Journal of Molecular Graphics and Modelling 76 (2017) 342–355

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Similarity/dissimilarity calculation methods of DNA sequences: A survey

Xin Jin [a], Qian Jiang [a], Yanyan Chen [b], Shin-Jye Lee [c,d], Rencan Nie [a], Shaowen Yao [c,*], Dongming Zhou [a,*], Kangjian He [a]

[a] School of Information, Yunnan University, Kunming, Yunnan Province, China
[b] School of Life Sciences, Yunnan University, Kunming, Yunnan Province, China
[c] School of Software, Yunnan University, Kunming, Yunnan Province, China
[d] Queens' College, University of Cambridge, Cambridge CB3 9ET, U.K.

ARTICLE INFO

Article history:
Received 25 May 2017
Received in revised form 17 July 2017
Accepted 18 July 2017
Available online 20 July 2017

Keywords:
DNA sequence analysis
Similarity analysis
Graphical representation
Evolutionary relationship
Feature extraction

ABSTRACT

DNA sequence similarity/dissimilarity analysis is a fundamental task in computational biology, which is used to analyze the similarity of different DNA sequences for learning their evolutionary relationships. In past decades, a large number of similarity analysis methods for DNA sequence have been proposed due to the ever-growing demands. In order to learn the advances of DNA sequence similarity analysis, we make a survey and try to promote the development of this field. In this paper, we first introduce the related knowledge of DNA similarities analysis, including the data sets, similarities distance and output data. Then, we review recent algorithmic developments for DNA similarity analysis to represent a survey of the art in this field. At last, we summarize the corresponding tendencies and challenges in this research field. This survey concludes that although various DNA similarity analysis methods have been proposed, there still exist several further improvements or potential research directions in this field.

1. Introduction

With the development of advanced sequencing technologies, a large volume DNA sequence data are available in database with very low cost, and the cost will continue to be lower [1,2]. Genetic information of different species is stored by the permutations and compositions of nucleic acids in DNA sequences; and the genetic diversity is the basis of biological diversity. However, it still is a challenge for biologists to understand the biological functions and significances of DNA primary sequence [3–6]. Similarity analysis of DNA sequence is a process to compare unknown DNA sequences with known ones for inferring the functions of unknown ones [7]; it also can provide evolutionary information of the same gene sequence in different species, which is basic approach to comprehend the biological information in DNA sequence [7,8]. Computational method for DNA sequences similarity analysis is one of the most important technologies in the emerging interdisciplinary field of bioinformatics.

In order to accurately and effectively analyze the similarities of DNA sequences, the method should consider the following critical problems: (i) How to effectively represent the DNA primary sequence by numeral sequence. (ii) How to obtain and select suitable invariants (descriptors) which can be regarded as the features of DNA sequences to characterize them according to the numeral sequence. (iii) How to effectively process DNA sequences with different length and keep its consistency [9,10]. In order to solve above three problems, this work considers these DNA similarities analysis methods should have the following abilities: (i) the methods should be able to avoid information loss when DNA primary sequence was transformed into numeral sequence, and the process should not generate artifacts or inconsistencies which can distract or mislead a human observer or any subsequent steps; (ii) the invariants should effectively represent the genetic information of different species, and it should effectively compress the information of DNA primary sequences and conveniently quantitative calculated its similarities; (iii) the methods should adapt to DNA sequences with different lengths, and they should take into account the local features and global features to effectively analyze DNA sequences.

In the last decades, many scholars have proposed lots of methods to analyze the similarity of DNA sequences to get the corresponding genetic information. According to different strategies, these methods could be classified several subcategories. For example, most of DNA similarities analysis methods could be classified into graphical representation and others schemes according

* Corresponding authors.
E-mail addresses: yaosw@ynu.edu.cn (S. Yao), zhoudm@ynu.edu.cn (D. Zhou).

http://dx.doi.org/10.1016/j.jmgm.2017.07.019
1093-3263/

to their numeral ways [33]; and the graphical representation based methods could be ulteriorly classified into several categories ranging from 2D to 3D and others according to the dimensions of space in which the sequences are plotted [11,33,39]. And these analysis approaches also could be regarded as pairs of mononucleotide, dinucleotides [4,23,26,32,51] and trinucleotide (triplet codon) based methods [3,27,33,52]. Besides, some of these methods took into account the chemical properties of single base or specific base compositions [12,23–25,32,39,38,48,50], but the others do not.

In all of these approaches, graphical representation based methods are a popular research tendency for DNA sequences similarities analysis and take up most of the whole field. The breakthrough started in 1983 due to the graphical representations of DNA sequences were first initiated by Hamori and Ruskin [7], and then it was expanded by Nandy and Randic in 1994 and 2003 [13,14]. Henceforth, graphical representation based schemes were widely used in DNA sequences similarity analysis field [3], such as Hamori and Ruskin developed H-curve [11]. After that Liao et al. created L-curve to represent DNA sequences in a 3D space [15]. The advantages of graphical representation based methods are as follows: (i) it will allow visual inspection of DNA sequences, directly; (ii) it will help to recognize the major differences among different DNA sequences [31]. However, these methods still have some weakness, such as overlapping and intersections of the curves, which would lose some important information of DNA sequences and have not high resolution and accuracy [3].

In addition to graphical representation based methods, there are many other methods to analyze the similarity of DNA sequences. In Ref. [16,17,36,49], some approaches are proposed to represent DNA sequences by mapping nucleotide sequences into a DNA walk to get its numeral sequences; in Ref. [1,56,59,60], the bases of DNA sequences were transformed (or mapped) into four-number sequences, then the similarities information of the numeral sequences would be extracted by other given methods; a non-return-to-zero (NRZ) line code method, named coded mark inversion (CMI) [31], was also used to analyze DNA sequences similarities in Ref. [28]; besides, there were many other methods to digitize DNA sequences for their similarities analysis, such as the combinations of chemical properties and 2-compositions of four bases [51], 64 triplets based 4×4 matrices [9], Lempel-Ziv complexity based method [10] and so on.

DNA sequences have various lengths: the shorter gene sequence may have 10^2 bases and some of them may be 10^5 or more bases, the changing lengths of different DNA sequences present a big challenge for their similarities analysis. Unfortunately, some of these methods cannot effectively deal with the short sequence and long sequence simultaneously, because their poor consistency. The similarities of DNA sequences were used to learn their biological function and evolutionary information; accordingly, their similarities would be represented by their biological mechanism; however, how to reasonably take the biological properties and chemical properties into consideration is another problem. Besides, how to effectively identify a few site mutations in DNA sequences similarities analysis is also a need to be focused problem. For the moment, most of DNA sequences analysis methods were not perfect, such as high computational complexity and large processing memory for long DNA sequences. In addition, the often-used DNA sequence similarities analysis tools by biologists still have some weakness [18,19], such as low precision and time consuming. And the automatic performance of most methods is weak which should be improved in the future researches. More accurate and convenient similarity analysis methods can more effectively help us find the functions information and evolutionary information of unknown DNA sequences [7]. So there are lots of works should be done in DNA sequences similarities analysis field.

Motivated by the ideas mentioned above, this paper provides a review on the development of DNA sequence similarity analysis methods, and it is mainly focused on the progress of the past decades. This survey first introduces the related knowledge of DNA similarities analysis. Then, it reviews recent advances for DNA similarity analysis. At last, it summarizes the corresponding tendencies and challenges in the research field. This survey concludes that the demands and the slow progress of DNA sequence similarity analysis methods would drive further advances in the field. And we hope that these presented guidelines will help nonspecialists and specialists to learn the critical progress in this field in recent years.

2. Related knowledge of DNA similarities analysis

This section introduces the related knowledge of DNA similarity analysis, including data sets, similarity distances, output data and effectiveness assessment, which are supposed to help researchers understand the foundations of DNA similarity analysis.

2.1. Data sets of DNA sequences

The experimental samples for testing the validity of DNA similarity analysis methods should be elaborately selected. However, there are only few available DNA sequences sets for testing the effectiveness of similarity analysis methods at present. More unfortunately, it still is a problem to find much more suitable DNA sequences sets which are widely adopted by most computer scientists and biologists to verify the performance of similarity analysis method. This also is a limitation for the development of this field [20].

2.1.1. Data 18: whole mitochondrial genomes of 18 eutherian mammals

The whole mitochondrial genomes include abundant genetic information of 18 eutherian mammals, which are frequently-used in the recent years [3]. In this set, the average, the longest and the shortest length of these sequences are 16572 bases, 17019 bases and 16 295 bases, respectively, as shown in Table 1.

2.1.2. Data15: the β-globin gene detailed information of 15 species

The sequences of three exons of β-globin gene from 15 species are most often-used DNA sequences. The three gene sequences include the first, second and third exon, and the average lengths of these sequences are 92 bases, 222 bases and 114 bases, respectively [20], which are listed in Table 2. Among of them, the first exons of β-globin gene of eleven different species were the most widely used DNA sequences. Their lengths are in the range of 86–105 bases which are listed in Table 3.

2.1.3. Data 8: DNA sequence of albumin from 8 different species

The DNA sequences of albumin from 8 different species were first used in Ref. [43], and there do not many works use this data set at present. The DNA sequences have slightly different lengths, from 1830 bases (human albumin) to 1803 bases (monkey), as shown in Table 4. Besides, there are some other data sets used in Ref. [1,44] only, including the NADH dehydrogenase subunit 4 genes of 12 primate species, 80 different human rhinovirus (HRV) genomes, influenza A virus neuraminidase (NA) gene and human papillomavirus (HPV) genomes of 12 genotypes, which could be considered to be the experimental samples in future researches.

2.2. Calculation methods of similarity distances

The calculation of the distances among DNA sequences are the foundation of DNA similarities analysis, and it is a key step to rep-

Table 1
The mitochondrial genome detailed information of 18 eutherian mammals.

No.	Species	GenBank accession No.	No.	Species	GenBank accession No.
1	Human	V00662	10	Harbor seal	X63726
2	Common chimpanzee	D38116	11	Gray seal	X72004
3	Pygmy chimpanzee	D38113	12	Cat	U20753
4	Gorilla	D38114	13	Fin whale	X61145
5	Orangutan	D38115	14	Blue whale	X72204
6	Gibbon	X99256	15	Cow	V00654
7	Baboon	Y18001	16	Rat	X14848
8	Horse	X79547	17	Mouse	V00711
9	White rhinoceros	Y07726	18	Platypus	X83427

Table 2
The β-globin gene detailed information of 15 species.

No.	Species	GenBank accession No.	No.	Species	GenBank accession No.
1	Human	U01317	9	Gorilla	X61109
2	Goat	M15387	10	Bovine	X00376
3	Opossum	J03643	11	Chimpanzee	X02345
4	Gallus	V00409	12	Sheep	DQ352470
5	Lemur	M15734	13	Mouflon	DQ352468
6	Mouse	V00722	14	European hare	Y00347
7	Rabbit	V00882	15	Muscovy duck	X15739
8	Rat	X06701			

Table 3
The DNA sequences of the first exon of β-globin gene of 11 different species.

No.	Species	Sequence
1	Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
2	Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
3	Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
4	Gallus	ATGGTGCACTGGACTGCTCAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
5	Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
6	Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
7	Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
8	Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
9	Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
10	Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
11	Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCA-GGTTGGTATCAAGG

Table 4
The DNA sequences of albumin from 8 different species.

No.	Species	EMBL-Bank accession No.	No.	Species	EMBL-Bank accession No.
1	human	CAA00606	5	monkey	AAA36906
2	cattle (serum albumin)	AAN17824.1	6	dog	CAB64867
3	cat	CAA59279	7	pig	AAT98610
4	rat	AAH85359	8	rabbit (serum albumin)	AAB58347

resent the results. Euclidean distances and correlation angles are the most often used distances calculation methods.

2.2.1. Euclidean distance

Suppose the $V_1 = (x_{i1}, y_{i1})$ and $V_2 = (x_{j2}, y_{j2})$ in 2-D space are two vectors of different DNA sequences, respectively. Euclidean distance can be defined as follows:

$$D_{12} = \sqrt{\sum_{ij} \left(x_{i1} - x_{j2}\right)^2 + \sum_{ij} \left(y_{i1} - y_{j2}\right)^2} \quad (1)$$

The smaller Euclidean distance between the sequences is, the more similar the DNA sequences are.

2.2.2. Correlation angle

The value of correlation angle represents the angle relation between two vectors. The same as the Euclidean distance hypothesis, the correlation angle of two vectors can be defined as follows:

$$\theta_{ij} = \arccos \frac{\sum_{ij} x_{i1}x_{j2} + \sum_{ij} y_{i1}y_{j2}}{\sqrt{\sum_{i} {x_{i1}}^2 + {y_{i1}}^2} \times \sqrt{\sum_{j} {x_{j2}}^2 + {y_{j2}}^2}} \quad (2)$$

The smaller correlation angle between the two vectors is, the more similar the DNA sequences are [24]. In addition to the above-mentioned calculation methods, Hamming distances [7] and the number of the edges (the so-called graph theoretical or topological distance) between the two vertices are also used by some works [27,29,36].

2.3. Output data

The output data of DNA sequences similarity calculation methods are similarity matrix (distance matrix) which could be used to generate unweighted pair group method with arithmetic mean (UPGMA). The similarity matrix and UPGMA also could be utilized to evaluate the performance of DNA similarity analysis methods.

2.3.1. Similarity matrix

Similarity matrix is used to represent the similarities of different DNA sequences. It could reveal the biological evolutionary or genetic relationship of different species by quantitative value. According to the similarities distances, the similarities degree of different DNA sequences could be described by a similarity matrix. As a result, similarity matrix could be used to evaluate the effectiveness of DNA similarities analysis algorithm. Besides, the similarity matrix also can be used to build UPGMA phylogenetic tree which would be introduced in Section 2.3.2. And the often-used similarity matrixes are introduced as follows.

The Euclidean matrix represented by Euclidean distance between a pair of vertices (dots), which is a symmetric matrix and is the most often-used matrix, $E = E^T$ [14].

The length/length (L/L) matrix also is symmetric matrix whose off-diagonal elements are given as a quotient of the Euclidean distance between a pair of vertices and the sum of geometrical lengths of edges between the same pair of vertices [14,29]. The ${}^kL/{}^kL$ matrix is constructed from the L/L matrix by raising its individual matrix elements to the kth power [14].

The distance/distance (D/D) matrix associated with a directed graphical representation is an upper triangular matrix; it is defined by Euclidean-distance matrix and graph theoretical distance matrix [27].

The M/M matrix is symmetric matrix whose off-diagonal elements are given as a quotient of the Euclidean distance between two vertices and graph theoretical distance between the two vertices. The entries on the main diagonal are defined by zero [14,29].

Some other similarities matrixes are also defined, such as a correlation of two y-coordinate matrixes (y-Ms) [26]. However, when DNA sequence is very long, most of these matrices would become too large to calculate the eigenvalues, and the computations would be very complex as well [34]. As a result, some alignment-free DNA similarities analysis methods are proposed to avoid the problems of similarity matrixes [1,20].

2.3.2. Phylogenetic tree

UPGMA is calculated from the pairwise distances between the species or products, and it needs input distances or vectors which are represented as similarity matrix generated by distance calculated methods. UPGMA is a binary tree whose branches are defined in chronological, and the branch of the tree contains two pointers to the branches or leaves nodes which are its children. Parent-child distances are set to the unit or by the ultrametric condition if child is a node branch [1].

Phylogenetic tree is used to construct the evolutionary relationships of different species. It sets all creatures on a tree with branches which could obviously distinguish the biological evolution and genetic relationship of different species. Every branch node represents a species, and the length of this branch represents the difference between two species.

3. Different analysis methods of DNA sequences similarities

DNA sequences are composed of four bases over the four-letters alphabet A, T, G and C, which represent the adenine, thymine, guanine, and cytosine, respectively. Since most methods could not directly deal with DNA primary sequence due to it is letters sequences; therefore it should be first converted into numerical sequences, which is called DNA sequence signalization or digitization [3]. Graphical representation methods were most widely used numerical methods for DNA sequences similarities analysis, which allow visual inspection of them, and it could help to recognize the differences among the analyzed DNA sequences as well [20]. Except graphical representation methods, the chemical properties and the composition of bases also were used to digitize DNA sequences for its similarities analysis. Thus, this survey classifies the similarities/dissimilarities analysis methods of DNA sequences into several categories according to the digitization technologies, and tries to describe the classification categories in a simple and sketchy way. This section first introduces the frequently-used biological basis and chemical basis in this field, and then makes an overview on different similarities analysis methods in detail.

Firstly, all forms of species are generated by the different permutations and compositions of four bases. The compositions of the bases will reflect the biological information of DNA sequences, and the connections of the bases will represent its functional characteristics on some level. As a result, many works utilized the theoretical basis of different permutations and compositions to represent the DNA sequences and used them to analyze the similarities of DNA sequences, such as mononucleotide, dinucleotides and trinucleotid based methods.

Secondly, it is well known that four bases of DNA sequence could be classed into several groups according to their chemical properties: purine (A, G)/pyrimidine (C, T), amino (A, C)/keto (G, T) and week-H bond (A, T)/strong-H band (G, C). These chemical properties also can reflect the structural information of DNA sequence, which subsequently reflects their biological information. Therefore, the chemical properties also can be used to quantize DNA sequence and analyze its similarities.

Thirdly, it is well-known that there are many graphical representation methods for transforming DNA sequences into numerical sequences. According to the dimensionality of space in which the sequences are represented, most of the graphical representation methods can be classified into three categories, including 2-D, 3-D and other graphical representation methods [11,13,21–24,39,42].

From the discussion above, it can be known that there are several different strategies can be used to classify these methods, and this work tries to classify them in widely adopted strategies for keeping with the most literatures. The numerical method is the foundation of DNA sequences similarities analysis, and it has a great influence on the similarities results. Hence, this survey classifies all the methods by different numerical strategies of DNA sequences, and takes into consideration the chemical properties and the composition of bases, simultaneously. The classification strategies schematic of this work is shown in Fig. 1, which also is the arrangement of this section.

3.1. Graphical representation based methods

Graphical representation based methods are the most popular research tendency for DNA sequence similarities analysis, and take up most of the whole field. The advantage of graphical representations in DNA sequence is that it allows visual inspection of the data. And it also creates a possibility of numerical characterization to obtain a quantitative measure of the degree of similarity or dissimilarity of different DNA sequences [43]. However, these methods may have some shortcomings, such as overlapping and intersections of the curves, which will lose some important information of DNA sequences and have not high resolution and accuracy. Thus, these analyzing conclusions may be caused a suspicious reaction from biologists [3]. But it still is an important research field in DNA sequences analysis. This section will introduce this kind of method

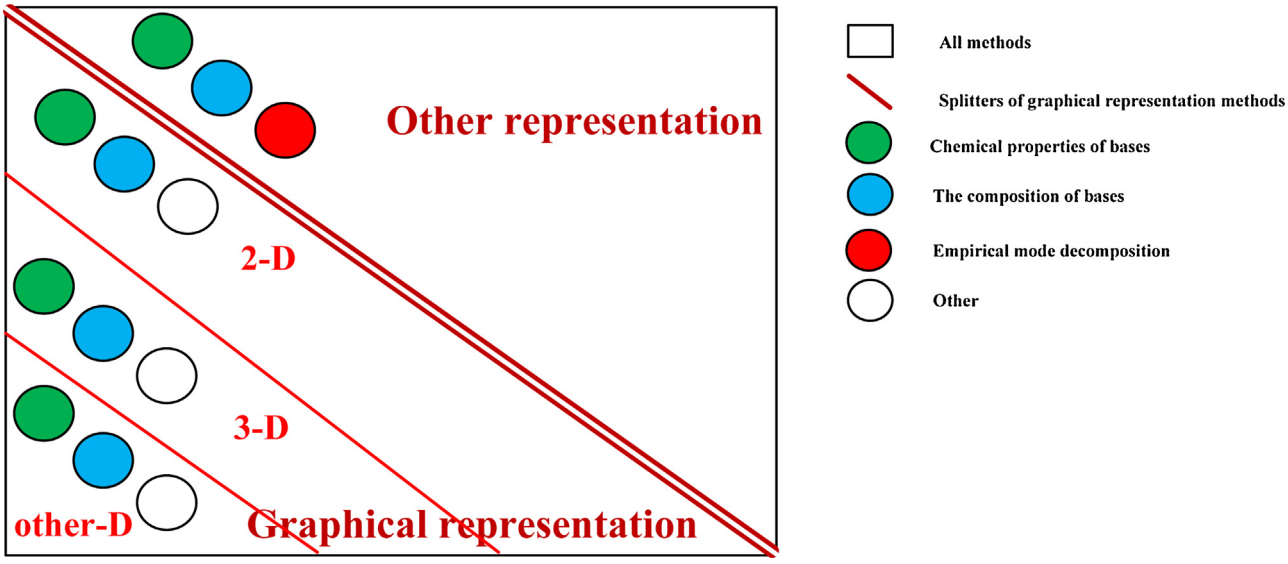

Fig. 1. The classification strategies schematic of DNA sequence similarities analysis methods in this work.

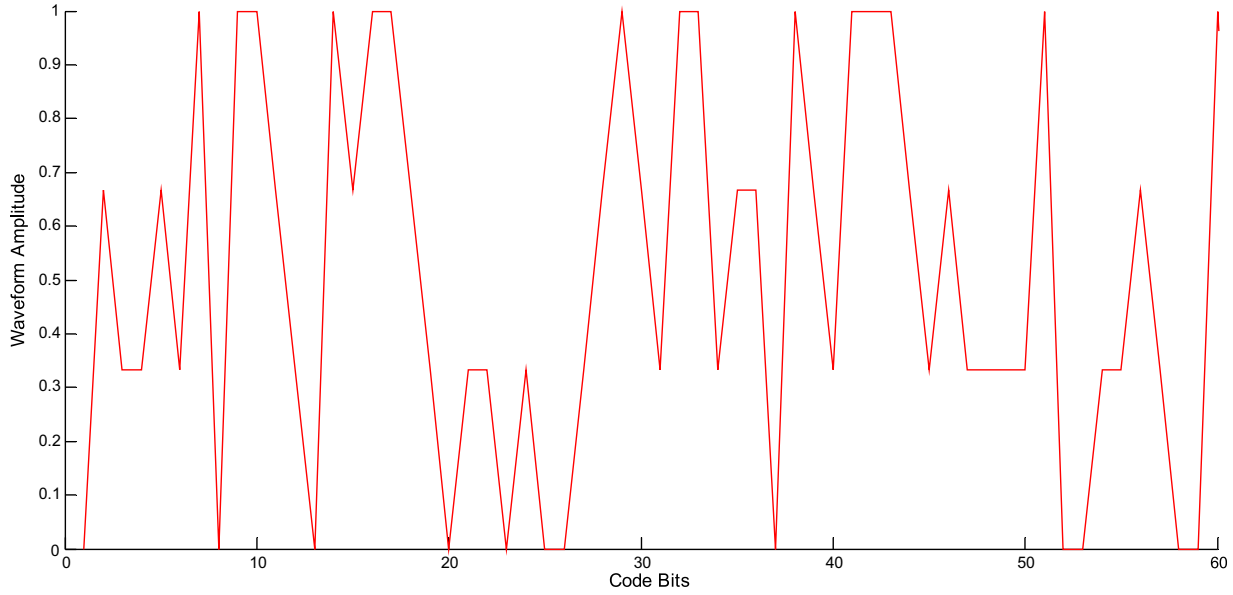

Fig. 2. A diagram of 2-D graphical representation method for DNA sequence.

which consists of three subsections, including 2-D, 3-D and other graphical representation methods.

3.1.1. 2-D graphical representation based methods

The 2-D graphical representation method was one of the most often-used schemes for DNA sequences similarities analysis. It involves a specific mapping of four-type bases to various geometrical alternatives, such as the four directions of Cartesian coordinate axes or the four symmetry non-equivalent horizontal lines [29]. And the mapping methods could be based on the chemical properties of bases, the compositions of bases (such as neighboring nucleotides, dinucleotides and codon) and four-number indicator (or mapping) of bases. The summary of 2-D graphical representation methods for DNA sequences similarities analysis is shown in Table 5. An example of 2-D graphical representation method is shown in Fig. 2 which shows the former 60 bases of Human first exon of β-globin in Table 2. And A = 0, T = 0.75, G = 0.25 and C = 1 in y-axis which marked by "Waveform Amplitude"; and the position of the base in the sequence is x-axis which marked by "Code Bits", such as A, T, G, G, T, G, C, A, C, C... is represented by a 2-D sequence (1, 0), (2, 0.75), (3, 0.25), (4, 0.25), (5, 0.75), (6, 0.25), (7, 1),(8, 0),(9, 1),(10, 1). . ..

3.1.1.1. Chemical properties of bases. The chemical properties of four bases was widely used in DNA sequences similarities analysis due to it is one of the key point to form DNA structure. In 2006, three chemical structure based methods were proposed in 2-D space. Based the chemical structures of DNA sequence, Wang et al. described it as non-A = G, C, T, non-G = A, C, T and non-C = A, G, T; and the non-A, non-G and non-C were represented by 1 and A, G, and C by 0, accordingly. By this way, three (0, 1)-sequences would be obtained to transform them into ${}^{inf}L/{}^{inf}L$ matrices. The sum of the maximal and minimal eigenvalues of these matrices was calculated as a mathematical invariant to describe DNA sequences. The similarities were calculated by Euclidean distance and the correlation angle between the end points of the vectors of different DNA sequences [23].

Table 5
Summary of 2-D graphical representation methods for DNA sequences similarities analysis.

Year	Ref.	Distance	Key words	Dataset
2002	[29]	Euclidean distances	2-D graphical representation, L/L matrix, zigzag curve	the first exon of β-globin gene of 11 species
2006	[23]	Euclidean distances and correlation angle	^{k}L/^{k}L matrix, 3-component vectors, chemical properties	the first exon of β-globin gene 11 species
2006	[25]	Euclidean distances and correlation angle	Cartesian coordinate system, chemical properties	the first exon of β-globin gene of 11 species
2006	[24]	Euclidean distances and correlation angle	eight-component vectors, chemical properties	the first exon of β-globin gene of 11 species
2006	[26]	correlation (RM)	pairs of the neighboring nucleotides, dinucleotides, chemical properties	the first exon of β-globin gene of 11 species
2008	[30]	Euclidean distances	2D graphical representation zigzag curve, Cartesian coordinates system	the first exon of β-globin gene of 11 species
2010	[4]	Euclidean distance	6-D vector and 16-D vector, dinucleotides	the complete of β-globin genes of 11 species
2012	[27]	Euclidean distance, theoretical (topological) distance matrix	ALE-index, distance/distance matrix (D/D), codon	the first exon of β-globin gene of 9 species.

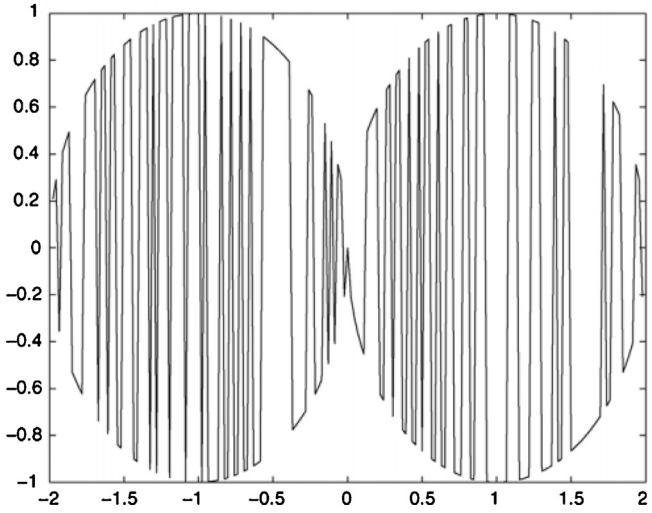

Fig. 3. W-curve of the first exon of human β-globin gene in Ref. [24].

Analogously, Dai et al. represented DNA sequence by two maps between the bases in 2-D space, named W-curve, which was embedded in two unit circles [24], and as shown in Fig. 3. In this method, an 8-component vector with entries was obtained by the average sums of abscissa and y-axis of the three chemical properties. The Euclidean distance and correlation angle were used to calculate the similarities of different DNA sequences. Besides, Yao et al. proposed another 2-D graphical representation method based on the chemical properties, which consisted of three characteristic curves on Cartesian coordinate system, named W-S curve M-K curve and R-Y curve. Afterwards, they presented a quantification of similarities analysis method based on the graph radius, angle and relative departure associated with DNA curve, respectively [25].

Generally, the chemical properties of four bases were often combined with their categories to digitize DNA sequences in 2-D space, which could reflect more useful information about their similarities. In Ref. [25], the space of the W-curve occupied was very small, just two unit circles. And the computational complexity of Yao's method was $O(N)$. All the researchers of Ref. [24] and [25] declared that their representation methods could avoid loss of information of DNA sequences, but it still is problem to prove this viewpoint due to there no appropriate assessment methods. And need to point out that the mentioned three works only used the first exon of β-globin 11 species that were all short sequences, which may be insufficient.

3.1.1.2. The composition of bases. The composition of bases in DNA sequence could reflect its local permutation and determine the biological meaning of short fragment. Therefore, Liu et al. proposed a 2-D graphical representation of DNA sequences based on the 16 kinds of the pairs of neighboring nucleotides and their chemical properties, denoted by PNN [26]. And a 4×4 cells and systems were designed to get the numerical sequences of DNA. Then the PNNs' distributions and a y-coordinate matrix (y-M) would be got, and correlation (RM) of two y-Ms was defined to get the similarities of DNA sequences.

Analogously, Yu et al. used a sliding window with 2-bases width to digitize the DNA sequences, which also was based on the 16 kinds of the pairs of neighboring nucleotides and the chemical properties [4]. The 2-D coordinated with its position could be regarded as D-curve which could convert a DNA sequence into a unique 3-D curve containing no loops based on (x, y, n). From this method, 6 components were used as quantitative descriptors of DNA sequence.

In 2012, Jafarzadeh et al. proposed a 2-D graphical representation method based on each kind of codon for DNA sequences, and the graphical representation data were transformed into a matrix to facilitate quantitative comparisons and computed D/D matrix. Then an invariant of this matrix was calculated, called "ALE-index" [27].

This kind of method often was combined with the chemical properties of four bases to reflect the local information and chemical information of DNA sequences. The advantages of PNN-curve based method was that it could avoid loss of information and did not overlap or intersect with itself in Ref. [26]. In the mentioned two methods in Ref. [26] and [24], a DNA sequences could be regarded as a digital sequence curve which based on dinucleotides; and then the dinucleotides based sequence curve were used to analyze DNA sequences similarities. However, the dinucleotides might do not describe the specific biological significance of DNA sequences. The features of the work in Ref. [27] were that it was simple for calculation lead to it could be easily used to handle long DNA sequences; and it also could use codon representation method to represent DNA sequence, which could more reasonably represent the biological significance of DNA sequences.

3.1.1.3. Other 2-D representation methods based on four horizontal lines. In 2003, Randic et al. proposed a 2-D graphical representation of DNA based on four horizontal lines involving an arbitrary assignment of the four types of bases to the lines, and the four horizontal lines could represent one of the four bases in a lines, as shown in Fig. 4 [28]. Thereafter, they introduced a DNA sequence similarity analysis method based on this representation method and *L/L* matrix. The method was based on the construction of a 12-component vector whose components were the leading eigenvalues of the *L/L* matrices associated with different DNA sequences [29]. Afterwards, Guo et al. proposed another DNA sequences sim-

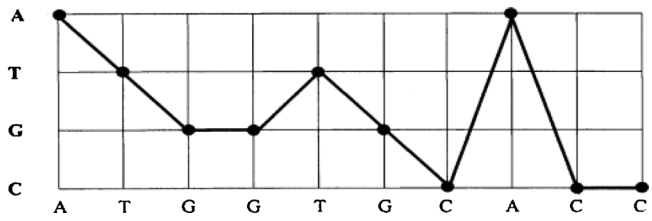

Fig. 4. 2-D graphical representation of sequence ATGGTGCACC in Ref. [28].

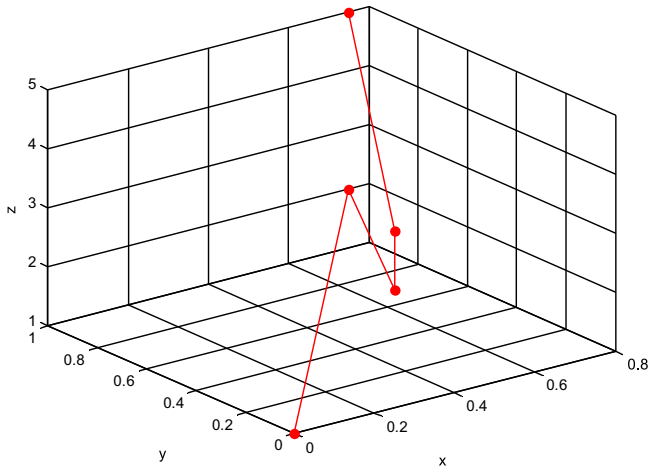

Fig. 5. A schematic diagram of 3-D graphical representation method for DNA sequence.

ilarity analysis method based on the 2-D graphical representation method. One DNA sequence was corresponded to 24 different curves which were set in 2-D Cartesian coordinates system. Their method smoothed the zigzag curve and calculated its curvature as the descriptors of DNA sequences for the similarities instead of calculating the leading eigenvalues of the matrix for graphical representation [30].

The above mentioned two methods could represent DNA sequences in a dynamic way which could provide more random information than traditional methods. However, it ignored the chemical structure of DNA sequences and would lose most of chemical information. Similar techniques by considering the dynamic behavior of DNA sequences also were researched in 3-D space which would be introduced in section 3.1.2.

3.1.2. 3-D graphical representation based methods

The 3-D graphical representation was another often-used DNA sequences similarities analysis method. These methods mapped DNA sequences into a specific 3-D space according to some rules. The mapping methods could be based on the chemical properties of bases or the compositions of bases. It should be pointed out that the mapping rules would produce a great influence on the effect of the similarities analysis method, especially the memory space and computation speed. The summary of 3-D graphical representation methods for DNA sequences similarities analysis is shown in Table 6. An example of 3-D graphical representation method is shown in Fig. 5 which shows the former 5 bases of human first exon of β-globin in Table 2. And A = 0, T = 0.75, G = 0.25 and C = 1 in x-axis; the chemical properties of bases, as purine (A = 0, G = 0)/pyrimidine (C = 1, T = 1), are represented as y-axis: and the position of the bases in the sequence is z-axis". Given A, T, G, G and T, which is represented by 3-D method as (0, 0, 1), (0.75, 1, 2), (0.25, 0, 3), (0.25, 0. 4) and (0.75, 1, 5,).

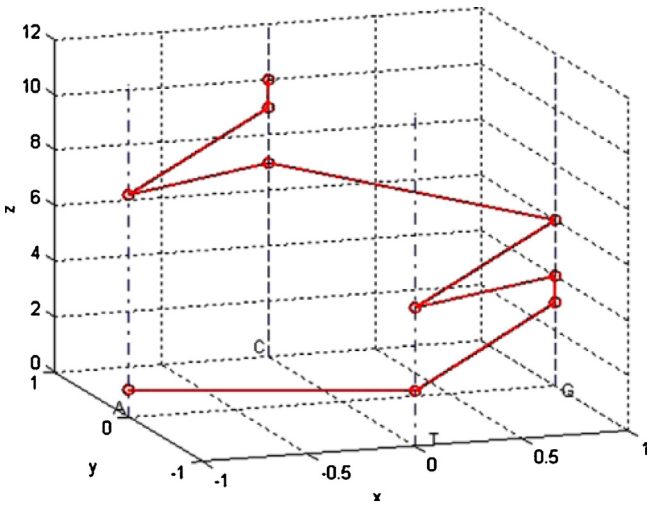

Fig. 6. 3-D graphical representation of sequence ATGGTGCACC in Ref. [31].

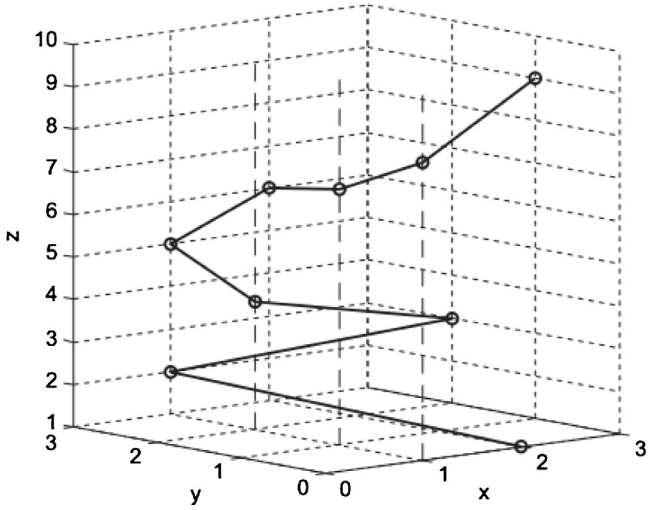

Fig. 7. 3-D graphical representation of sequence ATGGTGCACC in Ref. [32].

3.1.2.1. Chemical properties of bases. In 2004 and 2005, Liao et al. presented two similarity analysis methods of DNA sequences in 3-D space based on the chemical properties of bases [15,31]. In the two works, a 3-D curve of the DNA sequence was represented by three characteristic curves of chemical properties and the orders of bases, which could be regarded as a coarse grained description of the DNA primary sequence, as shown in Fig. 6. The former work constructed a 3-component vector whose components were the leading eigenvalues of the L/L matrices associated with DNA sequences; the later work constructed a 15-component vector which was made up of the average bandwidth of D/D matrices using full DNA sequence.

In Refs. [15,31], the DNA representation methods by the categories of each base also embody the corresponding chemical characteristics; as a result, these approaches considered not only sequences structure but also chemical structure for DNA primary sequences. All the two methods calculated similarities of DNA sequence by Euclidean distance and correlation angle, but these methods only used the first exon of β-globin gene of 11 species as their experimental sample.

3.1.2.2. The compositions of bases. Similar to the method of Ref. [26], Qi et al. proposed a 3-D graphical representation based on the 16 kinds of the pairs of neighboring nucleotides and the corresponding chemical properties, as shown in Fig. 7, and then a 4×4 matrix

Table 6
Summary of 3-D graphical representation methods for DNA sequences similarities analysis.

Year	Ref.	Distance	Key words	Dataset
2000	[36]	Euclidean distance	graphical representations	the first exon of β-globin gene of 8 species
2004	[31]	Euclidean distance	15-component vector, *D*/*D* matrices	the first exon of β-globin gene of 11 species
2005	[35]	Euclidean distance	*L*/*L* matrices, 3-component vector, 4-component vector	the first exon of β-globin gene of 11 species
2005	[15]	Euclidean distance and correlation angle	*L*/*L* matrices, 3-component vector	the first exon of β-globin gene of 11 species
2007	[32]	Euclidean distance	16-component vectors	part of β-globin gene of 9 species
2009	[34]	Euclidean distance	two vectors composed of 64 and six components, trinucleotides	the first exon of β-globin gene of 11 species
2013	[33]	Euclidean distance	codons, space-sum Matrix (s-M) and the distribution Matrix (p-M)	the first exon of β-globin gene of 11 species
2014	[37]	the proposed normalized similarity measure	graphical representations of DNA sequences, dynamic 3D representation	the first and the second exons in β-globin gene

was designed to get the numerical sequences (s-vector, p-vector): s-vector consisting of the 16 space-sums in the matrix s-M, p-vector consisting of the 16 distributions in the matrix p-M. And the similarities among such vectors could be computed in two ways: (i) the Euclidean distance between the end point of the s-vectors; (ii) the Euclidean distance between the end point of the p-vectors [32].

Based on the 64 kinds of codons and chemical properties of four nucleotides in DNA sequence, Jafarzadeh et al. proposed another 3-D graphical representation method (C-curve) which was transformed into two matrices to give the numerical representation of DNA sequences and facilitate its quantitative comparisons, then the space-sum matrix (s-M) and the distribution matrix (p-M) were calculated for analyzing DNA sequences [33]. Similarly, Yu et al. proposed a 3-D graphical representation method based on trinucleotides, named TN curve, and six descriptors were used to numerically represent the DNA sequence, then two Euclidean distance based methods were calculated by the two vectors composed of 64 and 6 components for the similarities analysis of different DNA sequences [34].

Besides, Yao et al. proposed other similarities analysis method based on a 3-D graphical representation method by considering a specific position correlation of DNA sequences to get the numerical characteristic curve of DNA sequence. The quantification of DNA sequence similarities by constructing a 3-component vector whose components were the geometrical center, and a 4-component vector consisting of the graph radius associated with DNA curves [35].

Among these methods, this paper considers the trinucleotid based methods are most suitable for biological significance, because the genetic information of DNA sequences is represented by three consecutive bases, namely, triplet code.

3.1.2.3. Other 3-D representation methods based on dynamic behavior. The dynamic 3-D representation methods also were proposed by some researchers. In 2000, Randic et al. proposed a graphical representation of DNA sequences on the basis of four bases, which were associated with a walk over integral points of a Cartesian coordinate system. By this graphical representation method, a *D*/*D* matrix and its eigenvalues were constructed as the features of individual DNA sequences, which were used for DNA similarity analysis [36]. However, the method of Randic et al. might need greater computational complexity, and the effect was not very good [36].

Piotr et al. presented a DNA sequence similarities analysis method based on 3-D-dynamic representation method [21,37], as shown in Fig. 8. An abstract mass was assigned to each base (point), which was shifted by a unit vector, namely A, G, C, and T were represented by (−1,0,1), (1,0,1), (0,1,1), and (0,−1,1), respectively. Based on the 3-D-dynamic representation method, a DNA sequence could be represented by unit-mass material points which were used to calculate the corresponding descriptors. And a normalized similarity computing method was defined to analysis the similarities

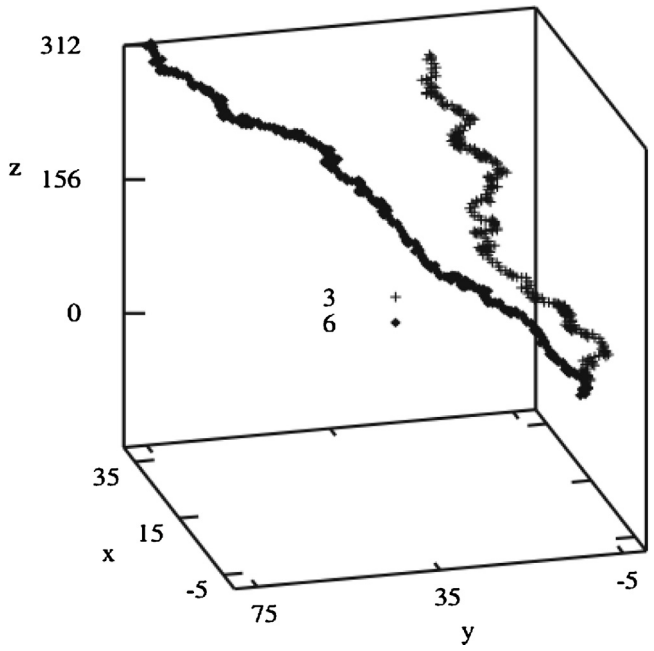

Fig. 8. 3-D graphical representation in Ref. [37].

of DNA sequences. Besides, the multidimensionality of similarity space of complex objects was also discussed in this work.

These methods represent four bases A, G, C, and T in DNA sequence by four different unit vectors which could be used to transform the whole DNA sequence into a dynamic numerical sequence. There are reasons to believe that the methods could contain more similarity information about the DNA sequences due to its dynamic behavior. Another advantage of the 3D-dynamic graph was that it could retain the history of DNA sequences for its similarities analysis.

The dinucleotides based representation method could not effectively describe the biological significance of DNA sequences. The trinucleotides based methods could reflect the biological information of trinuleotides in 3-D space, which was more suitable for biological significance than dividual and dual nucleotides. However, the 3-D space based methods may need more memory space and computation burden than that in 2-D based methods.

3.1.3. Other graphical representation based methods

Except conventional 2-D and 3-D graphical representation method, there were some other graphical methods for similarities analysis of DNA sequences, such as 4-D, 5-D, 6-D, chaos game, and chemical properties with special molecular structures based method. The summary of other graphical methods for DNA sequences similarities analysis is shown in Table 7.

Table 7
Summary of other graphical representation methods for DNA sequences similarities analysis.

Year	Ref.	Coding method	Distance	Key words	Dataset
2004	[14]	6-D representation of triplets of nucleotide	Euclidean distance and correlation angle	triplets, 3-COMPONENT VECTORS, chemical structure	the first exon of β-globin gene of 11 species
2006	[39]	4-D representation	Euclidean distance and correlation angle	chemical properties, categories and position of each base	the first exon of β-globin gene of 11 species
2007	[40]	5-D representation	Euclidean distance and correlation angle	dinucleotide absolute frequency	the first exon of β-Globin gene of 11 species
2010	[43]	chaos game representation	n-letter word similarity measure	a 3-dimensional representation, a numerical characterization, chemical properties	the albumin nucleotides sequences from 8 mammal species
2013	[38]	six different encoding structures	Euclidean distance	structure graph, chemical properties, six special molecular structures	the first exon of β-globin gene of 11 species
2016	[44]	chaos game representation	even scaling method	DFT and power spectrum	others

The three classifications of chemical properties of four bases had three corresponding representations; besides, there were six special molecular structures which also could be used to represent DNA sequences. As a result, Liao et al. proposed a graphical coding method by considering the properties of these three classifications and introduced a total molecular topological index of distance matrix. Then the DNA sequence was reduced into six molecular topological indices of six graphical structures. At last, Euclidean distance was applied to calculate the similarity of DNA sequences [38]. The advantage of this method was that it considered not only sequences structure, but also chemical molecular structure of DNA sequences. The feature of this method was that invariant was easily computed and applied to compare DNA sequences rather than strings sequences.

In a research of Liao et al., they proposed three DNA similarities analysis methods in 4-D, 5-D, and 6-D space, respectively. The 4-D based method represents DNA sequence by the chemical characteristics, categories and position of each base; and the geometrical centers of 4-D graph represented the distribution of base frequencies [39]. The 5-D representation method used five binary numbers points to represent the bases and its position, and the sequence invariants was constructed based on the entries by derived sequence matrices restricted to a selected width of a band along the main diagonal [40]. The 6-D representation of DNA sequences was based on its triplets of nucleotide bases, and the 6-D representation data was transformed another mathematical into a matrix as its invariants descriptors [41]. 4-D and 5-D representation based methods used Euclidean distance and correlation angle to calculate the DNA similarities, but 5-D methods only used Euclidean distance.

Chaos game representation method was first proposed as a scale-independent representation for genomic sequences by Jeffrey [42]. Then Stan et al. proposed a DNA sequences similarities analysis method based on chaos game representation image which was described by a numerical matrix, chaos game representation could evaluate the similarities between images corresponding to different DNA sequences based their patterns [43]. In 2016, Hoang et al. proposed another 2-D chaos game representation (CGR) method as complex numbers to encode DNA sequences, and applied discrete Fourier transform (DFT) to analyze the similarities of DNA sequences by the corresponding power spectrum [44]. The method in Ref. [43] had low computation complexity and required low computation effort, and it also could be used for DNA data classification; the method reported in Ref. [44] had high computation complexity, but the accuracy was better than the former. Besides, the experiments in Ref. [44] were more sufficient than the former.

3.2. Other representation based methods

Although the graphical representation based methods have certain advantages, there also need more innovative numerical methods which could provide more options for different demands. As a result, there are many other numeral methods of DNA sequence for its similarities analysis was studied to avoid the problems of graphical representation [33], such as, Lempel-Ziv complexity, convolutional code model, position-specific statistical, CMI coding and four binary indicators. And most of these methods could be divided into three strategies, including chemical properties of bases, the compositions of bases and empirical mode decomposition. The summary of other representation methods for DNA sequences similarities analysis is shown in Table 8. In this section, other representation based DNA similarities analysis schemes would be introduced in detail.

In 2006, Liu et al. proposed a relative similarity measure method by analyzing the local and global similarity of DNA sequences based on Lempel-Ziv complexity. The relative similarity matrix was calculated on the basis of relative similarity measure. Their method was fully automatic and it did not require sequence alignment, graphical representation and sequence quantification. It avoided the complex calculation and did not require any human intervention; besides, the method also was adaptive to both local and global analysis of DNA sequences [10].

Many classic technologies could also be used in DNA sequences analysis, such as Hamming distance and discrete Fourier transform. Hamming distance could be used to analysis the similarities of DNA sequences, but it was not effective when the length of DNA sequences was not equal. Therefore, Wang et al. defined a bilateral similarity function based method for DNA sequences similarities analysis by considering the normalized Hamming distances and normalized location differences; therefore, the Hamming distance based method could provide a comprehensive comparison of DNA sequences with different length [7]. In 2014, Yin et al. proposed a DNA sequence similarity analysis method based on discrete Fourier transform (DFT) which also could be applied on hierarchical clustering of DNA sequences simultaneously. In Yin's method, DNA sequences were mapped into four binary indicator sequences, and then DFT was applied to transform the indicator sequences into frequency domain. The similarity distance of the DNA sequences metric was calculated by Euclidean distance according to full DFT power spectra [1].

Position-specific statistical was widely used in bioinformatics, and Kuang et al. was the first to use position-specific statistical model in biological sequences analysis which was based on Markov model and the bases specific position matrices to describe randomness degree and local dynamic distribution [45]. Based on graph theory, Qi et al. presented a DNA similarities analysis method which utilized the adjacency matrix of weighted directed graph as the representative vector of DNA to calculate the distance by Euclidean distance, correlation angle and correlation coefficients [46]. Besides, Otsuka et al. investigated similarity relations of DNA

Table 8
Summary of other representation methods for DNA sequences similarities analysis.

Year	Ref.	Coding method	Distance	Key words	Dataset
2001	[9]	4×4 matrices	Euclidean distance and scalar product of vectors	64-component vector, 4×4 matrices, 64 triplets	the first exon of β-globin gene of 8 species
2002	[48]	chemical properties based method	Euclidean distance and correlation angle	chemical properties, 2×2 matrices, triplets	the first exon of β-globin genes for 8 species,
2005	[10]	Lempel-Ziv	the proposed relative similarity matrix	Lempel-Ziv complexity, chemical properties	the first exon sequences of β-globin genes of 11 species
2007	[49]	Euclidean distance	transition probabilities, transition entropies and mean square deviation	12-component vector and two 3-component vectors, chemical properties	the first exon of β-globin gene of 9 species
2008	[52]	codon usage based	Euclidean distance	3-component vectors, 64 triplets of four nucleotides	full β-globin genes of 10 species
2010	[7]	Hamming distance	bilateral similarity function	location difference	the first exon of β-globin gene of 11 species
2011	[46]	weighted graph	Euclidean distance, correlation angle, correlation coefficients	, Directed multi-graph, Graph Theory	the mitochondrial DNA sequences of 12 primate species
2011	[51]	chemical properties and 2-compositions of four bases	the proposed formula	2-compositions of four bases, 10-component vector, chemical properties, Weighted Pseudo-Entropy	the β-globin genes of 15 species
2011	[59]	four number mapping	–	EMD, IMF	the mitochondria of 4 species
2011	[60]	quaternary mapping and mathematization method	the comparing the of corresponding residues	IMF, EMD, MQ-RBF, Quasi-MQ EMD	the mitochondria of 4 species
2013	[20]	short *k*-word based	average distances between pairs of *k*-word	*k*-word, one-to-one mapping	the mitochondria of 18 species, the first exon of β-globin gene of 11 species
2014	[1]	binary indicator sequences	Euclidean distance	DFT, power spectra, frequency domain, alignment-free methods, phylogenetic trees	others
2014	[50]	CMI coding	cross correlation function	graphical representation, chemical properties	the mitochondria of 18 species, the first exon of β-globin gene of 13 species
2014	[57]	chemical properties and Shannon entropy of words	Euclidean Distance	category position frequency, word frequency, position and classification information of nucleotide bases, entropy based model, local word frequency and position information, chemical properties, DNA clustering, Shannon entropy of words	others
2015	[55]	segment of triplets	the designed	maximum segments of triplets	the first exon of β-globin gene of 11 species
2015	[58]	frequency patterns and entropy	Euclidean distance	frequent patterns and entropy, sequence blocking, maximal frequent patterns	b-globin genes of 11 species
2015	[45]	position-specific statistical model	Shannon's entropy	base-specific position matrices, Markov model,	the chromosomes 1 of 8 species, the first exon of β-globin gene of 11 species
2016	[3]	Huffman coding	Euclidean distance	simplified pulse coupled neural network, OTS	the first exon of β-globin gene of 11 species, the mitochondria of 18 species
2016	[56]	four number mapping	Euclidean distance	pulse coupled neural network, OTS	the first exon of β-globin gene of 11 species
2016	[54]	code of three nucleotides	Euclidean distance and correlation angle	characterization vector, complex networks	the 9 complete mitochondrial DNA

and RNA polymerases based on the principal component analysis of amino acid sequences [47].

3.2.1. Chemical properties of bases

Similar to the research of Wang et al. [23], He et al. first represented DNA sequence by three (0, 1) sequences according to its chemical properties, and then a set of 2×2 matrices were constructed to represent DNA sequence by calculating the occurrence frequency of its (0, 1) triplets. Afterwards, the leading eigenvalues of these matrices were calculated as invariants or vectors which could be used to compute the similarities of DNA sequences by Euclidean distance and correlation angle [48]. On the basis of three different curves of chemical properties in Fig. 9, Bai et al. represented a DNA sequence as two random sequences {Ym} and {Xn} which were both Markov chains. And they found that some numerical characterizations of its transition probability distributions could be used as the invariants of DNA sequences. At this point, Euclidean distance between the invariants was employed to desirable the similarities of DNA sequences [49].

Hou et al. adopted CMI coding and six mappings [38] from DNA sequence to translate it into 0-1 coding sequence according to the chemical properties of bases; and then cross correlation function was utilized to calculate the similarities of DNA sequences [50]. This approach considered not only the structures of the DNA sequence, but also the chemical structures for its similarity analysis. However, the differentiation of the Phylogenetic tree generated by this method from different species was not obvious in part.

Based on the chemical properties, 10 different kinds of 2-words-letters (2-compositions of four bases) in total would be obtained, here the repetition was allowed. On the basis of this, Li et al.

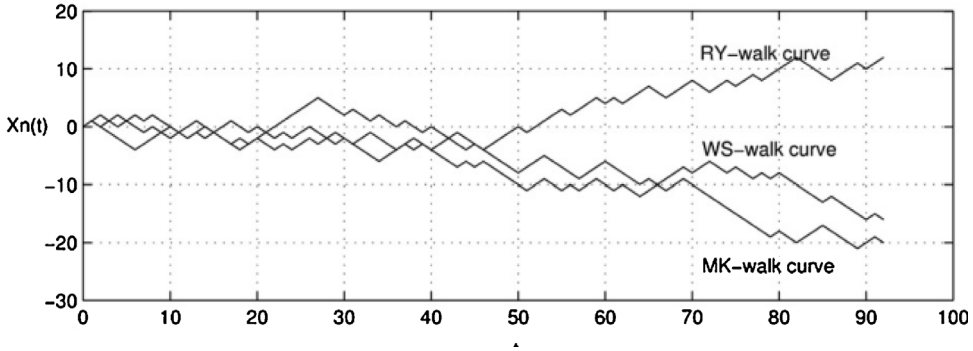

Fig. 9. Three walk curves of exon-1 of human β-globin [49].

transformed DNA sequence into 10-letter sequence; and then 10-component vector was constructed from the 10-letter sequence via weighted pseudo-entropy, which could reflect the element information and the order relation among them of a DNA sequence. The similarity between such two vectors was calculated by the proposed formula [51].

3.2.1.1. The compositions of bases. In the research of Randic et al., 64 triplets were constructed as a set of smaller 4×4 matrices to represent DNA sequences, and the leading eigenvalue of the matrices were selected to characterize DNA; besides, the condensed matrices of DNA sequences including a 64-component vector also were constructed, which consist of ordered triplets XYZ, with X, Y, Z = A, C, G, T; the similarities of DNA sequences were calculated by Euclidean distance and the scalar product [9]. In other work of Li et al., 64 triplets were first classified into 21 groups according to its corresponding amino acids, and defined the relative frequency of codon usage by the proposed formula to construct a line distance (LD) matrix. Then, the leading eigenvalues of LD matrices were chose as descriptors and obtained a 3-component vector accordingly, therefore the similarities could be calculated by Euclidean distance [52].

The advantage of Li's method was that it required neither graphical representation and calculations of invariants of higher order matrices, nor multiple sequence alignment [52]. Besides, other trinucleotide based methods were also proposed in recent years. Such as Liu et al. proposed a similarities analysis of DNA sequence with consideration of the effect of codon based on convolutional code model [53]; Zhou et al. represented a cis sequence complex network which was used to design a characterization vector of DNA sequence based on the three cis nucleotides [54], as shown in Fig. 10; Peng et al. utilized segment of triplets to transform a DNA sequence into a set of maximum segments and associated with a designed similarity calculated method for the similarities analysis of DNA sequences [55].

Jin et al. proposed a DNA sequence similarity calculation method based on simplified pulse coupled neural network (S-PCNN) and codon based coding scheme. Triplet code was took as a coding unit to transform DNA sequence into numerical sequence, and then S-PCNN was used to extract the features of the numerical sequence; at last, Euclidean distance applied to calculate the similarity of numerical sequence by the oscillation time sequence (OTS) from S-PCNN [3]. Afterwards another method was introduced by four number mapping DNA sequence representation method and PCNN [56]. These methods could deal with the DNA sequences with different lengths by the global processing mechanism of PCNN. However, the model had many parameters need to set, and the parameters would produce a great impact on the result. Besides, the methods had a large calculation amount.

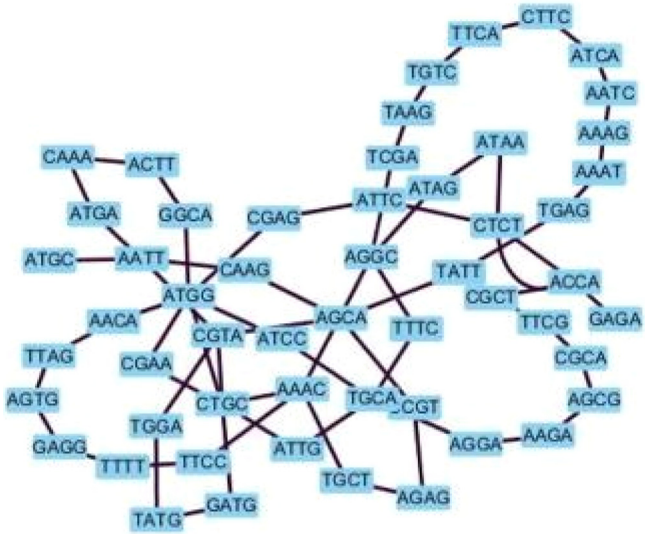

Fig. 10. Tetranucleotide cis sequences network based representation method Ref. [54].

In addition to the traditional compositions of bases, more other methods were proposed in recent years. Yang et al. proposed a biological sequences similarities analysis method which extracted sub-sequences of short *k*-word from biological sequences, and a similarity distance which was calculated by the average distances between pairs of *k*-word [20]. Afterwards, Bao et al. designed a category-position-frequency model which converted DNA sequence into three sequences according to the word frequency, position and classification of nucleotide bases [57]. Xie et al. proposed a method to represent DNA sequences based on their frequent sequential patterns and entropy, which divided DNA sequence into several blocks with the same length [58].

The trinucleotides based methods could more reasonably reflect the biological information of trinuleotides than dividual and dinucleotides. Therefore, the most works in this section were based on trinucleotides. Besides, the Ref. [20,57,58] adopted frequent sequential patterns based methods to represent DNA sequences, which could be regarded as more complex compositions of the bases than traditional methods. When compare to other traditional compositions of bases based representation methods, the three methods could find more long-range dependencies information and local information in DNA sequences.

3.2.1.2. Empirical mode decomposition. The empirical mode decomposition (EMD) method was a nonlinear, non-stationary complicated data processing method. Thus, Zhang et al. proposed two DNA sequences similarities analysis methods based on EMD.

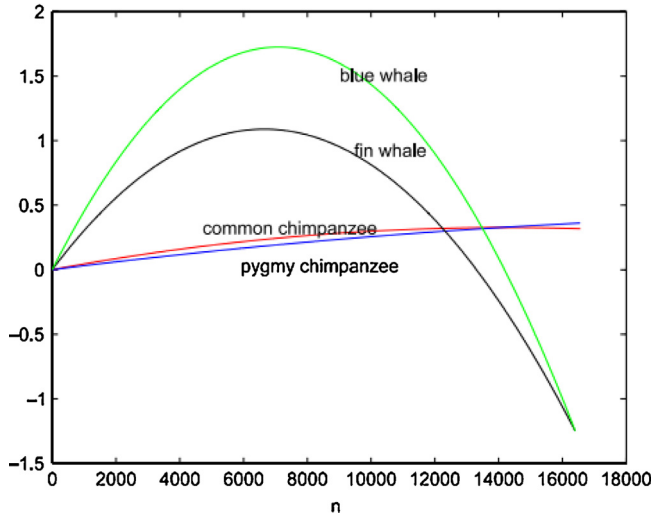

Fig. 11. The residue comparison in a single plot by the EMD method for long DNA sequences [60].

In these methods, DNA sequences were represented as numerical sequences by a mapping of (A→ t = + 1, T→ t = + 2, G→ t = − 1, C→ t = − 2). In the first method, EMD was used to divide nonlinear signal of DNA primary sequences into a group of well-behaved intrinsic mode functions (IMFs) and a residue which could be used to calculate the similarity of DNA sequences [59]. However, this method was affected by endpoints, which might not be used to compare short DNA sequences. As a result, another EMD method based on multi-quadrics radial basis function (MQ-RBF) quasi-interpolation (named Quasi-MQ EMD) was presented. Similar to the prior method, quasi-MQ EMD also could divide DNA numerical sequences into a set of IMFs and a residue which could be used to analyze the similarity of DNA sequences [60], as shown in Fig. 11. It could avoid loss of information when compared with matrix dimension reduction and compression based methods. These works provided new approaches for DNA sequences analysis. However, the two works just demonstrated the results by the plot of residue, which did not testify by similarity matrix or phylogenetic tree.

4. Tendencies and challenges

In recent years, more and more novel DNA representation methods are proposed for DNA similarities analysis; however, most of general works are based on the graphical representation method in the past. This is the obvious tendency in this field. And the invariants or descriptors of DNA sequence are becoming diversity, which could provide more possibilities for more accurate analysis of DNA sequence. But the challenge remains. How to more accurate and convenient analyze the similarity of DNA sequence still need to be advanced. And the various demands and slow progress of DNA sequence similarity analysis methods still are constantly put forward challenge in this field.

4.1. Data sets

DNA sequence test data sets are the most important foundation for verifying the effect of its similarity analysis, and it also can help to design more thorough methods. However, there are only several frequently-used data sets, and the early inchoate DNA similarity analysis methods just used the first exon of β-globin gene of 11 different species as test data. Obviously, it is unreasonable and insufficient for the works to declare the effectiveness of their methods due to the complexity DNA sequences with different lengths. The experimental datasets for testing the validity of DNA similarity analysis methods should be carefully to choose from biological databases, which is a challengeable task. Therefore, it still is a problem to find extra and more suitable DNA sequences samples for verifying the effectiveness of similarity analysis method; especially, it could be widely adopted by most computer scientists and biologists. As a result, there are only few often-used DNA sequence sets for similarity analysis method, and this also is the limitation of the development in this field [20]. Hence, more polytropic and systematized test data sets are needed in this field.

4.2. Numerical method of DNA sequence

It is obvious that the effectively numerical method of DNA sequence will provide more accurate sequence information for its similarity analysis and will play an important role in these schemes. It should be able to avoid losing structural and functional information when DNA sequence is transformed into numeral sequence, and the process should not generate any artifacts or inconsistencies which could distract or mislead a human observer or any subsequent steps. However, the frequently-used graphical representation methods for DNA sequence still have some shortcomings, such as overlapping and intersections of the curves, which would lose some important information of DNA sequences and have not high resolution and accuracy [3]. Although the graphical representation based methods have certain advantages, they also need more innovative numerical methods which could improve the performance of DNA similarities analysis methods and provide more options for different demands.

Analysis and research of DNA sequences should consider not only the strings' structures, but also their chemical structures [10,25]. The chemical properties also can be considered to integrate the compositions of bases. The biological function and evolutionary information of DNA sequences are represented by the permutation and composition of different bases, biological properties and chemical properties; as a result, this work considers that numerical method should take into consideration the above mentioned characteristics of DNA sequences in this research field.

4.3. Robustness and adaptability

The various lengths of DNA sequence would provide a big challenge for the robustness and adaptability of similarities analysis methods, because these methods may do not simultaneously and effectively deal with the short and long DNA sequence. Therefore, the methods should adapt to the DNA sequences with different lengths and take the local feature and global feature into account to effectively analyze the sequences. In some cases, there are only a few site mutations in DNA sequences. However, most of the current methods cannot accurately discover these features in DNA sequences, and represent them in the final similarities analysis results. Besides, a large volume DNA sequence data are available with the development of advanced sequencing technologies, therefore the characteristics of dynamic and complexity would be higher and higher in these data. The current DNA sequences similarities analysis methods seem inadequate for post-genomic studies due to the DNA sequences do not scale well with data size [30]. Therefore, the methods should be actively pursued with better robustness and adaptability.

4.4. Objective evaluation metrics

Many authors proclaimed their methods could avoid loss of information of DNA sequences and require low computation effort and memory space. However, there are no objective criterions to

intuitively prove it. As a result, the work considers that more effective objective indicators could be used to verify the performance of these algorithms such as, the often-used to evaluate the time complexity of an algorithm, "run time". The work also considers that the difference of the information entropy between the digitized DNA sequence and the DNA sequence could be used to measure the loss of the information in the numerical method. More reasonable and effective evaluation indexes for the performance analysis of DNA sequence similarities calculation method should be researched, and this may be a potential research area and would promote the advance in this field.

4.5. Cost-effective computational performance

The often-used DNA sequence similarities analysis tools by biologists still have some weakness, such as large memory, high computation effort and time consuming, especially for long DNA sequences or the whole genome analysis. Therefore, the cost-effective computational methods for DNA sequences similarity analysis are seriously required [7]. Firstly, numerical representations method should have low computational complex [20]. Besides, the extracted features by the similarities analysis methods should effectively represent and compress the genetic information of DNA sequences, which would be used to calculate the similarities analysis of DNA sequences, conveniently.

5. Conclusion

DNA sequence similarities analysis is an important research field in genetics and bioinformatics. In past decades, a large number of similarity analysis methods have been proposed for DNA sequence due to the ever-growing demands. However, it still cannot meet the needs of the biologists and medical scientists. This paper provides a report on the development of DNA sequence similarity analysis methods, and it is mainly focused on the progresses of the past decades. The related knowledge and advances are introduced in this paper, and the corresponding tendencies and challenges are also summarized in this research field. This survey concludes that the demands and slow progress of DNA sequence similarity analysis methods would need to further advance, and there are some possibilities to promote the advance of DNA sequence similarity analysis: (i) more test data should be provided to verify the effect of these methods; (ii) except graphical methods, more novel DNA representation methods should be proposed and took into consideration; (iii) the feature extraction methods should accurately reflect the biological nature; (iv) more effective objective indicators should be used to verify the performance of the algorithm. In the future, it can be inferred that more effective DNA similarities analysis methods would provide more accurate genetic information for experts in the field concerned.

Conflict of interests

The authors declare that there is no conflict of interests regarding the publication of this manuscript. This article does not contain any studies with human participants performed by any of the authors. Informed consent was obtained from all individual participants included in the work.

Acknowledgements

This study is supported by the National Natural Science Foundation of China (No.61640306), and Key Laboratory of Software Engineering of Yunnan Province in China (No.2017SE202). We also thank to the support of Scientific Research Fund of Education Department of Yunnan Province in China (No. 2017YJS108) and Doctoral Candidate Academic Award of Yunnan Province in China.

References

[1] C. Yin, C. Ying, S.T. Yau, A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering, J. Theor. Biol. 359 (24) (2014) 18–28.
[2] A.K. Alqallaf, A.K. Cherri, DNA sequencing using optical joint Fourier transform, Optik – Int. J. Light Electron Opt. 127 (4) (2016) 1929–1936.
[3] X. Jin, R. Nie, D. Zhou, et al., A novel DNA sequence similarity calculation based on simplified pulse-coupled neural network and Huffman coding, Phys.-A Stat. Mech. Appl. 461 (2016) 325–338.
[4] J.F. Yu, J.H. Wang, S. Xiao, Analysis of Similarities/Dissimilarities of DNA sequences based on a novel graphical representation, Match Commun. Math. Comput. Chem. 63 (2) (2010) 493–512.
[5] A. Saini, J. Hou, W. Zhou, Breast cancer prognosis risk estimation using integrated gene expression and clinical data, BioMed Res. Int. 2014 (13) (2014) 459203.
[6] C. Tang, G. Gu, B. Wang, et al., Design, synthesis, and biological evaluation of andrographolide derivatives as potent hepatoprotective agents, Chem. Biol. Drug Des. 83 (3) (2014) 324–333.
[7] S. Wang, F. Tian, Y. Qiu, et al., Bilateral similarity function: a novel and universal method for similarity analysis of biological sequences, J. Theor. Biol. 265 (2) (2010) 194–201.
[8] Z. Xu, B.R. Meher, D. Eustache, et al., Insight into the interaction between DNA bases and defective graphenes: covalent or non-covalent, J. Mol. Graph. Modell. 47 (1) (2014) 8–17.
[9] M. Randić, X. Guo, S.C. Basak, On the characterization of DNA primary sequences by triplet of nucleic acid bases, J. Chem. Inf. Comput. Sci. 41 (3) (2001) 619–626.
[10] N. Liu, T.M. Wang, A relative similarity measure for the similarity analysis of DNA sequences, Chem. Phys. Lett. 408 (4-6) (2005) 307–311.
[11] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. 258 (2) (1983) 1318–1327.
[12] F. Kabli, R.M. Hamou, A. Amine, Similarity analysis of DNA sequences based on the chemical properties of nucleotide bases: frequency and position of group mutations, Comput. Sci. Inf. Technol. 6 (1) (2016) 1–10.
[13] A. Nandy, A new graphical representation and analysis of DNA-sequence structure: 1 Methodology and application to globin genes, Curr. Sci. Assoc. Nandy A 66 (1994) 309–314.
[14] M. Randic, M. Vracko, N. Lers, et al., Novel 2-D graphical representation of DNA sequences and their numberical characterization, Chem. Phys. Lett. 368 (2003) 1–6.
[15] B. Liao, Y. Zhang, K. Ding, et al., Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, J. Mol. Struct. Theochem 717 (1-3) (2005) 199–203.
[16] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, et al., Fractal landscape analysis of DNA walks, Phys. A-stat. Mech. Appl. 191 (1-4) (1992) 25.
[17] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, et al., Analysis of DNA sequences using method of statistical physics, Physica A 249 (1998) 430–438.
[18] S. Kumar, K. Tamura, M. Nei, MEGA3. integrated software for molecular evolutionary genetics analysis and sequence alignment, Brief. Bioinform. 5 (2) (2004) 150–163.
[19] K. Tamura, G. Stecher, D. Peterson, et al., MEGA6: molecular evolutionary genetics analysis version 6.0, Mol. Biol. Evol. 30 (12) (2013) 2725–2729.
[20] X. Yang, T. Wang, Linear regression model of short k-word: a similarity distance suitable for biological sequences with various lengths, J. Theor. Biol. 337 (5) (2013) 61–70.
[21] P. Wąż, D. Bielińska-Wąż, 3D-dynamic representation of DNA sequences, J. Mol. Model. 20 (3) (2014) 2141.
[22] C. Yuan, B. Liao, T.M. Wang, New 3D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 379 (5-6) (2003) 1–6.
[23] J. Wang, Y. Zhang, Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation, Chem. Phys. Lett. 423 (1-3) (2006) 50–53.
[24] Q. Dai, X. Liu, T. Wang, A novel 2D graphical representation of DNA sequences and its application, J. Mol. Graph. Modell. 25 (3) (2006) 340.
[25] Y.H. Yao, X.Y. Nan, T.M. Wang, A new 2D graphical representation-Classification curve and the analysis of similarity/dissimilarity of DNA sequences, J. Mol. Struct. Theochem 764 (1-3) (2006) 101–108.
[26] X.Q. Liu, Q. Dai, Z. Xiu, et al., PNN-curve: a new 2D graphical representation of DNA sequences and its application, J. Theor. Biol. 243 (4) (2006) 555.
[27] N. Jafarzadeh, A. Iranmanesh, A novel graphical and numerical representation for analyzing DNA sequences based on codons, Commun. Math. Comput. Chem. 68 (2012) 611–620.
[28] M. Randic, M. Vracko, L. Nella, et al., Novel 2-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 368 (2003) 1–6.
[29] M. Randić, M. Vračko, N. Lerš, et al., Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem. Phys. Lett. 371 (1) (2003) 202–207.

[30] Y. Guo, T.M. Wang, A new method to analyze the similarity of the DNA sequences, J. Mol. Struct. Theochem 853 (1-3) (2008) 62–67.
[31] B. Liao, T.M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, Chem. Phys. Lett. 388 (1–3) (2005) 195–200.
[32] X.Q. Qi, J. Wen, Z.H. Qi, New 3D graphical representation of DNA sequence based on dual nucleotides, Chem. Phys. Lett. 440 (1) (2007) 139–144.
[33] N. Jafarzadeh, A. Iranmanesh, C-curve: a novel 3D graphical representation of DNA sequence based on codons, Math. Biosci. 241 (2) (2013) 217–224.
[34] J.F. Yu, X. Sun, J.H. Wang, TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications, J. Theor. Biol. 261 (3) (2009) 459–468.
[35] Yu-hua Yao, Xu-ying Nan, Tian-ming Wang, Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation, Chem. Phys. Lett. 411 (2005) 248–255.
[36] M. Randić, M. Vracko, On the similarity of DNA primary sequences, J. Chem. Inf. Comput. Sci. 40 (3) (2000) 599–606.
[37] P. Wąż, D. Bielińska-Wąż, Non-standard similarity/dissimilarity analysis of DNA sequences, Genomics 104 (6) (2014) 464–471.
[38] B. Liao, Q. Xiang, L. Cai, et al., A new graphical coding of DNA sequence and its similarity calculation, Phys. A Stat. Mech. Appl. 392 (19) (2013) 4663–4667.
[39] B. Liao, M. Tan, K. Ding, A 4D representation of DNA sequences and its application, Chem. Phys. Lett. 402 (4-6) (2005) 380–383.
[40] B. Liao, R. Li, W. Zhu, et al., On the similarity of DNA primary sequences based on 5-D representation, J. Math. Chem. 42 (1) (2007) 47–57.
[41] B. Liao, T.M. Wang, Analysis of Similarity/Dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases, J. Chem. Inf. Comput. Sci. 44 (5) (2004) 1666–1670.
[42] H.J. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Res. 18 (8) (1990) 2163.
[43] C. Stan, C.P. Cristescu, E.I. Scarlat, Similarity analysis for DNA sequences based on chaos game representation. Case study: the albumin, J. Theor. Biol. 267 (4) (2010) 513–518.
[44] T. Hoang, C. Yin, S.S. Yau, Numerical encoding of DNA sequences by Chaos Game Representation with application in similarity comparison, Genomics 108 (3) (2016) 134–142.
[45] C. Kuang, X. Liu, J. Wang, et al., Position-specific statistical model of DNA sequences and its application for similarity analysis, MATCH Commun. Math. Comput. Chem. 73 (2015) 545–558.
[46] X. Qi, Q. Wu, Y. Zhang, et al., A novel model for DNA sequence similarity analysis based on graph theory, Evol. Bioinf. 7 (7) (2011) 149–158.
[47] J. Otsuka, N. Kikuchi, S. Kojima, Similarity relations of DNA and RNA polymerases investigated by the principal component analysis of amino acid sequences, Biochimica et Biophysica Acta (BBA)-Protein Struct. Mol. Enzymol. 1434 (2) (1999) 221–247.
[48] P.A. He, J. Wang, Characteristic sequences for DNA primary sequence, J. Chem. Inf. Comput. Sci. 42 (5) (2002) 1080–1085.
[49] F.L. Bai, Y.Z. Liu, A representation of DNA primary sequences by random walk, Math. Biosci. 209 (2007) 282–291.
[50] W. Hou, Q. Pan, M. He, A novel representation of DNA sequence based on CMI coding, Phys. A Stat. Mech. Appl. 409 (3) (2014) 87–96.
[51] C. Li, H. Ma, Y. Zhou, et al., Similarity analysis of DNA sequences based on the weighted pseudo-entropy, J. Comput. Chem. 32 (4) (2011) 675–680.
[52] C. Li, X. Yu, N. Helal, Similarity analysis of DNA sequences based on codon usage, Chem. Phys. Lett. 459 (1-6) (2008) 172–174.
[53] X. Liu, F.C. Tian, S.Y. Wang, Analysis of similarity/dissimilarity of DNA sequences based on convolutional code model, Nucleosides Nucleotides Nucleic Acids 29 (2) (2010) 123–131.
[54] J. Zhou, P. Zhong, T. Zhang, A novel method for alignment-free DNA sequence similarity analysis based on the characterization of complex networks, Evol. Bioinf. Online 12 (2016) 229–235.
[55] H. Peng, L. Wang, J. Zheng, Analysis of Similarities/Dissimilarities of DNA sequences based on segment of triplets, J. Comput. Theor. Nanosci. 12 (9) (2015) 2601–2604.
[56] X. Jin, D. Zhou, S. Yao, et al., Analysis of Similarity/Dissimilarity of DNA sequences based on pulse coupled neural network. multi-disciplinary trends in artificial intelligence, Springer LNAI, in: 10th International Workshop, 10053, 2016, pp. 279–287.
[57] J. Bao, R. Yuan, Z. Bao, An improved alignment-free model for DNA sequence similarity metric, BMC Bioinf. 15 (1) (2014) 321, http://dx.doi.org/10.1186/1471-2105-15-321.
[58] X. Xie, J. Guan, S. Zhou, Similarity evaluation of DNA sequences based on frequent patterns and entropy, BMC Genom. 16 (3) (2015) 1–10.
[59] F. Bai, J. Zhang, J. Zheng, Similarity analysis of DNA sequences based on the EMD method, Appl. Math. Lett. 24 (2) (2011) 232–237.
[60] J. Zhang, R. Wang, F. Bai, et al., A Quasi-MQ EMD method for similarity analysis of DNA sequences, Appl. Math. Lett. 24 (12) (2011) 2052–2058.