used to represent the four nucleotides (A →NW (northwest); G →SE (southwest); C →NE (northeast); T → SW (southwest)), and the other direction is the position of nucleotides. At this time, the curve becomes a three-dimensional space curve, which is the H-Curve.

2.4.3.2 CGR Graph

CGR (chaos game representation) proposed by Jeffrey in 1990 is a combination of graphic and mathematical expression of DNA sequences. This method is based on Chaos Theory and corresponds the sequence to a graph that reveals its inherent fractal structure. Different DNA sequences show different shapes (such as mountains, clouds, corals, etc.) in the graph. It has been applied geometrically and has shown good results in dealing with genome analysis problems[84].

2.4.3.3 Graphic representation of two-dimensional coordinate axes

In 1986, M.A..Gates proposed the earliest two-dimensional graphic representation [86]: for the next gene locus, if the base on this locus is C, it can be viewed as the curve extend at the direction of +X-axis by one unit. And for the base G, the curve extend at the direction of -X-axis. And similarly, + Y-axis unit vector represents T, and -Y-axis unit vector represents A. And the coordinates can be illustrated in Figure 2.3.

In 1994, A. Nandy gave a two-dimensional graphic representation as [17]: + X-axis unit vector represents base G, -X-axis unit vector represents base A, +Y-axis unit vector represents base C, and -Y-axis unit vector represents base T.

In 1995, P.M.Leong and S. Morgenthaler proposed another two-dimensional graphic representation as [18]: + X-axis unit vector represents base A, -X-axis unit vector represents base C, +Y-axis unit vector represents base T, and -Y-axis unit vector represents base G.

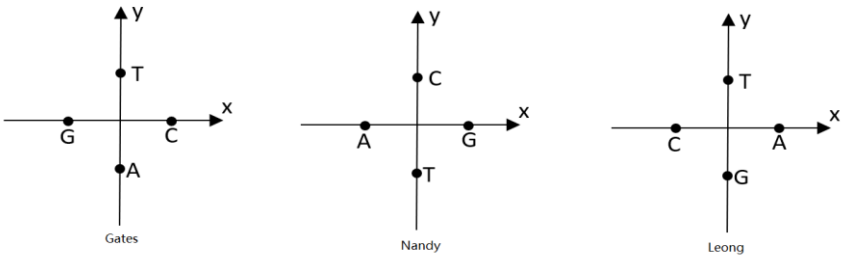

Figure 2.3 Three coordinates of two dimensional graphics

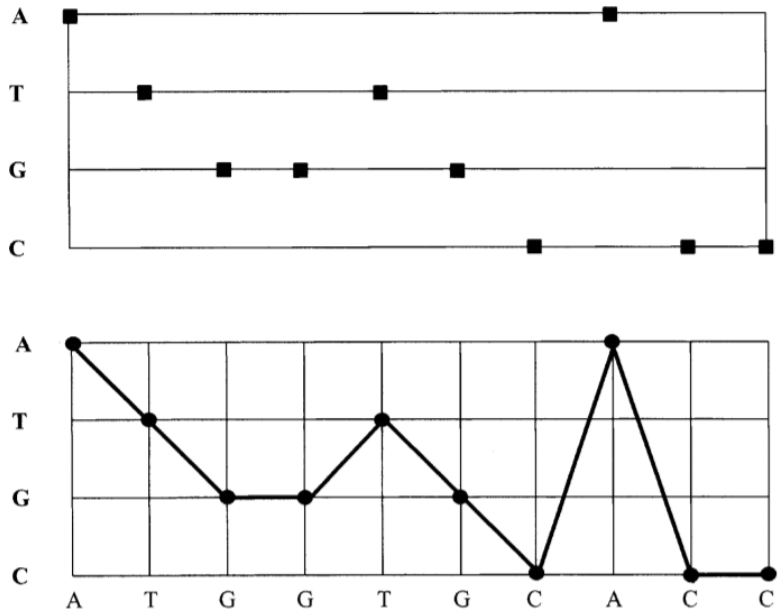

Figure 2.4 Graphical representation of a simple sequence using the 2D coordinate axe. The rectangles (dots) denote the bases making up the sequence.

Each of the three representations takes the origin as the initial point and adds a unit vector according to the given direction for each additional base as we can see in Figure 2.3 and 2.4. For example, the graphs of GC, GCG, GCGC and GGCGC will be difficult to distinguish under gates representation, while those of GA, GAG, GAGA and GGAGA under Nandy's representation will be difficult to distinguish.

In order to avoid the degradation of the above figure, X.F.Guo and Y.C.Liu[88] improved the above coordinate vector so that the four vectors A, G, C and T offset the X and Y axes of the two-dimensional Cartesian coordinate system. Randic expressed the four nucleotides in binary numbers: A = 00, G = 01, C = 10, T = 11, so as to obtain a DNA map [61]. Liao Bo also proposed several two-dimensional graphic representations [21], which well solved the degradation problem.

2.4.3.4 Z-curve

Z-curve theory is a geometric representation method of DNA sequence proposed by academician Zhang Chunting. It is an intuitive tool for displaying and analyzing DNA sequence. A given DNA sequence corresponds to a curve in three-dimensional space. The proposal of Z-curve opens up a new field for analyzing and studying DNA sequences by geometric methods [19].

Consider a single stranded DNA sequence of length n whose Z-curve contains a series of points, $P_0\ P_1\ P_2\ \ldots P_N$ and the corresponding coordinates (x_n, y_n, z_n) n = 0,1,2,…N,

$$\begin{aligned} x_n &= (A_n + G_n) - (C_n + T_n) \\ y_n &= (A_n + C_n) - (G_n + T_n) \qquad x_n, y_n, z_n \in [-N, N],\ n=0,1,2\ldots, N \qquad (2.8) \\ z_n &= (A_n + T_n) - (G_n + C_n) \end{aligned}$$

The number of times that the four bases appear in the subsequence from 1 to n are represented by A_n, G_n, C_n and T_n, respectively.

The three components of Z curve have clear biological significance:

(1) x_n represents the distribution of purine (A + G) / pyrimidine (C + T) bases along the sequence. When the number of purine base is more than pyrimidine base, $x_n > 0$; otherwise, $x_n < 0$; when both are equal, $x_n = 0$.

(2) y_n denotes the distribution of amino (A + C) / keto (G + T) bases along the sequence.

When the number of amino base is more than that of keto bases, $y_n > 0$; otherwise $y_n < 0$; when both are equal, $y_n = 0$.

(3) z_n denotes the distribution of strong hydrogen bond (A + T) / weak hydrogen bond (G + C) bases along the sequence. When the number of weak hydrogen bonding bases is more than that of strong hydrogen bonding bases, $z_n > 0$; when the number of weak hydrogen bonding bases is less than that of strong hydrogen bonding bases, $z_n < 0$; when the number of weak hydrogen bonding bases is equal to the strong hydrogen bonding bases, $z_n = 0$.

2.4.3.5 Other 3D-curves

Randic assigned the four bases to vector coordinates: (+1,-1,-1)→A, (-1,+1,-1)→

G, (-1,-1,+1)→C, (+1,+1,+1)→T, thus proposed a three-dimensional graphic representation of DNA sequence and protein sequence [61]. Li Chun gave a three-dimensional graphic representation [22], whose vector representation is: (1,0,0)→A, (0,1,0)→C, (0,0,1)→G, (1,1,1)→T. Liao Bo et al. assigned different vector coordinates to the four bases, and proposed a series of three-dimensional graphic representations[23-24], which generate the same type of curves following similar formula with only different parameters on X,Y and Z axis as formula 2.8.

And all the 2-D and 3-D curves contribute to the new 3-D graphical representation of DNA sequence in this thesis. And the 3-D dimensional alignment method will be introduced briefly in the next chapter.

2.4.4 Matrix invariant analysis of Graphs

2.4.4.1 Characteristic matrix of graph

The graphic representation of DNA sequence allows us to observe the sequence more intuitively, and the information of DNA sequence is represented by matrix, which makes it easier for us to extract sequence information. This work was first proposed by Randic [61]: Here we take a two-dimensional graph as an example, and the coordinates of the i_{th} and j_{th} point are (x_i, y_i) and (x_j, y_j).

E matrix is the matrix whose element is the Euclidean distance of the corresponding points of two bases on the curve.

$$E_{(i,j)} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \tag{2.9}$$

And the element of M/M matrix is the ratio of the Euclidean distance of the corresponding points of two bases on the curve to the unit line segment existing between them.

$$m_{(i,j)} = \frac{\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}}{|j - i|} \tag{2.10}$$

L/L matrix, also called D/D matrix, its element is the ratio of the Euclidean distance of the corresponding points of two bases on the curve and the graph theoretical distance between them (the sum of the length of line segments between two points on the curve). The main diagonal element of L/L matrix is 0, and all elements are less than or equal to 1.

$$l(i,j) = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{\sum_{k=1}^{j-1} e_{xy}(k,k+1)} \quad (2.11)$$

Also we have high order matrix of the M/M(L/L) matrix- ${}^{k}L/{}^{k}L$ is the matrix obtained by taking the power of k for each element of L/L matrix; ${}^{b}L/{}^{b}L$ matrix is a matrix filled by 0 and 1, which is obtained by taking the limit of each element in ${}^{k}L/{}^{k}L$ matrix when k tends to be positive infinity.

2.4.4.2 Matrix invariant

(1) The largest eigenvalue λ: this invariant has great advantages, that is, it can better reflect the information of the sequence, but it also has shortcomings, that is, with the increase of the length of the sequence, the calculation of the matrix eigenvalue will increase, so we seek new parameters to replace it in this article.

(2) Average sum of row $\frac{1}{n}\sum_j a_{i,j}$.

Using the invariants of matrix, we can compare the similarity of DNA sequences. Firstly, the corresponding matrices and matrix invariants are obtained for several DNA sequences to be compared, such as the maximum eigenvalue, determinant value, trace of matrix, maximum row sum, etc. Then some invariants are used as indexes to compare the sequences.

If the graphic representation of the same sequence must be completed by several graphs, each graphic representation extracts a matrix corresponding to a matrix invariant, and the obtained k feature invariants constitute a k-dimensional vector. Let the k-dimensional vector corresponding to two sequences a and b be: $U^a = (u_1^a, u_2^a, \dots, u_k^a)$, $U^b = (u_1^b, u_2^b, \dots, u_k^b)$. Generally, the closer the two vectors point, the more similar the two sequences are. Therefore, we have two methods to calculate: (1) $d(U^a, U^b)$, that is, the Euclidean distance between the end points of two vectors. The smaller $d(U^a, U^b)$, is, the more similar the two sequences are. (2) $\theta(U^a, U^b)$, which is the correlation angle of two vectors. If the correlation angle between two vectors is smaller, the two sequences are considered to be more similar.

The advantage of using matrix invariants to describe and compare biological sequences is that the description and comparison of invariants are very simple. The comparison of two biological sequences (character sequences) is transformed into the comparison of mathematical description sequences corresponding to biological

sequences. However, there are some structural information loss when the existing methods describe and compare biological sequences with invariants. Therefore, how to find a biological graphic representation and find better invariants to describe and compare biological sequences is a subject worthy of further study.

2.5 Conclusion

In this chapter, biological data mining and sequence graphic representation methods are reviewed. In bioinformatics, the most basic operation is the operation of DNA sequence. This chapter first introduces the concept and research purpose of bioinformatics, and then discusses the basic steps and main processes of database technology and biodata mining in bioinformatics processing. Finally, the graphic expression of several gene sequences and the matrix invariant analysis methods based on the graphic expression are summarized systematically.

This chapter mainly introduces the graphical representation of DNA sequence like Nandy's (AG-CT) graph, Gates' (AT-GC) graph and Leong and mogenthaler's (AC-GT) graph representation, and introduces people from the original DNA sequence represented by string into the DNA sequence represented by graph, which also points out a path for future generations to study DNA sequence. Although it is intuitive and convenient to compare and explain the similarity of species, there may be many intersections or overlaps in the graph, that is, in the sense of graph theory, the emergence of so-called multilateral or circle, the most direct impact is to make a DNA sequence uniquely correspond to a graph, but a graph does not uniquely correspond to a DNA sequence, In this way, the results of graphic analysis are somewhat inaccurate. Then, in the 3D graph proposed by Zhang(Z-curve), he has further improved on the basis of the graph mentioned above, that is, the direction given by the four bases slightly deviates from the original coordinate axis, which can greatly reduce the phenomenon of crossover or overlap. From the initial 2D to the 3D, 4D and even higher dimensional representation proposed by Liao and Chi.

In a word, there are many kinds of graphical representations of DNA sequences. Here are just a few typical methods. These methods are worthy of our careful consideration and thinking about how to improve them, and inspire us to propose some more accurate graphical representations as far as possible.

Chapter 3.Similarity Analysis Based on Graphical Representation of DNA Sequences

3.1 Introduction

With the growth of gene sequence, the graphic expression method of gene sequence has become an important means to study gene sequence. The intuitive description of gene sequence by graphic expression method is more conducive to the classification of genes and the analysis of gene evolution relationship. At present, researchers have given two-dimensional, three-dimensional or four-dimensional or even higher-dimensional graphical representations of gene sequences[16-18]. Among them, Academician Zhang Chunting's gene sequence research based on geometric methods has made a series of progress in the research of SARS sequences[19,20]. In addition, Nandy, Randic et al. [17,38,60-62]also proposed a coordinate axis-based 2D for DNA sequence, 3D graphic expression, on the basis of Nandy, Randic and others, Liao Bo has given a variety of graphic expression methods[21-25]. These expression methods have overcome the unfavourable factors existing in graphic expression. That is, the loss of gene information or the generation of loops in the expression of graphs. At the same time, the concept of geometric center is introduced, and various similarity matrices are constructed for similarity comparison and analysis under this kind of non-degenerate graphic expression. In this chapter, based on Academician Zhang Chunting's Z-curve and Liao Bo's graphic expression methods with zone parameters[21-25], a new graphic expression method of gene sequence-JZ-curve group is proposed. It is proved that the graphic curve not only has good non-degenerative characteristics, but also fully retains the biological characteristics of Z-curve. At the same time, this chapter takes sequence similarity as the research object, and constructs the feature matrix of similarity measurement between DNA sequences- J/J matrix. The J/J matrix combined with the JZ-curve group not only describes the chemical properties of sequence bases, but also extracts the biological significance of gene sequences. Through the similarity analysis of the coding sequence of the first exon of 11 organism globulin gene, the experimental results show that the similarity of DNA sequence can be simply and effectively analyzed on the basis of JZ curve group combined with J/J matrix.

3.2 Graphical representation of DNA sequences

3.2.1 A New Graphical Representation of DNA Sequence – JZ-curve

The DNA sequence consists of four kinds of bases A, T, G, C. According to the chemical properties of the four kinds of bases which is Purine (A + G) / pyrimidine (c + T), amino (A + C) / ketone (G + T), strong hydrogen bonding (A + T) / weak hydrogen bonding (G + C), Academician Zhang Chunting mapped the gene sub-sequence into the center point of a certain spatial regular tetrahedron with a height equal to the length of the sub-sequence, and thus established the spatial geometric curve（Z-curve）[26,27], which is equivalent to the gene sequence. The curve carries all the information of gene sequence. The Z curve has achieved very good results in sequence research and has caused extensive influence in the world. However, the expression method of Z curve also has certain shortcomings, that is, when the frequency of each base in the sequence are the same for the input sequences, the curve generates a loop which means even these two sequences are apparently different from each other, the eigenvalues of the two sequences are the same which is unable to display biological characteristics well leading to the loss of biological information[28-30]. And under this circumstance, the result we can obtain from these two sequences tend to be the same, which means these two sequences put into the algorithm should be the same, or at least homologous. And this is the shortcoming when we do the alignment to two known sequences which obviously come from different spicies. In dealing with the defect problem of graphic expression loop, Dr. Liao Bo proposed a series of effective methods[21-25], such as introducing parameter method and multi-mapping method. In this article, a new graphic expression method of gene sequence-JZ-curve is proposed on the basis of Academician Zhang Chunting's Z curve and the graphic expression method of Liao Bo zone parameters. The curve preserves the biological significance of Z curve and eliminates the questions brought by the loop at the same time[31-32].

Assume a single-stranded DNA sequence $G = g_1 g_2 g_3 g_4 \dots g_n$, starting from the first base, the sequence is examined in turn, and only one base is added at a time. When the i-th base (i = l, 2,..., n) is investigated, the number of occurrences of A ,T, G, and C in this *sub-sequence from 1* to i is respectively represented by A_i, T_i, C_i and G_i, $A_i + T_i + C_i + G_i = i$. We transform this gene sequence $G = g_1 g_2 g_3 g_4 \dots g_n$ into three spatial curves, JZ (1), JZ (2) and JZ (3), and the node coordinates corresponding to the curve

groups are respectively defined as follows:

$$\mathrm{JZ}(1):\begin{cases} x_i = mA_i - mT_i + nG_i - nC_i \\ y_i = nA_i + nT_i + mG_i + mC_i \\ z_i = m(A_i/i)^2 + m(T_i/i)^2 + n(G_i/i)^2 + n(C_i/i)^2 \end{cases}$$

$$\mathrm{JZ}(2):\begin{cases} x_i = mA_i - mC_i + nT_i - nG_i \\ y_i = nA_i + nC_i + mT_i + mG_i \\ z_i = m(A_i/i)^2 + m(C_i/i)^2 + n(G_i/i)^2 + n(T_i/i)^2 \end{cases} \quad (3.1)$$

$$\mathrm{JZ}(3):\begin{cases} x_i = mA_i - mG_i + nC_i - nT_i \\ y_i = nA_i + nG_i + mT_i + mC_i \\ z_i = m(A_i/i)^2 + n(T_i/i)^2 + m(G_i/i)^2 + n(C_i/i)^2 \end{cases}$$

Where m, n are positive numbers. At the same time, the JZ (1) curves, JZ (2) curves and JZ (3) curves corresponding to the same gene sequence are jointly named JZ curve groups. Compared with the Z-curve mentioned in Chapter 2, adding adjustable coefficient to each base can solve the shortcoming of the Z-curve while inherit its property of simplification of the sequence alignment process.

3.2.2 Features of New Graphics

Property 1 only a unique JZ curve set corresponding to any DNA sequence.

Proof: Make$(x_{JZ(1)}, y_{JZ(1)}, z_{JZ(1)})$ $(x_{JZ(2)}, y_{JZ(2)}, z_{JZ(2)})$ $(x_{JZ(3)}, y_{JZ(3)}, z_{JZ(3)})$ the coordinates of the position i of the graph of the sequence-the JZ curve group we have:.

$$\mathrm{JZ}(1):\begin{cases} x_i = mA_i - mT_i + nG_i - nC_i \\ y_i = nA_i + nT_i + mG_i + mC_i \\ z_i = m(A_i/i)^2 + m(T_i/i)^2 + n(G_i/i)^2 + n(C_i/i)^2 \end{cases} \quad (3.2)$$

$$\mathrm{JZ}(2):\begin{cases} x_i = mA_i - mC_i + nT_i - nG_i \\ y_i = nA_i + nC_i + mT_i + mG_i \\ z_i = m(A_i/i)^2 + m(C_i/i)^2 + n(G_i/i)^2 + n(T_i/i)^2 \end{cases} \quad (3.3)$$

$$\mathrm{JZ}(3):\begin{cases} x_i = mA_i - mG_i + nC_i - nT_i \\ y_i = nA_i + nG_i + mT_i + mC_i \\ z_i = m(A_i/i)^2 + n(T_i/i)^2 + m(G_i/i)^2 + n(C_i/i)^2 \end{cases} \quad (3.4)$$

x_i, y_i and z_i are rational numbers in the form of $um + vn$.

Rewrite the above formula to:

$$x_{JZ(1),i} = (A_i - T_i)m + (G_i - C_i)n \quad (3.5)$$

$$y_{JZ(1),i} = (G_i + C_i)m + (A_i + T_i)n \quad (3.6)$$

$$z_{JZ(1),i} = \left(\left(\frac{A_i}{i}\right)^2 + \left(\frac{T_i}{i}\right)^2\right)m + \left(\left(\frac{G_i}{i}\right)^2 + \left(\frac{C_i}{i}\right)^2\right)n \quad (3.7)$$

$$x_{JZ(2),i} = (A_i - C_i)m + (T_i - G_i)n \quad (3.8)$$

$$y_{JZ(2),i} = (G_i + T_i)m + (A_i + C_i)n \tag{3.9}$$

$$z_{JZ(2),i} = \left(\left(\frac{A_i}{i}\right)^2 + \left(\frac{C_i}{i}\right)^2\right)m + \left(\left(\frac{G_i}{i}\right)^2 + \left(\frac{T_i}{i}\right)^2\right)n \tag{3.10}$$

$$x_{JZ(3),i} = (A_i - G_i)m + (C_i - T_i)n \tag{3.11}$$

$$y_{JZ(3),i} = (T_i + C_i)m + (A_i + G_i)n \tag{3.12}$$

$$z_{JZ(3),i} = ((A_i/i)^2 + (G_i/i)^2)m + ((T_i/i)^2 + (C_i/i)^2)n \tag{3.13}$$

Adding formula 3.4, formula 3.5, formula 3.7, formula 3.8, formula 3.10 and formula 3.11 gives formula 3.14 as follows:

$$\begin{aligned} \sum_{k=1}^{3} x_{JZ(k),i} + \sum_{k=1}^{3} y_{JZ(k),i} &= (3A_i + T_i + C_i + G_i)m \\ &+ (3A_i + G_i + C_i + T_i)n \\ &= (2A + i)(m + n) \end{aligned} \tag{3.14}$$

Since the coordinate values of m, n and curve group are determined, the total number of times of base A at the i position of gene sequence can be known by formula (5.6).

Formula 3.15 is obtained by adding formula 3.5 and 3.9, formula 3.16 is obtained by adding formula 3.8 and 3.12, and formula 3.17 is obtained by adding formula 3.6 and 3.11, namely:

$$x_{JZ(1),i} + y_{JZ(2),i} = (A_i + G_i)(m + n) \tag{3.15}$$

$$x_{JZ(2),i} + y_{JZ(3),i} = (A_i + T_i)(m + n) \tag{3.16}$$

$$x_{JZ(3),i} + y_{JZ(1)i} = (A_i + C_i)(m + n) \tag{3.17}$$

In the same way, from the formula 3.15, the formula 3.16 and the formula 3.17, the total number of occurrences of the bases G, T and C at the position i of the gene sequence can be known, and the sequence can be constructed through the JZ curve group, which guarantees the uniqueness of correspondence between curve and sequence.

Property2 There are no rings in the JZ curve group.

Proof: Suppose: There is a ring constructed by l bases between point (x_i, y_i, z_i) and point (x_j, y_j, z_j). According to the concept of ring, there are $(x_{JZ(1),i}, y_{JZ(1),i}, z_{JZ(1),i})$ - $(x_{JZ(1),j}, y_{JZ(1),j}, z_{JZ(1),j})$ = (0, 0, 0) . Now let's assume that x'=$x_{JZ(1)}$, y'=$y_{JZ(1)}$, z'=$z_{JZ(1)}$, (x', y', z') can be expressed by the formula 3.1 as follows:

$$\begin{cases} x' = ma' - mt' + ng' - nt' = 0 \\ y' = na' + nt' + mg' + mc' = 0 \\ z' = m(a'/i)^2 + m(t'/i)^2 + n(g'/i)^2 + n(c'/i)^2 = 0 \end{cases} \tag{3.18}$$

Since the parameter m, n are positive numbers, if and only if a '= g' = c '= r' = 0,

y '= 0 z' = 0 the formula 5.18 can be established, l= 0, that is, there is no ring in the curve JZ(1), thus avoiding the loop and having good non-degenerate properties.

By the same token, it can be proved that there is no loop phenomenon in JZ (2), JZ (3) curves.

To sum up, it is proved that there is no ring in JZ curve group which means JZ curve is reliable when transforming sequences with similar frequency of appearance of each base like A,T,C,G into curves.

Based on the two properties above, JZ curve group contains the following biometrical features:

The new graph introduces two parameters m, n; when m=n=1 , the graph contains the following biological characteristics of gene sequences:

(1) Figure JZ (1) shows the distribution of purine(A+G) / pyrimidine(C+T) bases along the sequence. When the number of purine bases is more than pyrimidine bases, x_i> 0, otherwise $x_i \leq 0$;

(2) Figure JZ (2) shows the distribution of amino(A+C) / ketone (G+T) bases along the sequence. When the number of amino bases is more than ketone bases, x_i> 0, otherwise $x_i \leq 0$;

(3) Figure JZ (3) shows the distribution of strong hydrogen bond (A+T) / weak hydrogen bond (G+C) bases along the sequence. When the number of strong hydrogen bond bases is more than weak hydrogen bond bases, x_i> 0, otherwise $x_i \leq 0$;

(4) In the figures JZ (1), JZ (2), JZ (3), the sequence length is represented by y_i.

(5) In the graphs JZ (1), JZ (2) and JZ (3), z_n represents the genome sequence index S, ($S = a^2 + c^2 + g^2 + t^2$ where a, c, g, t are the frequencies of base A,C,G,T appearing in the sequence), which is consistent with Shannon entropy by the similar definition of the larger the index is, the larger the system or the sequence is more complex. This index can measure the heterogeneity of the use of four nucleic acids in DNA sequences and usually has important biological meanings.

3.2.3 JZ Curve Group Graph of 11 Species Gene Sequences

3.2.3.1 Experimental Data

As there is a commonly used database for testing DNA sequence alignment all used in the 2-D,3-D graphic algorithms and the Z-curve and other graphic representation algorithms, the DNA sequences of the first exon of the β globulin gene

of 11 species from the NCBI database was obtained as experimental data. By using this database, we can more intuitively see the difference between the JZ-curve and other algorithms on sequence alignment. Table 3.1 lists the base codes of the 3 sequences of 11 species. These sequences are very conservative, i.e. evolution is very slow, and many documents[21,59,76-78] have adopted these sequences for comparison and analysis of DNA sequences.

3.2.3.2 Experimental Results and Analysis

We can obtain different graphs by assigning values to m and n in the JZ curve group. Therefore, it is very important to choose the appropriate value of m, n for drawing the appropriate figure. In this article, the JZ (1) curves, JZ (2) curves and JZ (3) curves obtained by the m=5 and n=1/11 (the value of m and n are selected randomly) are given as Figure 3.1, Figure 3.2 and Figure 3.3 respectively. Observing each sequence curve, it is found that the curves in Figure 3.1, Figure 3.2 and Figure 3.3 all show the trend of extending in the same direction, which is preliminary compliance with that all these sequences are from the same gene. At the same time, it is also found that the curves of Chimpanzee. are obviously different from other curves (the reason of this is because Chimpanzee. is the only oviparous species), and the other 10 curves are similar (all viviparous born and share a similar functioning organ). Human and Gorilla are very similar in graphics. Based on graphics, we can intuitively and roughly judge the similarity of genes of various species. However, the similarity between gene sequences of different species cannot be accurately explained here[34-37] only by primarily comparing these curves. We need to extract appropriate characteristic parameters from graph curves to analyse the degree of similarity between different gene sequences. The calculation of matrix invariant parameters based on graph will be described in detail in the next section.

Table 3.1: DNA Sequence of Exon 1 of β Globulin Gene of 11 species.

Species	Sequence coding
Human (1)	ATGGTGCACTGACTCCTGAGGAGAAGTTCTGCCGTTTACTGCCCTGTGTGGGCAA GGTGAACGTGGATTAAGTTGGTGGTGGTGAGGCCCTGGGCAG
Goat (2)	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACGGGCTTCTGGGGCAAGGTGA AAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum (3)	ATGGTGCACTTGACTTCTGAGAAGAACTGCATCACTACCATCTGGTCTAAG GTGCAGGTTGACCAGACTGGTGGTGTGAGGCCCTTGGCAG
Gallus (4)	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGGCCTGGGGGCAA GGTCAATGTGGCCGAATGTGGGGGCCGAAGCCCTGGCCAG
Lemur (5)	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACTCTCTCTGTGTGGGGCAAG GTGGATGTAGAGAAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse (6)	ATGGTTGCACTGACTGATGCTGAGAAGTCTGCTGTTCTCTTGCCTGTGGGCAAA GGTGAACCCGATGAAGTTGGTGGTGAGGGCCTGGGCAGG
Rabbit (7)	ATGGTGCATCTGTCCAGTGAGGAGAGTCTGCGGTCACTGCCCTGTGGGGGCAA GGTGAATGTGGAAGAAGTTGGTGGTGGTGAGGCCTGGGC
Rat (8)	ATGGTGCACCCTAACTGATGCTGAGAAGGCTACTGTTAGTGGGCCTGTGGGGAAA GGTGAACCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla (9)	ATGGTGCACTGACTCCTGAGGAGAAGTTCTGCCGTTTACTGCCCTGTTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGGTGGAGGCCTGGGCAGG
Bovine (10)	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCCCACGCCTTTGGGGCAAGGTGAA AGTGGATGAAGTTGGTGGTGGTGAGGGCCCTGGGCAG
Chimpanzee (11)	ATGGTGCACTGACTCCTGAGGAGAAGTTCTGCCGTTTACTGCCCTGTTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGGTGGAGGCCTGGGGCAGGTTGGTATCAAGG

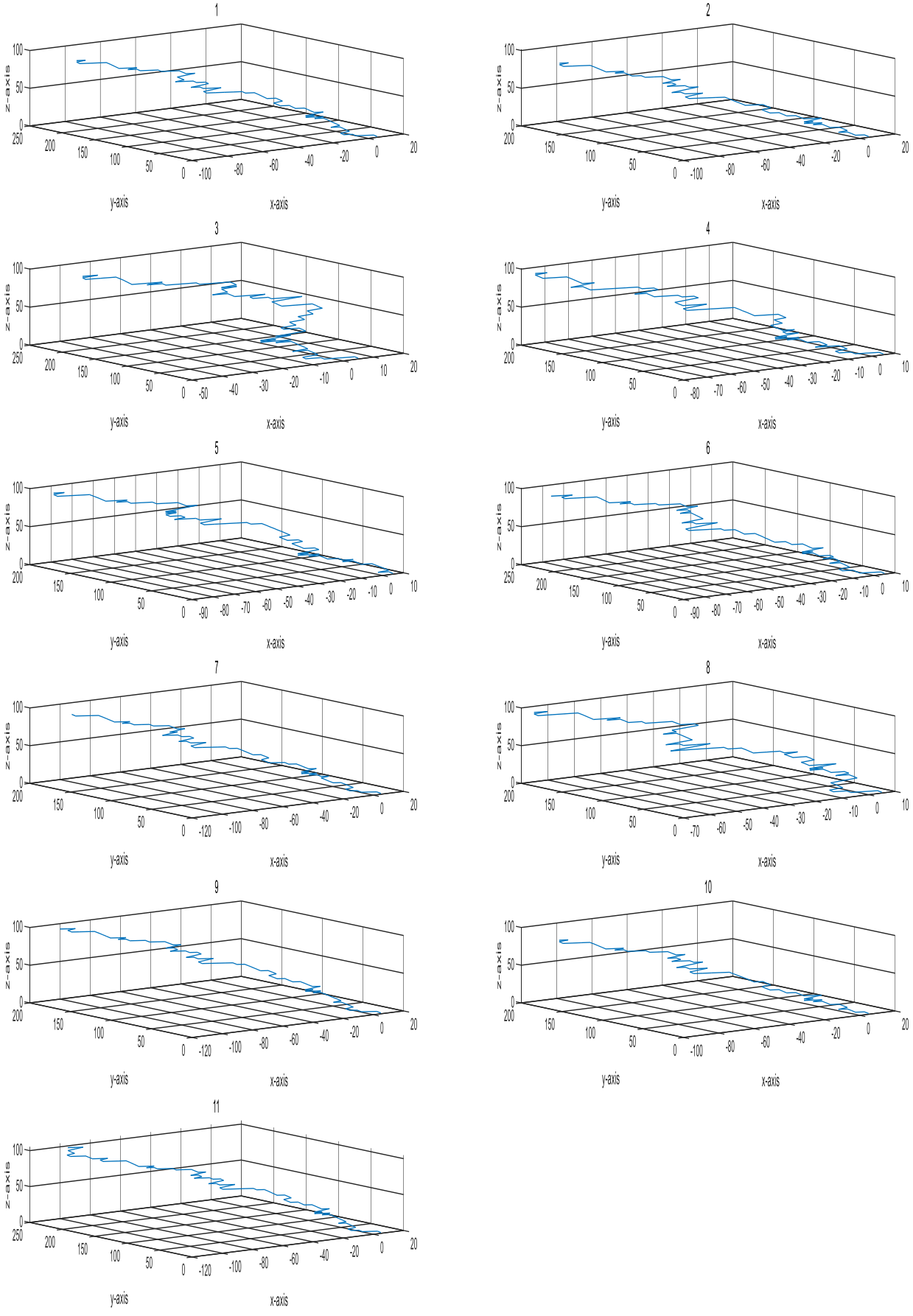

Figure 3.1 JZ (1) curve of the first exon of β globulin gene of 11 species

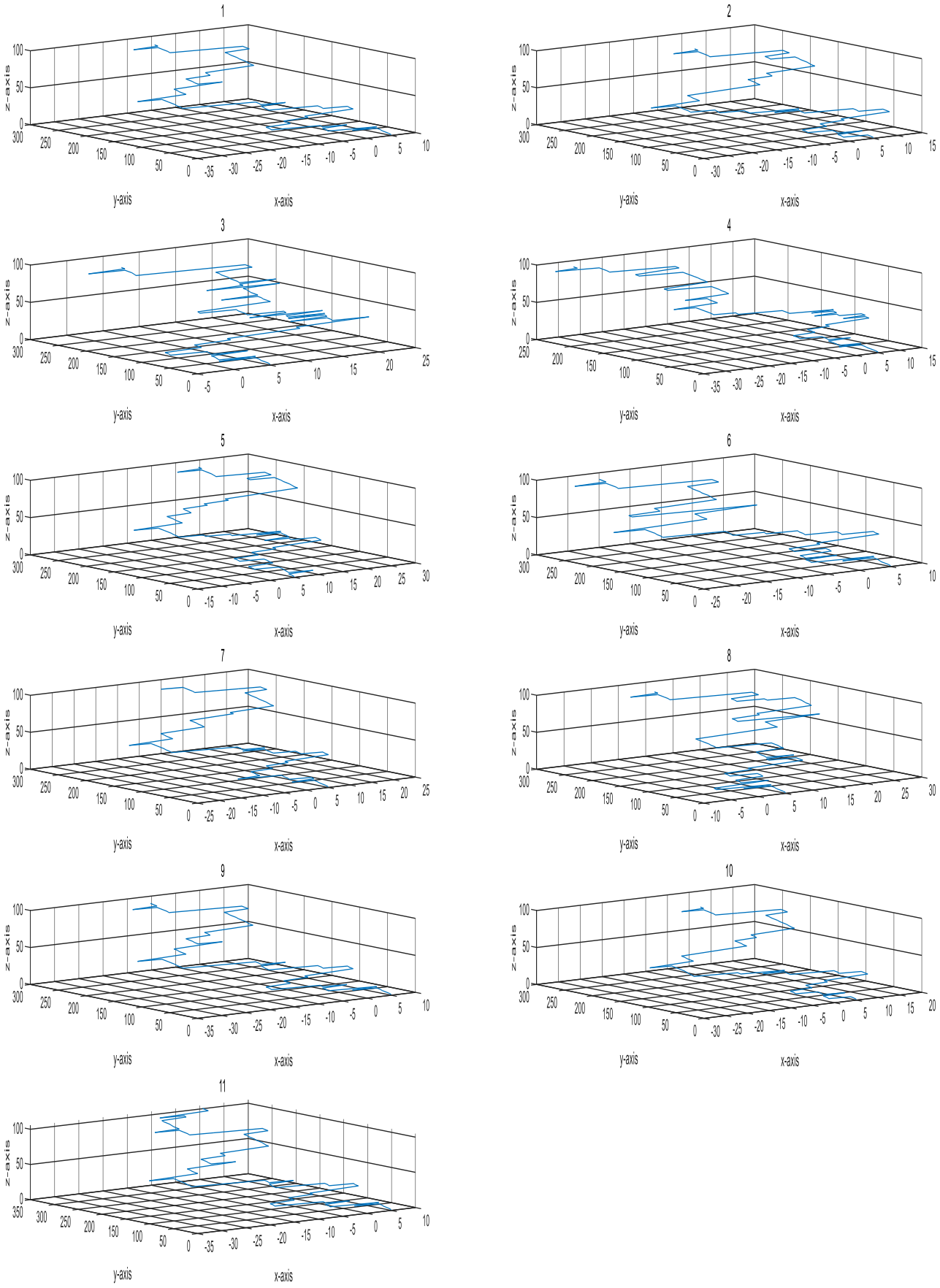

Figure 3.2 JZ (2) curve of the first exon of β globulin gene of 11 species

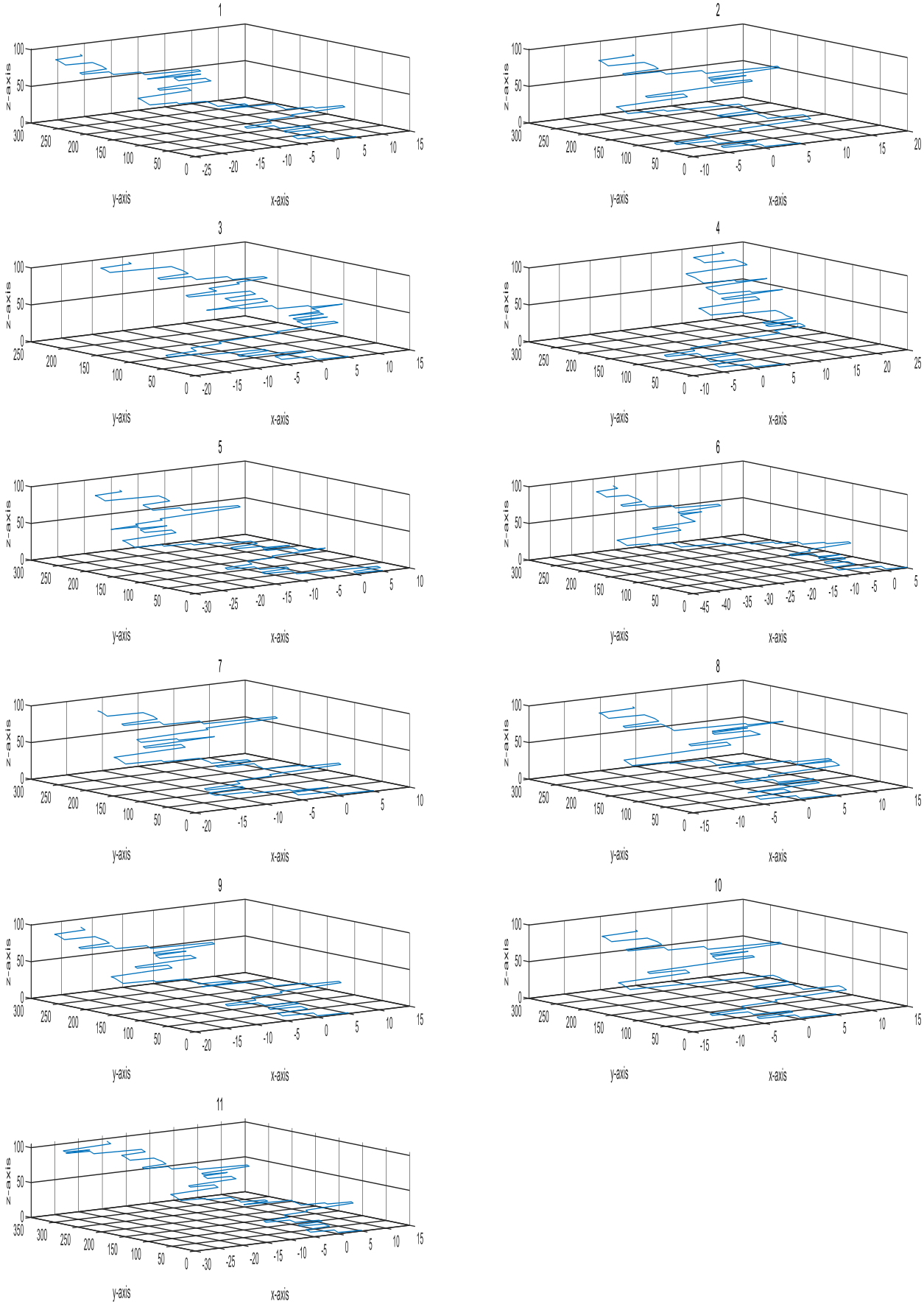

Figure 3.3 JZ (3) curve of the first exon of β globulin gene for 11 species

3.3 Sequence Similarity Analysis of Based on JZ Curve Group

Evolution theory[39] holds that if there is sufficient similarity between the two sequences, it is speculated that they may have the same evolutionary ancestor[40]. In bioinformatics, sequence similarity analysis is multi-faceted, which may be similar in the composition of base sequences, structure or function. Sequence similarity analysis can be used for quantitative analysis[55-58]. Generally, scoring matrix is used to describe the two-to-two comparison of sequences, and the best comparison path is found by constructing scoring matrix.

Recently, the matrix invariant method based on graphical representation has been used to analyze the similarity of DNA sequences[38]. Its idea is to transform DNA sequences into some graphical representations, use graphical representations to construct matrices, and then use matrix invariants (such as maximum eigenvalues, maximum row sums, trace of matrices, etc.) to compare the similarity of biological sequences.

3.3.1 A Feature Matrix Based on JZ Graphic Group

According to the characteristics of JZ curve group, this paper proposes a new characteristic matrix, J/J matrix, based on L/L matrix. The matrix value consists of the sum of two parts. The first part is the ratio of the Euclidean distance between the corresponding points x, y axes of the two bases on the curve and the graph theory distance between them (where $\sum_{k=1}^{j-1} e_{xy}(k,k+1)$ represents the sum of the line segment distances between the two points (i,j)of the curve calculated on the axis), which examines the chemical properties of the base. The second part is the ratio of the Euclidean distance of the corresponding point Z axis of the two bases on the curve to the graph theory distance between them (where $\sum_{k=1}^{j-1} e_z(k,k+1)$ represents the sum of line segment distances between two points (i,j) of the calculated curve on the Z axis, which characterizes the biological significance of the sequence. Compared with the traditional matrix, the J/J matrix not only examines the chemical properties of the sequence bases, but also examines the biological significance of the sequence, so it can describe the characteristics of the sequence more comprehensively.

$$J(i,j) = \frac{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}}{\sum_{k=1}^{j-1} e_{xy}(k,k+1)} + \frac{\sqrt{(z_i-z_j)^2}}{\sum_{k=1}^{j-1} e_z(k,k+1)} \tag{3.19}$$

3.3.2 Similarity Analysis Algorithm of Gene Sequences of 11 Species

The basic idea of similarity analysis of DNA sequences with invariant parameters of DNA sequences is to construct vector of sequences according to invariants of matrices. If two DNA sequences are similar, the smaller the Euclidean distance between their corresponding two vectors. In order to compare the differences in sequence similarity analysis, we construct 3-dimensional vectors with regularization parameters (λ/N) , where N is the number of nucleotide groups in the corresponding DNA sequence.

The Euclidean distance calculation formula between vectors based on regularization parameters of different curves is as follows:

$$D_{a,b} = \sqrt{\left(\frac{\lambda_{a,1}}{N_a} - \frac{\lambda_{b,1}}{N_b}\right)^2 + \left(\frac{\lambda_{b,2}}{N_a} - \frac{\lambda_{b,2}}{N_b}\right)^2 + \left(\frac{\lambda_{a,3}}{N_a} - \frac{\lambda_{b,3}}{N_b}\right)^2} \qquad (3.20)$$

Description of Similarity Analysis Algorithm

Input: 11 gene sequence coding files

Output: Similarity Distance Matrix of 11 Sequences

Begin

for (se=l; se<= num; se++) //num is the number of DNA sequences of a species

Converting sequence characters into JZ curve group graphics according to formula 3.1

for (k=1; k<=3; k++)// For each sequence diagram

for (i=l; i<=n; i++) //n is the sequence length

for (j=i+l; j<=n; j++)

Calculate J/J matrix, L/L matrix according to formula 3.19, formula 2.11

end

end

Calculating the maximum eigenvalue of matrix as matrix invariant λ_k

Regulate the parametric λ_k/n

end

end

According to the formula 3.20, the distance value is calculated, and the similar distance matrix is constructed.

END

3.3.3 Experimental Results and Analysis

This paper lists the experimental results of (m=29, n=3) and (m=5, n=1/11). The maximum eigenvalues of the regularization of the L/L matrix and the J/J matrix based on the JZ curve group (m=29, n=3) are shown in Table 3.2, and then the similarity comparison data between 11 species are obtained in Table 3.3, and Table 3.4. Finally, the similarity comparison data between each species and other 10 species are added to obtain Table 3.5, which can simply judge the genetic distance relationship between one species and other species. Based on the regularized maximum eigenvalues of the L/L matrix and the J/J matrix of the JZ curve group (m=5, n=1/11) , see Table 3.6, and then obtain the similarity comparison data between 11 species, see Table 3.7 and Table 3.8. Finally, the similarity comparison data between each species and other 10 species are added to obtain Table 3.9, which can simply determine the genetic distance between a species and other species.

Table 3.2: regularized maximum eigenvalues based on L/L matrices and J/J matrices of JZ curve sets (m=29, n=3)

λ		Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
L/L	$\lambda_{\varphi 1}$	0.6324	0.6570	0.5881	0.6764	0.6093	0.6279	0.6388	0.6013	0.6442	0.6456	0.6415
	$\lambda_{\varphi 2}$	0.6729	0.6744	0.5954	0.5985	0.6819	0.6777	0.6985	0.6531	0.6730	0.6869	0.6807
	$\lambda_{\varphi 3}$	0.5523	0.5158	0.5493	0.5291	0.5308	0.5542	0.5358	0.5254	0.5465	0.5227	0.5451
J/J	$\lambda_{\varphi 1}$	0.8128	0.8921	0.7633	0.8826	0.8258	0.7831	0.8431	0.7888	0.8156	0.8696	0.8014
	$\lambda_{\varphi 2}$	0.9099	0.8740	0.8283	0.8076	0,8860	0.8768	0.9121	0.8669	0.9083	0.9202	0.9140
	$\lambda_{\varphi 3}$	0.7957	0.7470	0.7636	0.7289	0.8152	0.7799	0.7726	0.7246	0.8065	0.7876	0.7895

The similarity distance matrix (Table 3.3) obtained by constructing the L/L matrix on the JZ curve group can be found that Human and Mouse (0.0069),Gorilla and Chim.(0.0083), Human and Gorilla (0.0132), Human and Chim.(0.0140) have the smallest difference index, indicating that they are relatively similar, which is due to these species share the same primate ancestor. Gallus and Opos have the largest difference index with other species, indicating that these two organisms are the least similar to other species, which is basically consistent with the biological facts(labelled data). This shows that JZ curve group can reasonably extract biological sequence information and is effective.

However, through observing table 3.3, also found some differences with the actual situation. For example, the difference index between Human and Mouse is the smallest

(0.0069), and the difference index between Human and Rat (0.0457), Rat and Mouse (0.0463) is roughly the same. It may be that some sequence information is lost when the sequence is converted into a graph, it may be that the biological information extracted by the matrix is not complete enough, or it may be that there is some unknown similarity between species, which has been similar in other literatures[20-22].

At the same time, observing the table 3.5, it is found that the sum of (0.9601) based on the matrix L/L is greater than the sum of (0.9303). This is contrary to the fact that Gallus is the only non-mammal species among the 11 species, and Opos is the farthest mammal species. Many literatures have also reached the conclusion that is inconsistent with the fact[16-22].

Table 3.3: calculates similarity distance matrices of 11 species based on JZ curve groups and L/L matrices(m=29, n=3)

	Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
Human	0	0.0441	0.0894	0.0895	0.0328	0.0069	0.0311	0.0457	0.0132	0.0353	0.0140
Goat		0	0.1101	0.0794	0.0505	0.0484	0.0362	0.0604	0.0333	0.0184	0.0338
Opos.			0	0.0906	0.0910	0.0916	0.1157	0.0639	0.0958	0.1113	0.1008
Gallus				0	0.1070	0.0962	0.1070	0.0929	0.0830	0.0939	0.0908
Lemur					0	0.0302	0.0342	0.0303	0.0393	0.0375	0.0353
Mouse						0	0.0298	0.0463	0.0187	0.0373	0.0167
Rabbit							0	0.0598	0.0282	0.0187	0.0202
Rat								0	0.0518	0.0557	0.0526
Gorilla									0	0.0276	0.0083
Bovine										0	0.0236
Chim											0

Table 3.4: calculates the similarity distance matrix of 11 species based on JZ curve group and J/J matrix(m=29, n=3)

	Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
Human	0	0.0997	0.1008	0.1407	0.0334	0.0473	0.0382	0.0865	0.0113	0.0582	0.0137
Goat		0	0.1377	0.0695	0.0958	0.1139	0.0671	0.1059	0.1027	0.0655	0.1078
Opossum			0	0.1260	0.0995	0.0548	0.1161	0.0605	0.1048	0.1425	0.0972
Gallus				0	0.1297	0.1315	0.1199	0.1111	0.1437	0.1276	0.1469
Lemur					0	0.0562	0.0529	0.0997	0.0259	0.0620	0.0451
Mouse						0	0.0701	0.0564	0.0526	0.0971	0.0425
Rabbit							0	0.0854	0.0439	0.0314	0.0451
Rat								0	0.0956	0.1155	0.0811
Gorilla									0	0.0584	0.0229

Bovine										0	0.0684
Chim											0

From the similarity distance matrix obtained from J/J matrix (Table 3.4), it is found that Human and Gorilla (0.0113), Human and Chim. (0.0137), Gorilla and Chim. (0.0229) have the smallest difference index, and Gallus and Opos have the largest difference index with other species, which is more in line with biological significance(which is already labelled on the database by NCBI). Different from Table 3.2, the difference index between Human and Mouse (0.0473) has been significantly increased which is a 585.51% increase, which remedy the defects of the L/L matrix. At the same time Human and Rat difference index increased to (0.0865) which is 89.28% increase than under L/L matrix, much larger than the difference index between Rat and Mouse (0.0564) which is more in line with the biological significance. At the same time, observing the table 3.5, it is found that the sum of the difference indexes (1.2466) of Gallus and other organisms based on the J/J matrix is greater than the sum of the difference indexes (1.0398) of Opos., which is 19.89% increase than the L/L matrix and is closer to their biological definition. This proves the validity of the J/J matrix.

Table 3.5: Sum of Similarities between 11Species and Other 10 Species (m=29, n=3)

	Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
L/L	0.4018	0.5146	0.9601	0.9303	0.4881	0.4221	0.4809	0.5593	0.3991	0.4591	0.3960
J/J	0.6298	0.9657	1.0398	1.2466	0.7002	0.7225	0.6702	0.8977	0.6617	0.8266	0.6717

Table 3.6: regularized maximum eigenvalues based on L/L matrices and J/J matrices of JZ curve sets(m=5, n=1/11)

λ		Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
L/L	$\lambda_{\varphi 1}$	0.6314	0.5583	0.6530	0.5802	0.6013	0.6133	0.5729	0.6188	0.6194	0.6163	0.6415
	$\lambda_{\varphi 2}$	0.6520	0.5674	0.5683	0.6614	0.6562	0.6784	0.6291	0.6508	0.6663	0.6590	0.6807
	$\lambda_{\varphi 3}$	0.4812	0.5156	0.4966	0.4955	0.5202	0.5005	0.4896	0.5129	0.4877	0.5108	0.5451
J/J	$\lambda_{\varphi 1}$	0.8590	0.7323	0.8541	0.7952	0.7540	0.8121	0.7597	0.7868	0.8365	0.7734	0.8014
	$\lambda_{\varphi 2}$	0.8483	0.7997	0.7757	0.8631	0.8543	0.8908	0.8423	0.8856	0.8972	0.8920	0.9140
	$\lambda_{\varphi 3}$	0.7139	0.7284	0.6958	0.7777	0.7452	0.7364	0.6867	0.7708	0.7522	0.7532	0.7895

Table 3.7: Based JZ Curve Group and L/L Matrix Calculate Similarity Distance Matrix of 11 Species(m=5, n=1/11)

	Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
Human	0	0.0451	0.0962	0.0969	0.0368	0.0079	0.0341	0.0493	0.0136	0.0372	0.0152
Goat		0	0.1171	0.0878	0.0540	0.0495	0.0374	0.0634	0.0342	0.0198	0.0340
Opossum			0	0.0967	0.0987	0.0988	0.1249	0.0685	0.1031	0.1196	0.1086
Gallus				0	0.1182	0.1046	0.1171	0.1008	0.0908	0.1040	0.0989
Lemur					0	0.0329	0.0375	0.0337	0,0436	0.0402	0.0392
Mouse						0	0.0321	0.0498	0.0197	0.0386	0.0179
Rabbit							0	0.0647	0.0307	0.0187	0.0222
Rat								0	0.0558	0.0596	0.0568
Gorilla									0	0.0296	0.0088
Bovine										0	0.0244
Chim.											0

Looking at table 3.7, we found that the similarity distance matrix obtained based on L/L matrix has the above similar problems, such as: the difference index of Human and Mouse (0.0079) is too small, and the difference index (0.0493) of Human and Rat is slightly smaller than the difference index (0.0498) of Rat and Mouse. At the same time, observing the table 3.8, it is found that the sum of the difference indexes between Opos. and other species (1.0322) in the L/L matrix is greater than the sum of the difference indexes between Gallus and other species (1.0157). This is inconsistent with the biological definitions. The above-mentioned problems can be well solved through the J/J matrix. The difference index of Human and Mouse (0.0230)which is 191.14% increase than under L/L matrix, has been significantly improved, and the difference index of Human and Rat has been increased to (0.0886), much larger than the difference index (0.0600), of Rat and Mouse, which is more in line with our intuition. At the same time, observing the table 3.9, it is found that the sum of the difference indexes between the J/J matrix Gallus and other species (1.3018) is greater than the sum of the difference indexes between Opos and other species (1.0785), which also corrects the bad results generated from L/L matrix. This shows that the J/J matrix applied to the JZ curve group can simply and effectively analyse the similarity between sequences.

Table 3.8: calculates the similarity distance matrix of 11 species based on JZ curve group and J/J matrix(m=5, n=1/11)

	Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
Human	0	0.0970	0.1050	0.1466	0.0317	0.0455	0.0381	0.0886	0.0115	0.0555	0.0129
Goat		0	0.1364	0.0749	0.0915	0.1097	0.0672	0.1031	0.0993	0.0661	0.1039
Opossum			0	0.1283	0.1020	0.0611	0.1213	0.0655	0.1102	0.1446	0.1040
Gallus				0	0.1335	0.1365	0.1291	0.1158	0.1491	0.1351	0.1528
Lemur					0	0.0532	0.0525	0.0999	0.0250	0.0593	0.0437
Mouse						0	0.0692	0.0600	0.0520	0.0933	0.0432
Rabbit							0	0.0870	0.0430	0.0298	0.0422
Rat								0	0.0984	0.1149	0.0841
Gorilla									0	0.0543	0.0230
Bovine										0	0.0633
Chim											0

Table 3.9: Sum of similarities between 11 species and other 10 species

	Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
L/L	0.4322	0.5423	1.0322	1.0157	0.5348	0.4517	0.5193	0.6023	0.4298	0.4917	0.4260
J/J	0.6323	0.9490	1.0785	1.3018	0.6923	0.7236	0.6794	0.9174	0.6658	0.8163	0.6732

In recent years, there have been much research on the similarity relationship between species. Taking 11 species in Table 3.1 as an example, Table 3.10 lists the targets of this article and other documents comparing the similarity between Opos, Gallus and other species[21,59,76-78], we can find that the similarity between Opos, Gallus and other species exists in general. The method proposed in this article can correct the similarity effectively, and the results obtained are closer to the facts. In the table 3.10, the A1 column is from the document table 3.4(J/J Matrix), A2 column is from the document table 3.3(L/L Matrix), A3 column is from the document table 3.8(J/J Matrix), A4 column is from the document table 3.7(L/L Matrix), B column is from the document [76] table4 (Zhang YC's), C column is from the document [57] table4 (2D Z-curve), D column is from the document [59] table4 (4D Z-curve), E column is from the document [77] table4(Wang J's), F column is from the document [78] table4(Song J's). Through this table, only column A1 and A3 (which are under J/J matrix)give the result that the sum of similarities of Gallus and other 10 species has am approximately +20%

difference than the one of Opossum, while other results either shows a +0.09% or negative difference rate from -0.03% to -40.9% which is contrary to biological fact which Opossum is mammal while Gallus is the only oviparous specie among the 11 species.

Table 3.10: Sum of similarities between Gallus, Opossum and other 10 species

	Al	Al	A3	A4	B	C	D	E	F
Gallus	1.2466	0.9303	1.3018	1.0157	0.2786	0.0597	0.9913	164.7	2.8247
Opossum	1.0398	0.9601	1.0785	1.0322	0.4374	0.1010	0.9045	183.4	3.1334
Difference Rate (Gallus/Opossum)	0.1989	-0.0310	0.2070	-0.0160	-0.3631	-0.4089	0.0960	-0.1020	-0.0985

3.4 Summary

This chapter presents a new graphical representation of JZ curve set based on Z curve and Liao Bo's 3D graphical expression method, which makes it have the characteristics of non-degeneration while retaining the biological characteristics of gene sequence. First of all, by assigning different vectors to the four bases of the DNA sequence, we can convert the character coding of the DNA sequence into spatial curves, and there is a unique JZ curve group corresponding to any DNA sequence. Then it is proved that there is no ring and degeneration phenomenon in the JZ curve group, and the JZ curve group contains some biological characteristics. In section 3.2.2, we draw the JZ curve set for the DNA sequence of the first exon of β globulin gene gene of 11 species.

Our graph can bring visual observation of gene sequences, but in order to analyze sequence similarity, we need to further study the matrix invariant calculation of the graph. This chapter presents a method to measure the similarity of DNA sequences. Compared with the method based on character comparison, our method takes into account the character encoding information of DNA sequences and the characteristics of sequences at the visual graphic level. The method is simple, avoids the limitation of multiple sequence alignment, and can accurately judge the similarity between DNA sequences. Throughout the tests on the DNA database by NCBI and compared with the results from other articles using other 3-D graphic alignment algorithms, the one-dimensional gene sequence can have visualization characteristics. Also we can use feature matrix to easily calculate and realize the DNA sequence comparison. Through

the generation of feature matrix, we can obtain 1-D, 2-D, 3-D and even high dimensional statistics of a gene sequence, and it shall benefit in future research of the gene sequence.

Chapter 4. Evolutionary Tree Construction Algorithm

4.1 Introduction

It is a very important issue in modern molecular evolution research to reconstruct the evolutionary history of organisms according to the existing biological genes or species diversity. Since the evolution rates of different genes or DNA fragments are quite different , the evolutionary relationship between organisms at almost all levels can be studied by these genes or DNA fragments.

This chapter briefly introduces the method of building phylogenetic tree[41-43], the main steps of constructing sequence phylogenetic tree in chapter 4.2.1 and 4.2.2.

Then a new fuzzy hierarchical clustering algorithm based on atlas theory to cluster on the basis of sequence graph representation to construct species phylogenetic tree is introduced in chapter 4.2.3.

And the results and discussion of experiments of 3 sets of data are shown in chapter 4.3.

4.2 Construction of Biological Phylogenetic Tree Based on Fuzzy Clustering Transfer Algorithm of Spectrum Theory

4.2.1 Spectrum Theory

The idea of the earliest spectrogram theory[79] comes from the graph partition theory. It is assumed that each data sample is regarded as a vertex ω in the graph, and the edges E between vertices are assigned weight values according to the similarity between samples, thus obtaining an undirected weighted graph G= (V, E) based on sample similarity. In the graph partitioning problem[80], the graph G is divided into A, B subgraphs (where $A \cap B=\varphi$, $A \cup B = V$), and the formula 4.1 is the cost function of graph partitioning, that is, it requires the maximization of the connection weight in the subgraph and the minimization of the edge weight between the subgraphs:

$$\begin{gathered} min\, N\, cut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)} \\ vol(A) = \sum_{i\in A}\sum_{i=j} w_{ij} \\ cut(A,B) = \sum_{i\in A}\sum_{j\in B} w_{ij} \end{gathered} \tag{4.1}$$

Where cut (A,B) is the edge between subgraphs A,B , also called ‘edge tangent

set'[44]. From formula (4.1), we can see that the objective function not only satisfies the small similarity between samples of different classes, but also satisfies the large similarity between samples of different classes.

Due to the essence of graph partitioning problem, it is a difficult problem to find the optimal solution of graph partitioning criterion[80]. A good solution method is to consider the continuous relaxation form of the problem[45-46]. Spectral graph theory is to use matrix theory and linear algebra theory to study the adjacency matrix of graph[47-49]. The graph partition problem is converted into spectral decomposition for solving similarity matrix or Laplacian matrix, That is, the eigenvector corresponding to the first K minimum eigenvalues, The realization of the best approximation to the graph partition criterion is called that it represents a solution of the best graph partition[50]. Compared with the traditional fuzzy similarity matrix, the new matrix constructed by spectrogram theory not only considers the dispersion degree between classes, but also considers the compactness degree of the same class, which can better describe the evolutionary distance between sequences and improve the accuracy of clustering.

The similarity between N elements in the graph partitioning problem is defined as follows:

$$W_{ij} = exp\left(-\frac{d(s_1,s_j)}{2\varepsilon^2}\right) \tag{4.2}$$

Where d (s_i, s_j) is the Euclidean distance between elements s_i, s_j, and the calculation formula is shown in (6.1), W matrix is also known as similarity matrix or affinity matrix.

The theoretical analysis of spectrogram is based on the Laplacian matrix of graph, which was proposed by Fiedler [81] in 1973. There are nonstandard Laplacian matrices and standard Laplacian matrices, which are respectively defined as follows:

$$L = D - W \tag{4.3}$$

$$L = D^{-1/2}WD^{-1/2}$$
$$L(i,j) = W(i,j)/(\sqrt{D(i,i)}\sqrt{D(j,j)}) \tag{4.4}$$

Among them, $D_{n\times n}$ matrix is a diagonal matrix formed by adding each row of elements of the $W_{n\times n}$ matrix to obtain degree values of diagonal elements, which is called degree matrix.

The eigenvectors x_1, $x_2, x_3 \ldots, x_k$ corresponding to the first K minimum eigenvalues of the affinity matrix or the Laplacian matrix are calculated, and the matrix X = $[x_1, x_2, x_3 \ldots, x_k]_{n\times k}$ is constructed. Researches have shown that these

eigenvectors represent a solution of the best graph decomposition[82,83].

Scott G [82] uses n by K-dimension vectors to represent N elements in the graph partition problem and calculates the cosine value between K dimension eigenvectors to be the similarity between the n elements.

$$\cos(A,B)=\frac{A\times B}{|A|\times|B|} \tag{4.5}$$

The element values of the new matrix based on atlas theory proposed by Weiss Y [83]are between [-1, 1] , and the closer the element values are to 1, the smaller the difference degree between the corresponding clustering samples, and vice versa. At the same time, compared with the original affinity matrix, the new matrix constructed based on the graph theory considers the graph partition problem and is a solution method of the cost function (formula 4.1). Therefore, the new matrix not only increases the similarity of the same kind, but also reduces the similarity of the different kinds, and has better clustering resolution.

4.2.2 Transfer Algorithm of Fuzzy Clustering Analysis

The general steps of fuzzy clustering need to first calculate the fuzzy similarity relation r ($0 \le r \le l$) between classified objects to describe the correlation degree between classified species, then establish the paste equivalent matrix center according to the fuzzy similarity relation t(r) , and finally select a suitable parameter $\lambda \in (0,1)$ to obtain the equivalent cut set of the fuzzy equivalent matrix t(r), which is based on λ level of equivalent clustering results. Wang Y [73] thinks that the method of fuzzy clustering based on equivalence thought needs people to further construct fuzzy equivalence relation on the basis of fuzzy similarity matrix, that is, symmetric matrix satisfying reflexivity, symmetry and transitivity at the same time. The transitive closure method[73] can find the minimum element of the transitive fuzzy similarity matrix r , that is, t(r) through continuous calculation of the fuzzy similarity matrix r . Obviously, the working time of transitive closure method for calculating closures occupies most of the calculation time of the whole fuzzy clustering algorithm, and the clustering efficiency is low.

The transfer algorithm proposed by Wu Fubao [74] realizes the purpose of clustering analysis directly from fuzzy similarity matrix. Its basic idea is to introduce confidence level λ on the basis of fuzzy similarity matrix, the common similarity relation r_λ is

obtained as follows: cluster analysis is directly carried out from the common similarity matrix by setting the level λ. By setting confidence level λ, the transmission method directly obtains clustering results from fuzzy similarity relation r ; The transitive closure method must synthesize fuzzy similarity matrices to construct fuzzy equivalent relation t(r), and then cluster them through confidence level λ. Then, what is the relationship between them? In this article[72], through mathematical reasoning, the following conclusions are proposed and proved: under the same confidence level, the clustering results obtained by the transitive method and the transitive closure method are equivalent to each other for fuzzy similarity relations.

4.2.3 A New Algorithm Description for Constructing Evolutionary Tree

The gene sequence consists of four characters A,G,T and C, and the fuzzy clustering algorithm requires that the initial data processed must be numerical data, so the next step of fuzzy clustering can only be carried out after the numerical processing of the gene sequence is completed. Graphic representation of gene sequence is a numerical processing method.

4.2.3.1 Graphic Representation and Extraction Features of Gene Sequence

Graphic representation of gene sequence is a numerical processing method. Since the JZ curve group proposed in chapter 3 is the only representative of genes and can be used to reflect the characteristics of gene sequences, the differences between JZ curve groups of gene sequences are the basis for constructing evolutionary trees.

According to the formula 3.1, each gene sequence can be expressed by a group of JZ curves, i.e. JZ (1), JZ (2), JZ (3), three gene sequence characteristic matrices can be constructed from each curve of the formula 3.19, and the regularized matrix invariant λ/N (here, λ takes the maximum eigenvalue of the matrix, N is the sequence length), and the characteristic information of the gene sequence can be described by a three-

Algorithm 1: Description of Numerical Processing Algorithm for Gene Sequence

(1) Converting each gene sequence into JZ curve group graph according to formula 3.1;

(2) Calculate the J/J matrix corresponding to a graph according to formula 3.19;

(3) Calculate the regularized matrix invariant of J/J matrix, λ/N:

dimensional vector composed of the corresponding regularized matrix invariants;

4.2.3.2 Constructs a New Clustering Matrix Based on Spectrum Theory

According to incomplete statistics, there are 13 methods to construct fuzzy similarity matrix[52-54]. It is generally believed that the absolute value reciprocal method has good clustering resolution. Based on spectral graph theory, the spectral decomposition of fuzzy similarity matrix is solved, clustering samples are described by spectral decomposition, and the distance between sample vectors is calculated to construct a new distance matrix. Compared with the original fuzzy similarity matrix, the new clustering matrix is a solution method of the cost function 4.1. It not only considers the dispersion degree between classes, but also considers the compactness degree of the same class, which can better describe the evolution distance between sequences and has better clustering resolution.

Algorithm 2. Constructs a New Clustering Matrix Based on Atlas Theory

1. Construct Laplacian Matrix

(1) construct a similarity distance matrix $D(a,b)_{n\times n}$ of N species accord to formula 3.20. If that element value of the similarity matrix of species is larger, the similarity of species is higher, and vice versa.

(2) build a fuzzy similarity matrix $W_{n\times n}$ by the formula

$w_{ij}=\begin{cases}1, i=j\\ \frac{c}{\sum_{k=1}^{m}|x_{ik}-x_{jk}|}, i\neq j\end{cases}$ and the element values of fuzzy similarity matrix are between 0 and 1.

(3) construct a degree matrix $D_{n\times n}$, which is a diagonal matrix formed by adding each row of elements of the degree matrix to obtain degree values and taking all degree values as diagonal elements;

(4) constructs a standardized Laplacian matrix $L_{n\times n}$ according to the formula 4.4.

2. Calculating eigenvectors of Laplacian matrix and constructing a new clustering matrix

(1) obtain the eigenvector X corresponding to the minimum first K eigenvalues of Laplacian matrix L,

(2) orthogonalizes the X matrix and uniformizes the X row vector to obtain the matrix $N_{n\times k}$,

(3) recalculates the similarity matrix according to the formula 4.5, and outputs a new clustering matrix $Q_{n\times n}$

4.2.3.3 Transfer Algorithm for Fuzzy Clustering Analysis Hierarchical Generation of Evolutionary Tree

(1) Pick Cut Set

In this article, the element values of clustering matrix are selected to construct the cut set $\lambda=\{\lambda_1,\lambda_2,\dots,\lambda_n\}(\lambda_1<\lambda_2<\cdots<\lambda_n)$, and different classification is obtained when taking different confidence level λ. When $\lambda=\lambda_1$ are used, all species are classified into the same kind, and with the increasing of λ value, they are clustered from coarse to fine levels, and finally when λ =1, all species were classified into one category and formed an evolutionary tree.

(2) Transfer Algorithm of Fuzzy Clustering Analysis

The basic idea of fuzzy transfer algorithm [75] is to introduce confidence level on the basis of fuzzy similarity matrix[81], get the common similarity relation, and then cluster analysis is directly carried out from the common similarity matrix by setting horizontal people. The transfer method realizes the purpose of clustering analysis directly from fuzzy similarity matrix. Compared with the "transfer closure method", the clustering results of the two methods are the same, but the time complexity and space complexity of the transfer method are far less than those of the transfer closure method. Therefore, this paper adopts the transfer algorithm of fuzzy clustering analysis to generate evolutionary tree.

Algorithm 3. Description of Transfer Algorithm of Fuzzy Clustering Hierarchical Generation Evolutionary Tree(based on the new cluster matrix)

1. Select Cut Set

(1) Selects the different element values of the matrix Q to construct a cut set $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_n\}(\lambda_1 < \lambda_2 < \cdots < \lambda_n,)$, n is the size of the cluster sample

2. Generating Evolutionary Tree by Fuzzy Hierarchical Clustering

(1) $i = l,\ bef = -1, af = -1$

(2) Take the confidence level $\lambda = \lambda_1$, $c_{i-1} = 0$;

(3) Initialize the cluster sample $X = (x_1, x_2, x_3, \dots, x_n)$;

(4) Take $x = X\ (1)$, k =1; $A[k] = \emptyset$, AA =$\emptyset$;

(5) Find out that all the similarity with the sample x_i is not less than λ of x samples; If there is no sample x_i in $A[k]$, $A[k] = A[k] + x_i$, AA=AA+x_i;

(6) When AA is not empty, take x=AA(1), AA=AA-x; Jump to (4);

(7) When AA is empty, the result is the next cluster sample $A[k]$ with the confidence level λ , $X = X - A\ [k]$;

(8) If X is not empty, k=k+l, jump to (3);

(9) When X is empty, c_i = k, all clustering samples under the confidence level λ are $AK = A[1] + \cdots + A[k]$;

(10) If c_{i-1}=c_i, af=$\lambda(i)$, i=i+l, jump to (2);

(11) If $c_{i-1} \neq c_i$, output clustering result is between the confidence level bef to af, the clustering result is $A[k]$; $bef = \lambda(i)$

4.3 Experimental Results and Discussion

In order to test the effectiveness of fuzzy clustering transfer algorithm based on spectrogram theory in constructing biological phylogenetic tree, three different groups of data were used for experiments. And all 3 sets of experiment are implemented via Matlab.

4.3.1 Clusters the First Exon of β Globulin Gene of 11 Species to Generate Phylogenetic Tree

4.3.1.1 Experimental Data

In order to compare the difference between the traditional fuzzy clustering transfer algorithm and the fuzzy clustering transfer algorithm based on spectrogram theory in constructing biological evolutionary trees, the coding sequence of the first exon of β globulin gene of 11 species in chapter 3 is taken as experimental data (see Table 3.1),) to construct evolutionary trees respectively.

4.3.1.2 Experimental Results and Discussion

Here, the graphic representation proposed in the third chapter is used to construct JZ curve group, and the parameter m= 5 and n= 1/11, JZ curve group is shown in figure 3.3. Then, the similarity matrix of the sequence is calculated by using the J/J matrix proposed in chapter 3 and the regularization matrix invariant λ/N, as shown in table 3.7. The transfer algorithm of traditional fuzzy clustering uses the fuzzy similarity matrix as shown in table 4.1, to select parameters c=0.01, the clustering results are shown in table 4.3, and figure 4.1, which give the structure diagram of evolutionary tree constructed according to the transfer algorithm of traditional fuzzy clustering; On the basis of fuzzy similarity matrix, a new clustering matrix is constructed according to algorithm 2. See Table 4.2, for parameter K=8. The clustering results are shown in Table 4.4, and Figure 4.2, which give the structure diagram of evolutionary tree constructed based on the new clustering matrix.

Table 4.1: Construct Fuzzy Similarity Matrix of 11 Species

	Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
Human	1	0.1031	0.0952	0.0682	0.3155	0.2200	0.2627	0.1129	0.8725	0.1802	0.7744
Goat		1	0.0733	0.1335	0.1093	0.0911	0.1489	0.0970	0.1008	0.1512	0.0963
Opos.			1	0.0779	0.0981	0.1637	0.0824	0.1526	0.0908	0.0691	0.0962
Gallus				1	0.0749	0.0732	0.0775	0.0863	0.0671	0.0740	0.0655
Lemur					1	0.1881	0.1904	0.1001	0.4000	0.1686	0.2286
Mouse						1	0.1445	0.1666	0.1922	0.1072	0.2317
Rabbit							1	0.1149	0.2324	0.3361	0.2369
Rat								1	0.1017	0.0870	0.1189
Gorilla									1	0.1840	0.4342
Bovine										1	0.1579
Chim.											1

Table 4.2: Construct a New Cluster Matrix of β Globulin Genes of 11 Species Based on Map Theory

	Human	Goat	Opos.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
Human	1	-0.006	-0.005	-0.003	0.007	-0.035	0.179	-0.012	0.924	-0.112	0.992
Goat		1	-0.001	-0.002	0.004	-0.001	0.020	-0.001	-0.008	-0.011	-0.006
Opos.			1	0.002	0.001	-0.002	-0.001	0.000	-0.003	0.001	-0.005
Gallus				1	0.001	0.001	0.003	0.001	-0.003	-0.002	-0.003
Lemur					1	0.026	-0.135	0.007	0.380	0.078	0.007
Mouse						1	0.050	-0.002	-0.093	-0.029	-0.035
Rabbit							1	0,037	0.106	0.927	0.179
Rat								1	-0.017	-0.020	-0.012
Gorilla									1	-0.076	0.924
Bovine										1	-0.112
Chim.											1

Table 4.3: Cluster Results of β Globulin Genes of 11 Species Based on Fuzzy Similarity Matrix

confidence level λ	Clustering results										
0.065458 - 0.13346	{1	2	3	4	5	6	7	8	9	10	11};
0.14453 -0.15123	{4};	{1	2	3	5	6	7	8	9	10	11};
0.15257 -0.16374	{4};	{2};	{1	3	5	6	7	8	9	10	11};
0.1666	{4};	{2};	{3};	{1	5	6	7	8	9	10	11};
0.16855 -0.23172	{4};	{2};	{3};	{8};	{1	5	6	7	9	10	11};
0.23242 - 0.26267	{4};	{2};	{3};	{8};	{6};	{1	5	7	9	10	11};
0.31545 -0.33609	{4};	{2};	{3};	{8};	{6};	{7	10};	{1	5	9	11};
0.4000	{4};	{2};	{3};	{8};	{6};	{7};	{10};	{5	1	9	11};
0.43424 - 0.77441	{4};	{2};	{3};	{8};	{6};	{7};	{10};	{5};	{1	9	11};
0.8724	{4};	{2};	{3};	{8};	{6};	{7};	{10};	{5};	{11};	{1	9};
1	{4};	{2};	{3};	{8};	{6};	{7};	{10};	{5};	{11};	{1};	{9};

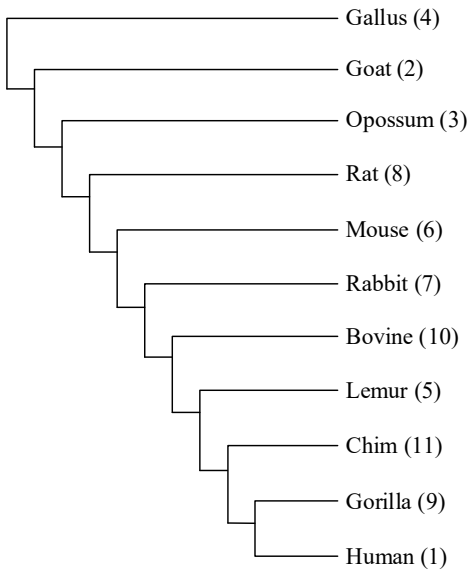

Figure 4.1: Phylogenetic tree obtained by clustering β globulin genes of 11 species based on fuzzy similarity matrix

Observing figure 4.1, it is found that Gallus was first isolated from several evolutionary species, indicating that Gallus is far related to other species, which is consistent with the fact that Gallus is the only non-mammal in 11 biological clocks. (Human, Gorilla, Chim.) these pairs of creatures were finally separated from several evolutionary species, indicating that they were relatively close in genetic relationship in evolutionary history. This is basically consistent with the results generated from chapter 3 and the biological significance (which is labelled in the database in the form of manually generated evolutionary tree). However, Goat was separated from the evolutionary tree before Opos, which is contrary to the result in chapter 3(in which the sum value of Goat is 0.9490 while Opossum is 1.0785) and the fact that Opos is the farthest related species of mammals. From the above, it can be seen that fuzzy hierarchical clustering based on fuzzy matrix has shortcomings.

Table 4.4: Cluster Results of 11 Species β globulin Genes Based on New Clustering Matrix

confidence level λ	Clustering results										
-0.25533 - 0.0047971	{1	2	3	4	5	6	7	8	9	10	11};
0.0065144 - 0.0069031	{4};	{1	2	3	5	6	7	8	9	10	11};
0.013893 - 0.020316	{4};	{3};	{1	2	5	6	7	8	9	10	11};
0.025779 - 0.037072	{4};	{3};	{2};	{1	5	6	7	8	9	10	11};
0.050196 - 0.092088	{4};	{3};	{2};	{8};	{1	5	6	7	9	10	11};
0.10632 - 0.24773	{4};	{3};	{2};	{8};	{6};	{1	5	7	9	10	11};
0.3804	{4};	{3};	{2};	{8};	{6};	{7	10};	{1	5	9	11};
0.77509 - 0.92446	{4};	{2};	{3};	{8};	{6};	{7	{10};	{5};	{1	9	11};
0.9274	{4};	{2};	{3};	{8};	{6};	{7	{10};	{5};	{1	11};	{9};

0.9550	{4};	{2};	{3};	{8};	{6};	{7};	{10};	{5};	{1	11};	{9};
1	{4};	{2};	{3};	{8};	{6};	{7};	{10};	{5};	{1};	{11};	{9};

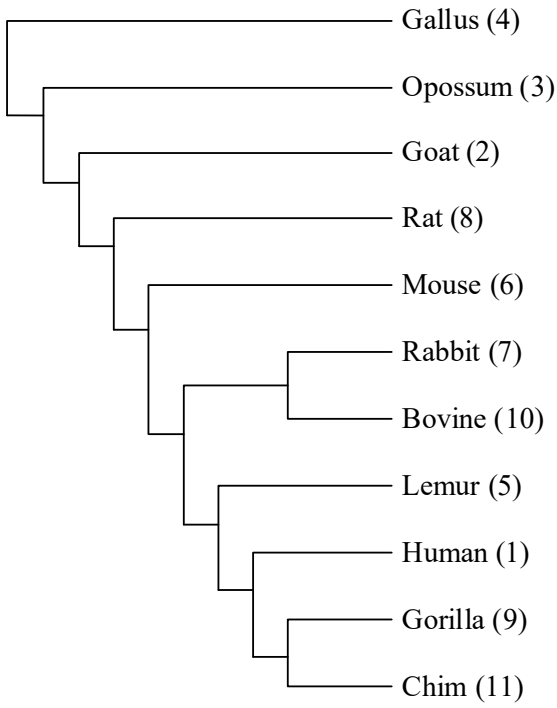

Figure 4.2: Generates Phylogenetic Tree by Clustering β Globulin Genes of 11 Species Based on a New Clustering Matrix

Looking at figure 4.2, it is found that Gallus was first isolated from the evolutionary tree, which is consistent with the fact that Gallus is the only non-mammal in 11 biological clocks. {Human, Gorilla, Chim.} these pairs of creatures were finally separated from several evolutionary species, indicating that they were relatively close in genetic relationship in evolutionary history. This is basically consistent with the reality. At the same time, compared with figure 4.1, it is found that Opos is isolated from the evolutionary tree before Goat, which conforms to the fact that Opos is the farthest mammalian species and is consistent with the results from chapter 3. From the above, it can be seen that the transfer method of fuzzy clustering based on atlas theory can effectively correct this error, and the obtained results are closer to the facts.

4.3.2 Clusters 11 H1N1 Virus NA Genes to Generate Phylogenetic Tree

4.3.2.1 Experimental Data

This group of experiments were carried out from the NCBI database (http://www.ncbi.nlm.nih.gov/) Download 11 Influenza a NA gene. Details of sequence number, access number, length, year and virus gene name are shown in Table 4.5.

4.3.2.2 Experimental Results and Discussion

In this article, the JZ curve group is constructed by using the graphic representation proposed in the third chapter, and the JZ curve group graphs are formed by selecting parameters m=l, n= l, JZ (1) curve, JZ (2) curve and JZ (3) curve, as shown in Figure 4.3, Figure 4.4 and Figure 4.5 respectively. Then the evolutionary tree is constructed by the transfer method of fuzzy clustering based on the graph theory, in which the new clustering matrix constructed based on the graph theory is shown in Table 4.6, and the parameters C=0.02 and K=5 are selected, and the clustering results are shown in Table 4.7. Finally, the structure diagram of the phylogenetic tree constructed according to the new clustering matrix of 11 H1N1 virus NA gene sequence is given as shown in Figure 4.6.

Table 4.5: 11 H1N1 Virus NA Gene Sequence

No	Accession	Length	Year	Virus Name
1	CY022063.1	1056	1976	Infuenza A virus (A/swine/Tennessee/19/1976 (H1N1))
2	AM777828.1	1056	2005	Infuenza A virus (A/swine/Cotes d*Armor/0227/2005 (H1N1))
3	AM777825.1	1056	2001	Infuenza A virus (A/swine/Cotes d 'Armor/98574/2001 (H1N1))
4	AM777830.1	1056	2005	Infuenza A virus (A/swine/Cotes d*Armor/016007/2005 (H1N1))
5	AM777829.1	1056	2006	Infuenza A virus (A/swine/Cotes d*Armor/002007/2006 (H1N1))
6	GQ150330.1	1056	2009	Infuenza A virus (A/swine/Alberta/OTH-33-8/2009 (H1N1))
7	GQ369426.1	1056	2009	Infuenza A virus (A/swine/Alberta/OTH-33-24/2009 (H1N1)
8	CY022359.1	1056	1976	Infuenza A virus (A/swine/Minnesota/27/1976 (H1N1))
9	CY022335.1	1424	1988	Infuenza A virus (A/swine/Iowa/17672/1988 (H1N1))
10	GQ369420.1	1056	2009	Infuenza A virus (A/swine/Alberta/OTH-33-22/2009 (H1N1))
11	AM777826.1	1056	2001	Infuenza A virus (A/swine/Cotes d1 Armor/60293/2001 (H1N1))

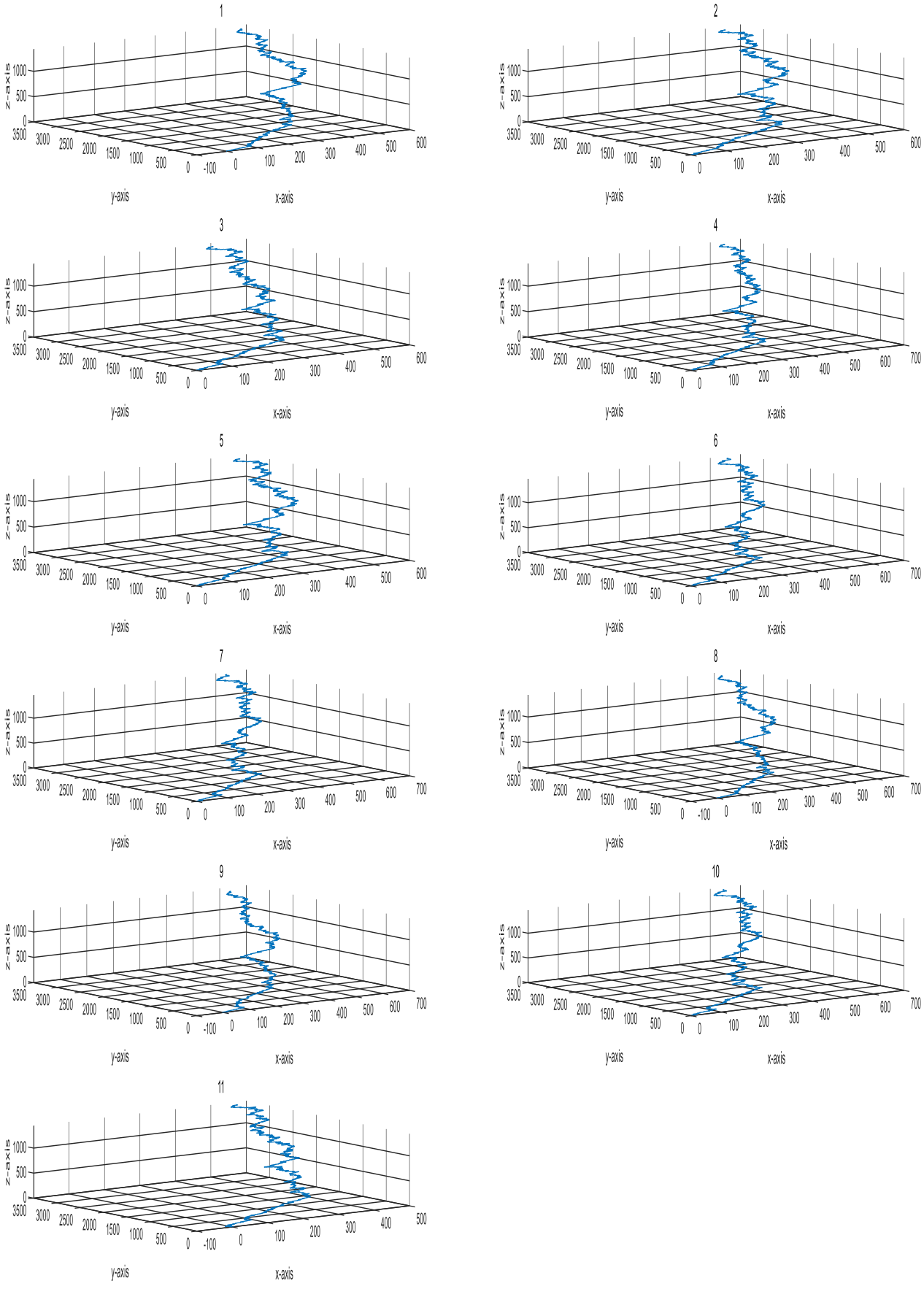

Figure 4.3 JZ (1) curves of 11 H1N1 virus sequences

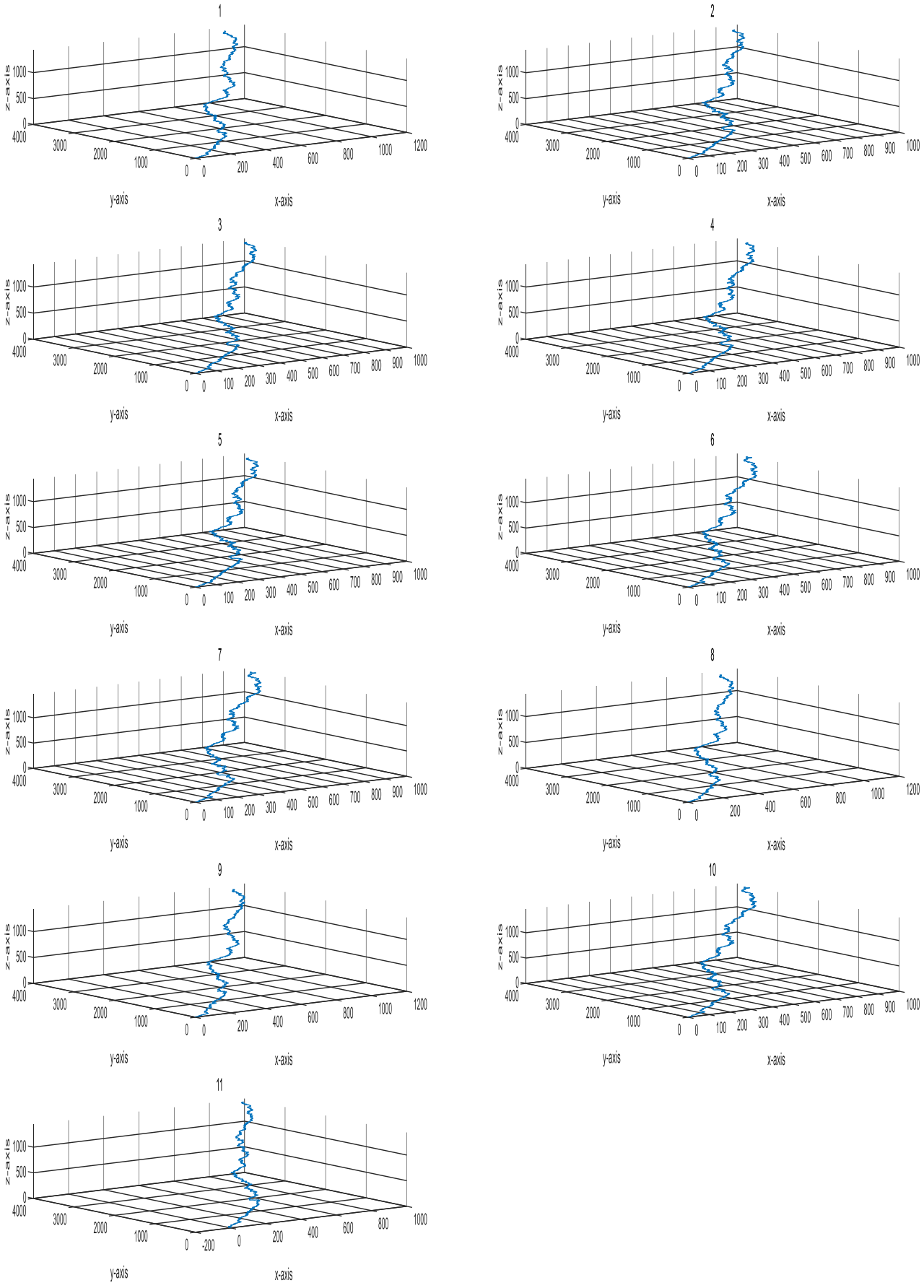

Figure 4.4 JZ (2) curves of 11 H1N1 virus sequences

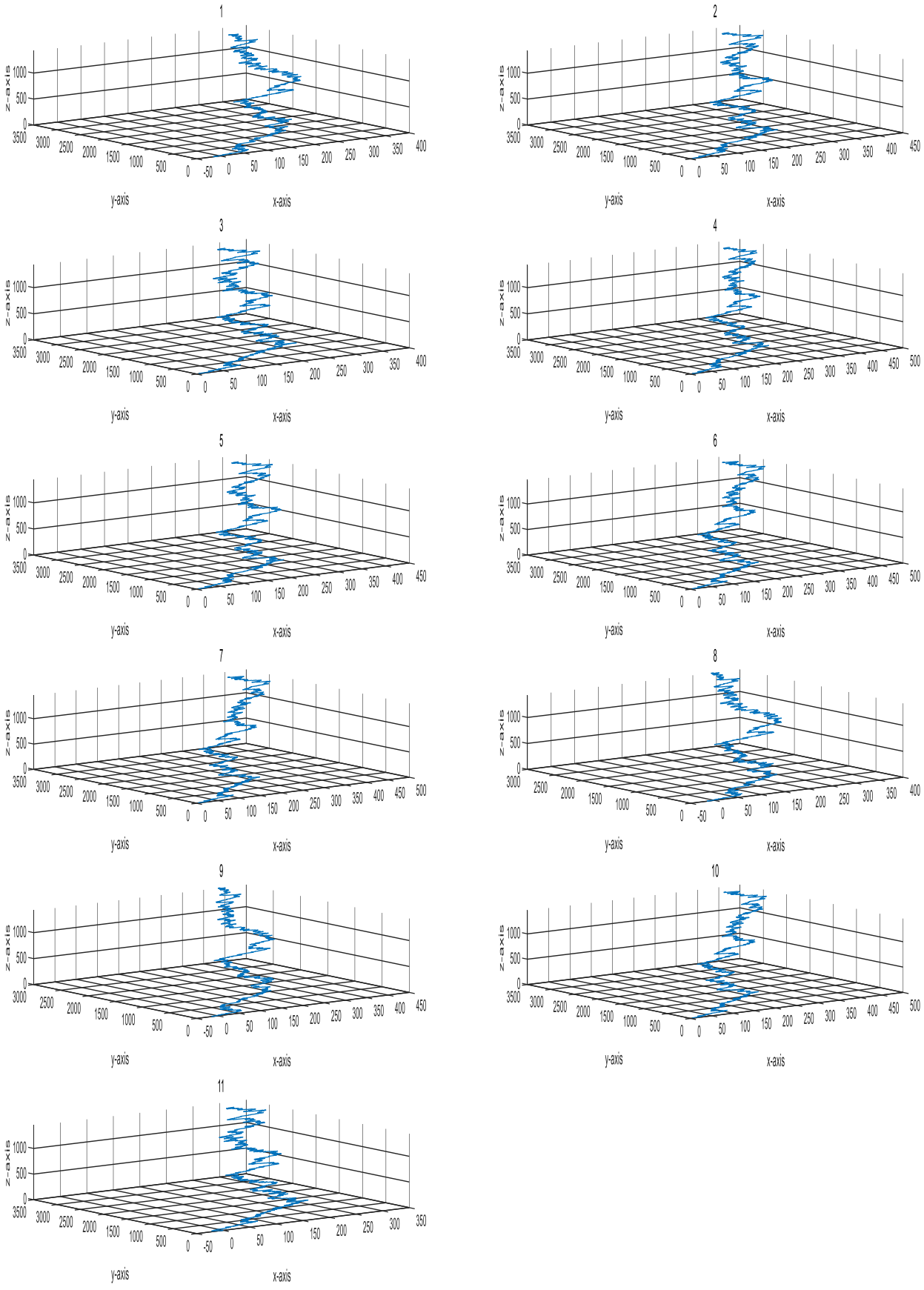

Figure 4.5 JZ (3) curves of 11 H1N1 virus sequences

Table 4.6: Construct a New Clustering Matrix of 11 H1N1 Viruses Based on Atlas Theory

	1	2	3	4	5	6	7	8	9	10	11
1	1	0.0019	-0.0011	-0.0029	-0.0020	-0.0034	-0.0027	0.9736	0.1275	0.0004	-0.0014
2		1	0.1106	0.9992	0.9935	-0.0060	0.1967	0,0034	0.0026	0.0124	-0.1272
3			1	0.1274	0.2164	-0.0148	0.0941	0.0018	-0.0006	0.0172	0.9717
4				1	0.9959	-0.0413	0.1629	0.0004	-0.0056	0.0477	-0.1103
5					1	-0.0454	0.1655	0.0015	-0.0057	0.0519	-0.0202
6						1	0.9697	0.0042	-0.0174	1.0000	-0.0102
7							1	-0.0190	0.0886	0.9681	0.0503
8								1	-0.1021	0.0072	0.0011
9									1	-0.0176	-0.0011
10										1	-0.0111
11											1

Table 4.7: Cluster Results of 11 H1N1 Viruses Based on New Clustering Matrix

confidence level λ	Clustering results										
-0.12716 - 0.088587	{1	2	3	4	5	6	7	8	9	10	11};
0.094047-0.12754	{1	8	9};	{2	3	4	5	6	7	10	11);
0.16288 - 0.19671	{1	8};	{9};	(2	3	4	5	6	7	10	11};
0.2163	{1	8};	{9};	(2	3	4	5	11};	{6	7	10};
0.06814-0.96967	{1	8};	{9};	{2	4	5};	{3	11};	{6	7	10};
0.9717	{1	8};	{9};	{2	4	5};	{3	11};	{6	10};	{7};
0.9736	{1};	{8};	{9};	{2	4	5};	{3};	{11}	{6	10};	{7};
0.99348 - 0.99592	{1};	{8};	{9};	{2	4	5};	{3};	{11}	{6	10};	{7};
0.9992	{1};	{8};	{9};	{2	4};	{5};	{3};	{11}	{6	10};	{7};
0.9997	{1};	{8};	{9};	{2};	{4};	{5};	{3};	{11}	{6	10};	{7};
1	{1};	{8};	{9};	{2};	(4);	{5};	{3};	{11}	{6};	{10}	{7};

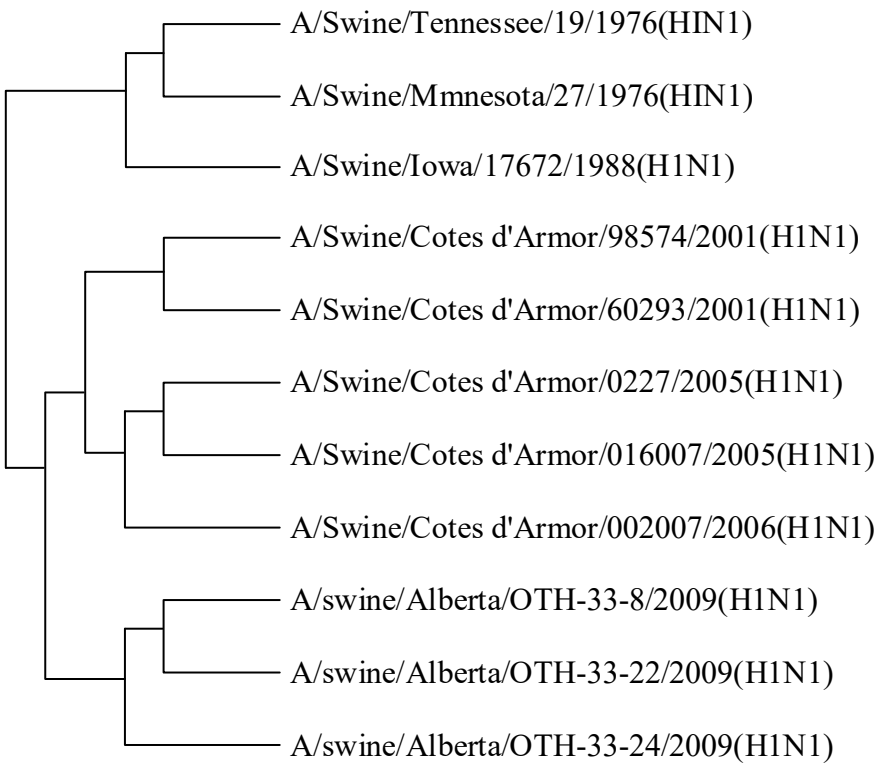

Figure 4.6: is an evolutionary tree generated by clustering 11 H1N1 viruses based on a new clustering matrix.

In this group of experiments, the phylogenetic tree of 11 H1N1 influenza neuraminidase (NA) genes is constructed. Observing the figure 4.4, it is intuitively found that in the first branch of the phylogenetic tree, the H1N1 influenza virus {1,8,9} in the 20 century and the H1N1 influenza virus 90 era are clustered into one group, and the H1N1 influenza virus {2,3,4,5,6,7,10,11} in the early 21 century are clustered into one group, which is consistent with the pathological inference that the virus changes greatly every ten years. At the same time, through observation, it is found that 1976 year H1N1 influenza virus {1, 8} begin separation after threshold 0.9736, 2001 H1N1 influenza virus {3, 11} after the threshold 0.9717, 2005 year H1N1 influenza virus {2, 4} after the threshold 0.9992, 2009 year H1N1 influenza virus {6, 7, 10} after the threshold 0.9717, which shows that the same year virus has high similarity and high homology, which is consistent with pathology, that is, the natural variation of the virus is slow, and the same year virus has high similarity. The experiment shows that the transfer algorithm of fuzzy clustering based on spectrogram theory is effective in constructing biological phylogenetic trees.

4.3.3 Clusters 8 H1N1 virus NA gene sequence to Generate Phylogenetic Tree

4.3.3.1 Experimental Data

This group of experiments were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov/) which is a 8 HINI NA gene fragment. Details of sequence number, access number, length, year and virus gene name are shown in Table 4.8.

4.3.3.2 Experimental Results and Discussion

Here, JZ curve group is constructed by using the graphic representation proposed in the third chapter, and parameter m = 1, n=1, JZ (1), JZ (2), JZ (3) are selected to form curve group graphs, which are respectively shown in figure 4.7, figure 4.8 and figure 4.9; Then, the evolutionary tree is constructed by the transfer method of fuzzy clustering based on the atlas theory, wherein the new clustering matrix constructed based on the atlas theory is shown in the table 4.9, selection parameters C=0.001 , K=4 , and the clustering results are shown in the table 4.10; Finally, the structure diagram of the phylogenetic tree generated by clustering 8 H1N1 viruses based on the new clustering matrix is given, as shown in Figure 4.10.

Table 4.8: 8 H1N1 virus NA gene sequence

No	Accession	Species	Year	Virus Name
1	CY022359	swine	1976	(A/swine/Minnesota/27/1976 (H1N1))
2	CY022335	swine	1988	(A/swine/Iowa/17672/1988 (H1N1))
3	CY022972	swine	1988	(A/swine/Iowa/31483/1988 (H1N1))
4	AB434298	swine	2003	(A/swine/Ratchaburi/NIAH550/2003 (H1N1))
5	GQ247842	duck	2005	(A/duck/Italy/1447/2005 (H1N1))
6	GQ150330	swine	2009	(A/swine/Alberta/OTH-33-8/2009 (H1N1)
7	GQ369426	swine	2009	(A/swine/Alberta/OTH-33-24/2009 (H1N1))
8	GQ351316	human	2009	(A/Hong Kong/2369/2009 (H1N1))

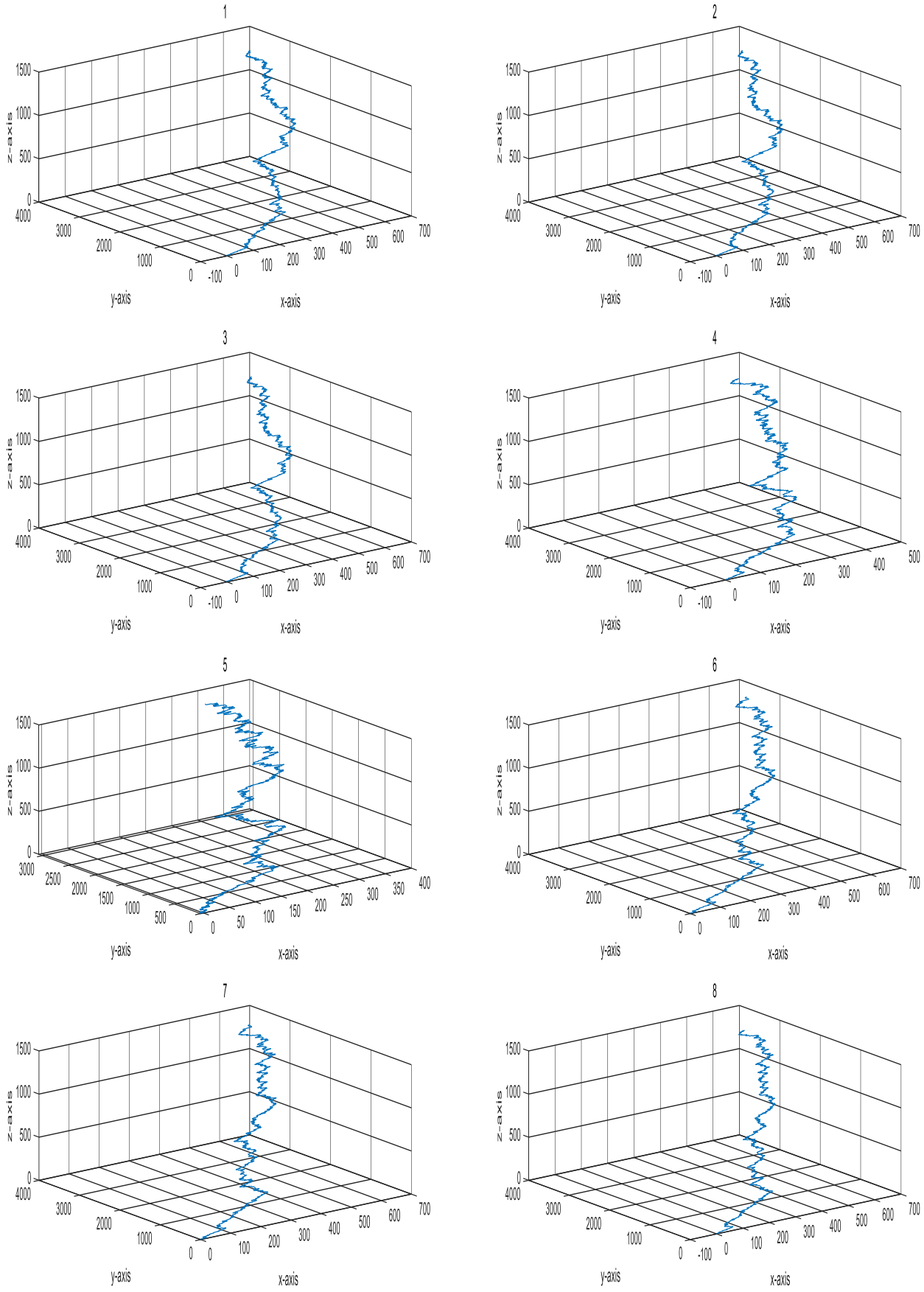

Figure 4.7 JZ (1) Curve of 8 H1N1 Virus Sequences

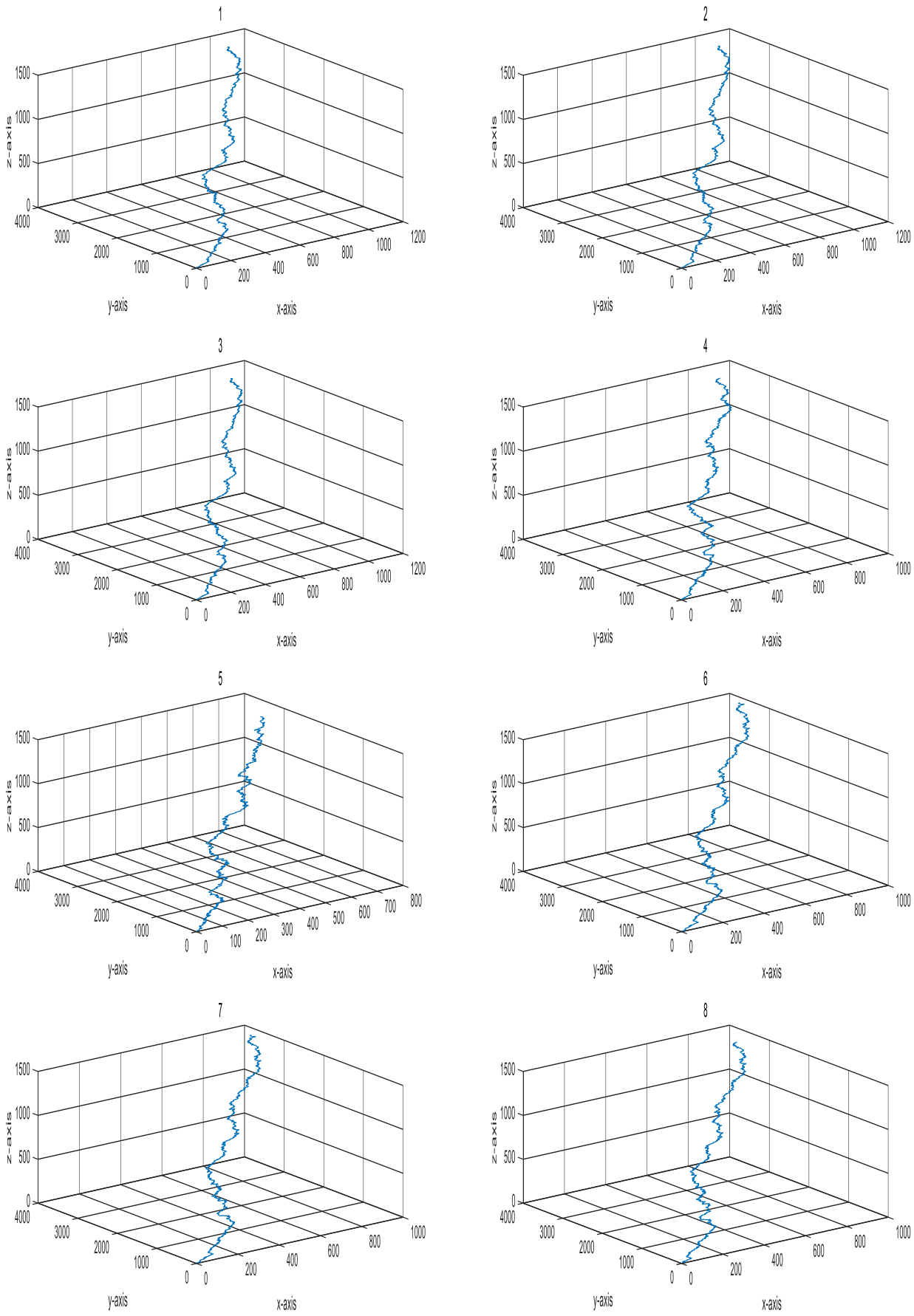

Figure 4.8 JZ (2) Curve of 8 H1N1 Virus Sequences

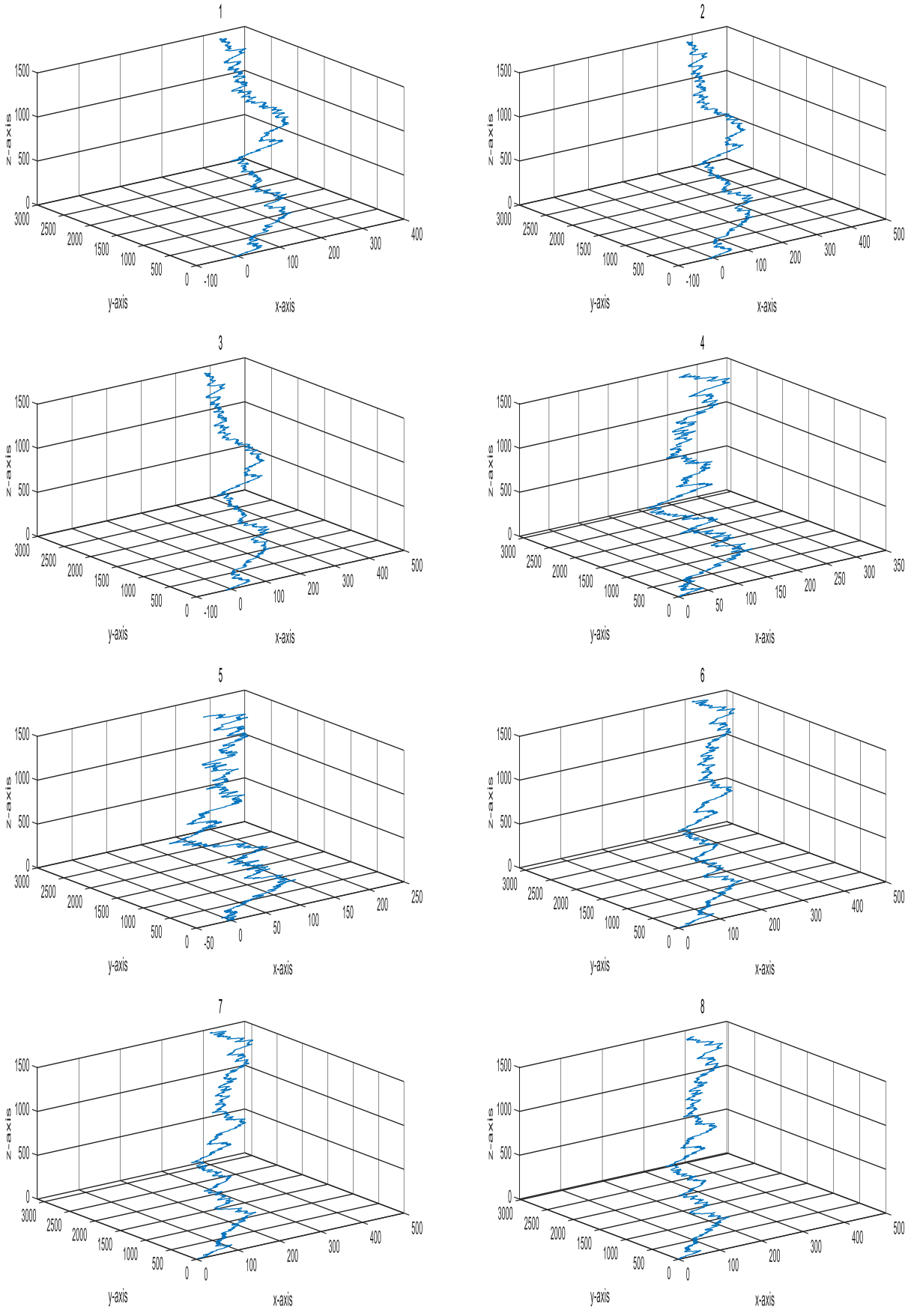

Figure 4.9 JZ (3) curves of 8 H1N1 virus sequences

Table 4.9: Construct a New Clustering Matrix of 8 H1N1 Viruses Based on Atlas Theory

	1	2	3	4	5	6	7	8
1	1	0.04494	0.03291	-0.29773	0.01518	0.03299	0,03341	0.99773
2		1	0.99992	0.80766	-0.02478	-0.09383	-0.10005	0.11071
3			1	0.81283	-0.02517	-0.09167	-0.09790	0.09876
4				1	0.09854	0.37712	0.37144	-0.24248
5					1	-0.01124	-0.01203	0.00079
6						1	0.99998	0.03095
7							1	0.03097
8								1

Table 4.10: Cluster Results of 8 H1N1 Viruses Based on New Clustering Matrix

confidence level λ	Clustering results							
-0.29773 - 0.09854	{1	2	3	4	5	6	7	8};
0.098756 -0.11071	{5};	{1	2	3	4	6	7	8};
0.37144 - 0.37712	{5};	{1	8};	{2	3	4	6	7};
0.80766 - 0.81283	{5};	{1	8};	{2	3	4};	{6	7};
0.9999	{5};	{1	8};	{2	3};	{4};	{6	7};

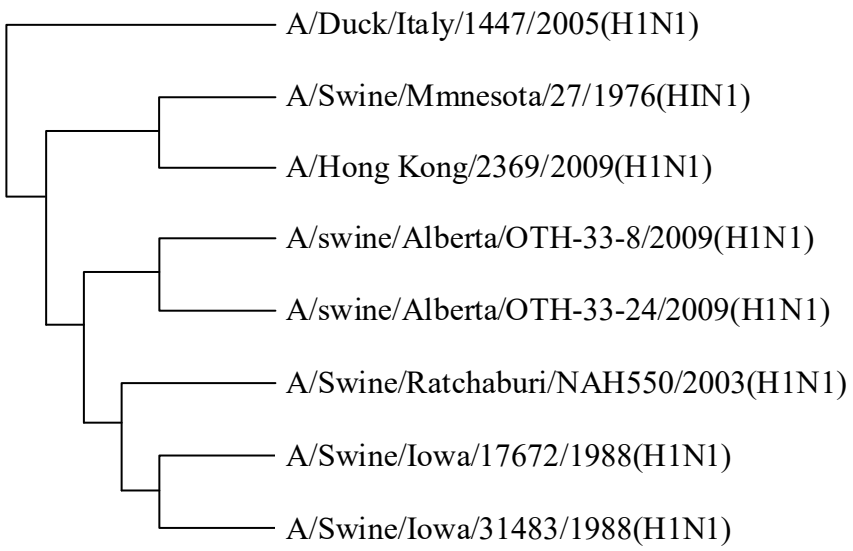

Figure 4.10: an evolutionary tree generated by clustering 8 H1N1 viruses based on a new clustering matrix.

In this experiment, the 8 H1N1 virus infected with pigs, human beings and livestock was constructed based on fragments. Observing figure 4.7, it is intuitively

found that at the first branch of the phylogenetic tree, the H1N1 influenza virus infected ducks, the H1N1 influenza virus infected humans and the H1N1 influenza virus infected pigs are obviously divided into two subtrees. This shows that the NA fragment of HINI infected with human beings has extremely high homology with the NA fragment of H1N1 infected with pigs, and relatively low homology with the NA fragment of H1N1 infected with livestock. This is consistent with the results of pathological studies, namely: Pigs and humans belong to the mammalian family and are easy to become the same or highly similar virus hosts, while ducks are avian animals, and the similarity between the virus infecting ducks and the virus infecting humans is low. Through observation, it is found that the H1N1 virus infected by 2009 in human and the H1N1 virus infected by 1976 in pigs have high homology, which is consistent with the virus gene sequencing results published at present, i.e. The influenza has a similarity of about 96% with the traditional swine influenza virus (classical swine influenza virus H1N1, which has also become the main reason why it was named swine influenza at the beginning. At the same time, we found that the homology between the H1N1 virus infected by 2009 and the H1N1 virus infected by 2009 was low. This is consistent with the fact that it is more scientific for people to change the name of the swine flu to influenza a H1N1 , that is, with the gradual deepening of understanding of this new influenza virus, people have found that even in Mexico, the birthplace of the influenza a H1N1 epidemic, no pandemic of pig disease has been found so far. At the same time, the flow of people who feel infected is more people-to-people contact infection.

4.4 Summary

In this article, a fuzzy hierarchical clustering algorithm based on spectrogram theory is proposed to construct biological evolutionary tree. In the process of sequence numerization, the graphic expression method is adopted. Firstly, the sequence is converted into JZ curve group, which intuitively depicts the characteristics of the sequence, but the graphic cannot quantitatively describe the evolution distance between the sequences. Next, in the process of constructing the clustering matrix, we combine the advantages of spectrogram theory and fuzzy analysis, considering not only the degree of dispersion between classes, but also the degree of compactness of the same class. Then we construct the cut set of fuzzy clustering with the element values of the

matrix, and construct the clustering results from top to bottom hierarchically to guide the production of constructing evolutionary trees, which clearly reflects the evolutionary relationship between each sequence. Experimental results and analysis show that the new method improves the accuracy of clustering compared with the results obtained through the fuzzy similarity matrix, which is shown in the discussion between the results from figure 4.1 and figure 4.2. This algorithm is a deeper research on the basis of chapter 3. It combines coding sequence, spatial curve, matrix and matrix characteristic parameters, atlas theory in machine learning and fuzzy clustering algorithm in data mining to study evolutionary tree construction algorithm from a new angle. In this chapter, fuzzy clustering algorithm and fuzzy hierarchical clustering method based on atlas theory are respectively adopted to construct biological evolution tree for the coding sequence of the first exon of β -globin gene of 11 organisms, and the effectiveness of fuzzy hierarchical clustering method based on atlas theory and fuzzy hierarchical clustering algorithm is compared. At the same time, this paper adopts fuzzy hierarchical clustering method based on atlas theory to construct phylogenetic tree for the full-length nucleotide sequence and coding protein amino acid sequence of the new type a H1N1 virus NA gene and carry out corresponding analysis, so as to provide reference for the establishment of effective vaccine for the new type a H1N1 influenza virus. The biggest limitation of this algorithm is that the branch length of the constructed evolutionary tree cannot represent the actual evolution time, but is only a representation of relative distance. This aspect needs to be further deepened in future research.

Chapter 5.Conclusions and Future Work

5.1 Summary of the conclusions

With the continuous development of various genome projects, more and more biological data have been generated. Only after these massive biological data are scientifically collected, analyzed and sorted out can we obtain the internal mystery of gene sequences of species. Under such circumstances, this thesis takes bioinformatics and data mining as the research background and studies and discusses the similarity of DNA sequences and the construction algorithm of biological evolutionary tree based on graphical representation.

Firstly, the development frontier of bioinformatics is summarized, and bioinformatics, biological data mining technology, sequence graph representation method and biological evolution relationship are reviewed. The main process of existing biological data mining is discussed, which lays a foundation for further research.

Secondly, the similarity analysis of DNA sequences and evolutionary tree construction algorithms are reviewed. Firstly, the 2-dimensional, 3-dimensional and high-dimensional graphic representation methods of DNA sequences are introduced, and the characteristics and advantages of various graphic representation methods are analyzed. Then the sequence similarity analysis based on graphic representation is introduced in detail, with emphasis on the calculation methods of eigenmatrix, eigenvalue of matrix and distance vector.

Thirdly, the novel work of this thesis is introduced:

(1) On the basis of Z curve, considering the graphic representation with parameters, a new graphic curve - JZ curve group is presented. It is proved that there is no loop in JZ curve group, and JZ curve group contains some biological characteristics.

(2) The three-dimensional graphics of DNA sequences are applied to quantitative analysis of sequence similarity, the transformation from graphics to matrix is studied, and a new sequence feature matrix J/J matrix is proposed. It not only describes the chemical properties of sequence bases, but also extracts the biological significance of gene sequences. The DNA sequence of the first exon of β globulin gene of 11 common species has been used for experimental verification. By adapting the new J/J matrix, the

experiment results show that several errors that L/L matrix brings can be effectively corrected. For example, compared with the result generated from L/L matrix, Human, Chim and Gorilla can be easily divided into the same genus by their smallest difference index, significantly widen the gap with Gallus and Opos. For the difference index between these species, significant increase has been brought from 19.89% to up to 585.51%. Also the sum of similarities of Gallus and other 10 species under J/J matrix shows an +20%-60% difference rate compared to other thesis using different variants of Z-curve and L/L matrix. This shows that the J/J matrix applied to the JZ curve group can simply and effectively analyse the similarity between sequences and fix the defects caused by other algorithms.

(3) Based on the graphical representation of DNA sequence, a transfer algorithm of fuzzy analysis clustering based on spectrogram theory is proposed to construct evolutionary tree hierarchically. By constructing the phylogenetic tree of the first exon of β globulin gene of 11 species, the new clustering method generates the more accurate evolutionary tree than the traditional clustering method by isolating Opossum before Goat on the evolutionary tree, which is in consistent with the experiment results in Chapter 3.

And in the construction of the phylogenetic tree of influenza virus HIN1, the evolutionary relationship between species is analysed and the feasibility of the algorithm is illustrated which shall help in the analysis and construction of nucleotide sequence and amino acid sequence of coding protein, or even tracing the source of new epidemic virus.

5.2 Future Work

Through the above research, we have a more in-depth understanding of graphic-based DNA similarity analysis and evolutionary tree construction algorithm, and further work prospects are as follows:

(1) A new graph curve- JZ curve group proposed in this article can avoid the phenomenon of graph loop and degeneration, and also consider the biological significance of sequence: genome sequence index. The J/J matrix combined with JZ curve group can simply and effectively compare the similarity between sequences. Therefore, the graph can only be used to observe the sequence globally, and the characteristics of the specific sequence need to be determined by calculating the

corresponding characteristic parameters. In the further research, more appropriate graphic parameters to solve the local mutation problem needs to be considered. For single nucleotide polymorphism (SNP), there has been several algorithms such as Smith-Waterman local comparison algorithm which can accurately determine the mutation locus. For structure variations which are more complicated than SNP, the quantity of calculation increases exponentially using local comparison algorithm. So the attempt to combine appropriate local comparison algorithm with JZ-curve will be the major research direction in the future.

(2) This thesis proposes an evolutionary tree construction algorithm based on graphical representation by combining spectrogram theory and fuzzy clustering transfer algorithm. Compared with the traditional closure algorithm of fuzzy clustering, this algorithm improves the accuracy of clustering and avoids the problem of multiple sequence alignment, but there is a gap between the branch length of the evolutionary tree constructed by the algorithm and the actual evolutionary distance, which is also a problem that needs to be further solved to extract more comprehensive evolutionary information from graphics.

References

[1] Mo Z, Zhu W, Sun Y, Xiang Q, Zheng M, Chen M, Li Z, One novel representation of DNA sequence based on the global and local position information. Scientific reports. 2018, 8(1):1-7.

[2] Zhong Yang, Zhao Liang, Zhao Qiong, Concise Bioinformatics. Beijing Higher Education Press, 2002.

[3] Mu Z, Li G, Wu H, Qi X. 3D-PAF curve: a novel graphical representation of protein sequences for similarity analysis. Match Commun Math Comput Chem. 2016, 75:447-62.

[4] J. W. Han, M. Kamber, Data Mining: Concepts and Techniques. Simon Fnsier University: Morgan Kaufoimm Publishers, 2000,12-20, 236-245.

[5] Zhang Chunting, Current Situation and Prospect of Bioinformatics. World Science and Technology Research and Development, 2000, 22 (6): 17-25

[6] Zhang Chunting, Analysis of DNA sequence by geometric method, China Science Foundation, 1999, 13 (3): 152-153.

[7] Lu Weiping, Zhou Yuanguo, Current Situation and Prospect of Bioinformatics. Foreign Journal of Clinical Biochemistry and Laboratory, 2002,23(5):254-255, 274.

[8] C. Zhang, "Current Situation and Prospect of Bioinformatics," World Science and Technology Research and Development, 2000, 22: 17-20.

[9] D. J. Parry-Smith. T. K. Attwood, Introduction to bioinformatics, 1999, pp. 168-196.

[10]Gao Y. A multiple sequence alignment algorithm based on inertia weights particle swarm optimization. Journal of Bionanoscience. 2014 ,8(5):400-4.

[11]W. R. Pearson, "Selecting the Right Similarity‐Scoring Matrix," Current Protocols in Bioinformatics, 43: 1-9.

[12]J. Gibbs, G. A. McIntyre, “The Diagram a Method for Comparing Sequences its Use with Amino and Nucleotide Sequences," Eur J Biochem, 1970, 16: 1-11.

[13]D. Wunsh, S. B. Needleman, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins," Journal of Molecular Biology, 1970, 48: 443-453.

[14]M. S. Waterman, T. F. Smith, "Identification of Common Molecular Subsequences," J Mol Biol, 1981, 147: 195-197.

[15]J. S. Richardson, M. Murata, J. L. Sussman, "Simultaneous Comparison of Three Protein Sequences," Proc Natl Acad Sci, 1985, 93: 3073-3077.

[16]Gates M.A, A simple way to look at DNA. Journal of Theoretical Biology, 1986, 119: 319-328.

[17]Nandy A, A new graphic representation and analysis of DNA sequence structure, Methodology and application to globin genes. Current Science, 1994, 66: 309-313.

[18]Leong P M, Morgenthaler S, Random walk and gap plots of DNA sequences, Computer Application Bioscience, 1995, 11 (5): 503-50.

[19]Zhang C T, Zhang R, Ou H Y, The Z curve database: A graphic presentation of gene sequences. Bioinformatics, 2003, 19 (5): 593-599

[20]Zhang C. T, A Symmetric Theory of DNA sequences and its applications, Journal of Theoretical Biology, 1997, 187 (3): 297-306

[21]Liao B, Zhang Y S, Zhang K Q, Wang T M, Analysis of similarity/sensitivity of DNA sequences based on a confined curve representation. Journal of Molecular Structure: THEOCHEM, 2007, 717 (1-3): 199-203

[22]Liao B, 3-D graphic presentation of DNA sequences and their numerical characterization. Journal of Molecular Structure: THEOCHEM, 2004, 681 (1-3): 209-212.

[23]Q Xiang, K Feng, B Liao, Y Liu, G Huang, Combinatorial chemistry & high throughput screening, 2017, 20 (7), 622-628.

[24]Zhang Y, Liao B, Ding K, On 3DD-curves of DNA sequences. Molecular Simulation, 2006, 32 (1): 29-34.

[25]B Liao, Q Xiang, L Cai, Z Cao, A new graphical coding of DNA sequence and its similarity calculation, Physica A: Statistical Mechanics and its Applications, 2013, 392 (19), 4663-4667.

[26]Delibaş E, Arslan A, Şeker A, Diri B. A novel alignment-free DNA sequence similarity analysis approach based on top-k n-gram match-up. Journal of Molecular Graphics and Modelling. 2020 Nov 1; 100:107693.

[27]Gibbs A J, McIntyre G A. A Method for Assessing the size of a protein from its composition: It uses in evaluating data on the size of the protein subunits of plant virus particles. Journal of General Virology, 1970, 9: 51-67.

[28]Needleman S B, Wunsch C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 1970, 48 (3): 443-453.

[29]Smith T F, Waterman M S, Identification of common molecular subsequences. Journal of Molecular Biology, 1981, 147 (1): 195-197.

[30]Pearson W R, Lipman D J, Improved tools for biological sequence comparison, Proceedings of the National Academy of Sciences of the USA, 1988, 4: 2444-2448.

[31]Altschul S F, Gish W, Miller W, et al, Basic local alignment search tool. Journal of Molecular Biology, 1990, 215 (3): 403-410.

[32]Zhang Min, Research status and prospect of biological sequence alignment algorithms. Journal of Dalian University, 2004, 25 (4): 75-79.

[33]Wu R, Liu W, Mao Y, Zheng JZ. 2D Graphical Representation of DNA Sequences Based on Variant Map. IEEE Access. 2020, 8:173755-65.

[34]Jafarzadeh N, Iranmanesh A. A new graph theoretical method for analyzing DNA sequences based on genetic codes. MATCH-Commun. Math. Comput. Chem. 2016,75(3):731-42.

[35]Mizuta S. Graphical Representation of Biological Sequences. Bioinformatics in the Era of Post Genomics and Big Data. 2018, 2:109.

[36]Suna D, Xua C, Zhanga Y. A novel method of 2D graphical representation for proteins and its application. RNA. 2016,18:20.

[37]Li Y, Liu B, Cui J, Wang Z, Shen Y, Xu Y, Yao K. Similarities and evolutionary relationships of COVID-19 and related viruses. arXiv preprint arXiv:2003.05580. 2020.

[38]Nandy A, Harle M, Basak SC, Mathematical descriptors of DNA sequences: Development and applications. ARKIVOC, 2006, ix: 211-238.

[39]Keeling PJ. Genomics: evolution of the genetic code. Current Biology. 2016, 26(18):R851-3.

[40]Solovyev V, Galeev T. The problem of interpretation of phylogenetic trees. National Academy of Managerial Staff of Culture and Arts Herald. 2018(3).

[41]Sharma A, Jaloree S, Thakur RS. Review of Clustering Methods: Toward Phylogenetic Tree Constructions. InProceedings of International Conference on Recent Advancement on Computer and Communication 2018, 475-480. Springer, Singapore.

[42]David A, Morrison, Phylogenetic tree-building. International Journal for Parasitology, 1996, 26 (6): 589-617.

[43]Fitch W M, Margoliash E, Construction of phylogenetic trees. Science, 1967, 155: 279-284.

[44]Saitou N, Nei M, The neighbor-joining method: A new method for constructing phylogenetic trees, Molecular Biology and Evolution, 1987, 4 (4): 406-425.

[45]Felsenstein J, Evolutionary trees from DNA sequences: A maximum likelihood approach, Journal of Moleculer Evolution, 1981, 17 (6): 368-376.

[46]Carnin J H, Sokal R R, A method for deducing branching sequences in phylogeny, Molecular Biology and Evolution, 1965, 19: 31 1-326.

[47]Soler S, Gramazio P, Figàs MR, Vilanova S, Rosa E, Llosa ER, Borràs D, Plazas M, Prohens J. Genetic structure of Cannabis sativa var. indica cultivars based on genomic SSR (gSSR) markers: Implications for breeding and germplasm management. Industrial Crops and Products. 2017, 104:171-8.

[48]Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, Suh A. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014, 346(6215):1320-31.

[49]Yang J, Grünewald S, Wan XF. Quartet-net: a quartet-based method to reconstruct phylogenetic networks. Molecular biology and evolution. 2013, 30(5):1206-17.

[50]Santander-Jiménez S, Vega-Rodríguez MA, Sousa L. Multiobjective Frog-Leaping Optimization for the Study of Ancestral Relationships in Protein Data. IEEE Transactions on Evolutionary Computation. 2017, 22(6):879-93.

[51]Zhang R, Zhang CT. A brief review: The z-curve theory and its application in genome analysis. Current genomics. 2014, 15(2):78-94.

[52]JIN YP, LI BL. Research on Multisource Information Fused Logging Curves Layering Algorithm. Measurement & Control Technology. 2014(1):6.

[53]Li Gangcheng, Liu Zanbo, Zeng Qingguang, A method of constructing evolutionary tree based on fuzzy clustering. Computer Applications, 2009, 29 (3): 836-839.

[54]Nanda S J, Panda G. A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm and Evolutionary computation. 2014, 16:1-8.

[55]Zhou Y, Luo J, Yuan X, Qiu G, Qu H, Zheng Z. The Visual Analysis of Variables Statistic Based on Arabidopsis DNA Sequences. 2019.

[56]Zheng Huaxian, and Jeffrey Zheng. 2D Similarity Map of Multiple Coronavirus Gene Sequences. 2020.

[57]Liao B, Tan M S, Ding K Q, Application of 2D graphical presentation of DNA sequence, Chemical Physics Letters, 2005, 414 (4-6): 296-300.

[58]Yuan C X, Liao B, Wang T M, New 3D graphic presentation of DNA sequences and their numerical characterization, Chemical Physics Letters, 2003, 379 (5-6): 412-417.

[59]Liao B, Tan M S, Ding K Q, A 4D presentation of DNA sequences and its application, Chemical Physics Letters, 2005, 402 (4-6): 380-383.

[60]He P A, Yan L, Zhu T. A Graphical Representation of Protein Sequences and Its Applications. InProceedings of the Fourth International Conference on Biological Information and Biomedical Engineering 2020 ,1-6.

[61]Randic M, On 3-D Graphical Representation of Proteomics Maps and the Numerical Characterization. Journal of Chemical Information and Computer Sciences, 2001, 41 (5): 1339-1344.

[62]Randic M, Vmcko M, On the similarity of DNA primary sequences, Journal of Chemical Information and Computer Sciences, 2000, 40 (3): 599-606.

[63]Chen C, Xing D, Xie XS, inventors; Harvard College, assignee. Methods of Amplifying Nucleic Acid Sequences. United States patent application US 15/745,251. 2019.

[64]Dorn M, e Silva MB, Buriol LS, Lamb LC. Three-dimensional protein structure prediction: Methods and computational strategies. Computational biology and chemistry. 2014, 53:251-76.

[65]Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Molecular biology and evolution. 2015, 32(10):2798-800.

[66]Zhang Y, Jia C, Fullwood MJ, Kwoh CK. DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. Briefings in Bioinformatics. 2020.

[67]Sharma A, Jaloree S, Thakur RS. Review of Clustering Methods: Toward Phylogenetic Tree Constructions. InProceedings of International Conference on Recent Advancement on Computer and Communication 2018, 475-480. Springer, Singapore.

[68]Chen MJ, Dixon JE, Manning G. Genomics and evolution of protein phosphatases. Science signaling. 2017, 10(474).

[69]Fagin R, Lotem A, Naor M. Optimal aggregation algorithms for middleware. Journal of computer and system sciences. 2003, 66(4):614-56.

[70]Li H, Chen CP, Huang HP. Fuzzy neural intelligent systems: Mathematical foundation and the applications in engineering. CRC Press; 2018.

[71]Helgason CM, Jobe TH, De Leon O, Mazumdar D. A structural representation of anticipatory thought process using the example of clinical medicine and the physician. In2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) 2013,1-4.

IEEE.

[72]Bian Z, Ishibuchi H, Wang S. Joint learning of spectral clustering structure and fuzzy similarity matrix of data. IEEE Transactions on Fuzzy Systems. 2018, 27(1):31-44.

[73]Wang Y, Ma X, Lao Y, Wang Y. A fuzzy-based customer clustering approach with hierarchical structure for logistics network optimization. Expert Systems with Applications. 2014, 41(2):521-34.

[74]Wu Fubao, Li Qi, Song Wenzhong, Transfer method of fuzzy cluster analysis. Journal of Southeast University, 1999, 29 (2): 22-26.

[75]Li Gangcheng, Liu Zanbo, Zeng Qingguang, A method for constructing evolutionary trees based on fuzzy clustering. Computer Applications, 2009,29(3):836-839.

[76]Zhang Y C, Luo J W, Zhang X Z, Research of gene sequence similarity analysis based on graphical representation. Science Technology and Engineering, 2007, 7 (21): 5593-5599.

[77]Wang J, Zhang Y, Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation. Chemical Physics Letters, 2006, 423 (1-3): 50-53.

[78]Song J, Analysis of similarity of DNA sequences based on function of degree of disagreement. Computers and Applied Chemistry, 2007, 24 (6): 729-733.

[79]Donth W E, Hoffman A J, Lower bounds for the partitioning of graphs. IBM Journal of Research and Development, 1973, 17 (5): 420-425.

[80]Gao Yan, Gu Shiwen, Research on spectral clustering in machine learning. Computer Science, 2007, 34 (2): 201-203.

[81]Fiedler M. Algebraic connectivity of graphs, Czechoslovak Mathematical Journal, 1973, 23 (98): 298-305.

[82]Scott G, Longuet H H, Feature grouping by relocalisation of eigenvectors of the proximity matrix. In: Proceedings of the British Machine Vision Conference. Oxford: BMVC, 1990, 103-108.

[83]Weiss Y, Segmentation using eigenvectors: A unified view. In: International Conference on Computer Vision. Kerkyra: IEEE, 1999, 975-983.

[84]Jeffrey HJ. Chaos game representation of gene structure. Nucleic acids research. 1990 Apr 25;18(8):2163-70.

[85]Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series

especially suited for long DNA sequences. Journal of Biological Chemistry. 1983 Jan 25;258(2):1318-27.

[86]Guo X, Randic M, Basak S C. A Novel 2-D Graphical Representation of DNA Sequences of Low Degeneracy, Chemical Physics Letters. 2002, (350): 106-112.