SIMILARITY ANALYSIS OF DNA SEQUENCES BASED ON THE CHEMICAL PROPERTIES OF NUCLEOTIDE BASES, FREQUENCY AND POSITION OF GROUP MUTATIONS

Fatima KABLI[1], Reda Mohamed HAMOU[2], Abdelmalek AMINE[3]

GeCode Laboratory, Department of Computer Science
Tahar MOULAY University of Saïda, Algeria.
[1]kablifatima47@gmail.com,[2]hamoureda@yahoo.fr,
[3]amine_abd1@yahoo.fr

ABSTRACT

The DNA sequences similarity analysis approaches have been based on the representation and the frequency of sequences components; however, the position inside sequence is important information for the sequence data. Whereas, insufficient information in sequences representations is important reason that causes poor similarity results. Based on three classifications of the DNA bases according to their chemical properties, the frequencies and average positions of group mutations have been grouped into two twelve-components vectors, the Euclidean distances among introduced vectors applied to compare the coding sequences of the first exon of beta globin gene of 11 species.

KEYWORDS

DNA sequence, chemical proprieties, DNA bases, position, frequency, group mutations.

1. INTRODUCTION

DNA Sequence similarity is fundamental challenge in bioinformatics to predicting unknown sequences functions or effects, constructing phylogenetic tree, and identify homologous sequences, several of DNA sequence similarity measuring approaches have been developed, divided into several categories, the alignment-based, alignment-free, statistics method and others.

Most methods based on the concept of the sequence alignment defined as a way of arranging the sequences of DNA, RNA, or proteins to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences, such as BLAST[1] (Basic Local Alignment Search Tool), was developed as a way to perform DNA and protein sequence similarity searches, is a heuristic method considered as a rapid approach for sequence comparison, based on the comparaison of all combinations of nucleotide or protein queries with nucleotide or protein databases. there are several types of BLAST, Another method call Fasta [2]

Jan Zizka et al. (Eds) : CCSIT, SIPP, AISC, CMCA, SEAS, CSITEC, DaKM, PDCTA, NeCoM - 2016
pp. 13–22, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60102

uses the principle of finding the similarity between the two sequences statistically. This method matches one sequence of DNA or protein with the other by local sequence alignment method. It searches for local region for similarity and not the best match between two sequences.

In addition to performing alignments, is very popular due to its availability on the World Wide Web through a large server at the National Center for Biotechnology Information (NCBI) and at many other sites. Has evolved to provide molecular biologists with a set of very powerful search tools that are freely available to run on many computer platforms.

Also, they are UCLUST [3] and CD-HIT [4] and many more, Obviously it consumes time while running however, the similarity can be quickly computed with the alignment-based method that converts each piece of DNA sequence into a feature vector in a new space. To generate feature vectors some algorithms exploit probabilistic models of which the Markov model [5-6], SVM-based approaches [7], widely used in bioinformatics applications.

Other technique used statistics method for sequence comparison, based on the joint k-tuple content in two sequences called K-typle Algorithm One of the very popular alignment-free methods [8, 9], in which DNA sequence is divided into a window of length k (word of length). The feature vector is generated by the calculated to the frequency value of each tuple; the similarity can be quickly measured by some distance metric between vectors. Such as KLD [10] from two given DNA sequences, was constructed two frequencies vectors of n-words over a sliding window, whereas was derived by a probabilistic distance between two Sequences using a symmetrized version of the KLD, Which directly compares two Markov models built for the two corresponding biological sequences.

On the other hand, these methods cannot completely describe all information contained in a DNA sequence, since they only contains the word frequency information, therefore, many researches modified k-tuple are proposed to contain more information. [11] used both the overlapping structure of words and word frequency to improve the efficiency of sequence comparison. [12] Transformed the DNA sequence into the 60-dimension distribution vectors.

In order to help improve DNA sequence analysis methods ,the graphical representations of DNA sequences on 2D or 3D space [13-14] applied by several researches, but there are some disadvantage as loss of information due to crossing and overlapping of the curve representing DNA with itself [15-16]. To avoid this problem many new graphical representation methods recently [17-14] have been invented.

Other works [18-19] have based on the dinucleotide analysis. To reveal the biology information of DNA sequences. Based on qualitative comparisons used the three classifications of the four DNA bases A, G, T and C, according their chemical properties.

[20] Present the DNA sequence by a 12-component vector consisting of twelve frequencies of group mutations, and calculated the similarity between deferent vectors by the Euclidian distance. While [21], converted a DNA sequence into three 2-dimension cumulative ratio curves the R/Y-ratio curve, the K/M-ratio curve and the W/S-ratio curve, the coordination of every node on these 2-D cumulative ratio curves have clear biological implication.

Li and Wang [22] presented a 16-dimension binary vector based also on the group of nucleotide bases. These methods give encouraging results, they are focused much more on the sequence frequency that the position for sequences analyses, Dong and Pei [23] argued that the position inside sequence is important information. Therefore, unsufficient information in a feature vector is important reason that causes poor similarity results.

In this paper, we combine the advantages of other methods with our own proposal. We presented each DNA sequence by three symbolic sequences according to their chemical proprieties, the group mutations have been grouped into two twelve-components vectors. The first represent the frequency and the second represent the average position, to compare the coding sequences of the first exon of beta globin gene of 11 different species, we applied the Euclidean distances among introduced vectors.

2. MODELLING

2.1 Data Set

We have used in our experimentation DNA sequences derived by the data Set obtained by [23] , the data set contain the first exon of beta globin genes of 11 different species in Table 1

Table 1. The first exon of beta globin genes of 11 different species

Species	Sequences
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGA AAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAA GGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCA AGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAA GGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAA AGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGACGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAA GGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAA GGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAA AGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

2.2 Proposed Method

It is difficult to obtain the information from DNA primary sequence directly; in our approach we based on the three classes of DNA bases, according the chemical properties, the purine group R = {A, G} and pyrimidine group Y = {C, T}; amino group M = {A, C} and keto group K = {G, T}; weak H-bond group W= {A,T} and strong H-bond group S= {C, G}. They call RY classification, MK classification and WS classification correspondingly. Whereas, for the primary sequence X = $S_1..S_2…S_3….\ S_n$, with length n, is presented by three different sequences according the three classification, RY, MK and WS by $\Phi_{RY}, \Phi_{MK}, \Phi_{WS}$

$\Phi_{RY}(X) = \Phi_{RY}(s_i)\ \Phi_{RY}(s_2)\ldots\ \Phi_{RY}(s_n)$, $\Phi_{MK}(X) = \Phi_{MK}(s_i)\ \Phi_{MK}(s_2)\ldots\Phi_{MK}(s_n)$, $\Phi_{ws}(X) = \Phi_{ws}(s_i)\ \Phi_{ws}(s_2)\ldots..\ \Phi_{WS}(s_n)$.

$$\text{Where}\quad \Phi_{RY}(s_i) = \begin{cases} R \text{ if } s_i \in R \\ Y \text{ if } s_i \in Y \end{cases}, i=1,2,\ldots\ldots,n \qquad (1)$$

$$\text{Where}\quad \Phi_{MK}(s_i) = \begin{cases} M \text{ if } s_i \in M \\ K \text{ if } s_i \in K \end{cases}, i=1,2,\ldots\ldots,n \qquad (2)$$

$$\text{Where}\quad \Phi_{WS}(s_i) = \begin{cases} W \text{ if } s_i \in W \\ S \text{ if } s_i \in S \end{cases}, i=1,2,\ldots\ldots,n \qquad (3)$$

Each DNA sequence represented by the three symbolic sequences according the three formula above.

2.2.1 Frequency Analyses

In each classification we focus on group mutation information, for the three symbolic sequences there are twelve group mutations**, R →R, R →Y, Y →R, Y →Y, M →M, M →K, K →M, K →K, W →W, S →W, W →S, S →S.**

As a first step, we calculated the frequency of each mutation information defined by the following formula used by [19].

$$f_{UV} = \frac{\text{the number of word UV}}{n-1} \qquad (4)$$

UV is the mutation information for RY classification, the frequencies denoted by $f_{RY}\ f_{RY}\ f_{YR}\ f_{YY}$, while the frequencies of MK classification denoted by $f_{MM}\ f_{MK}\ f_{KM}\ f_{KK}$, and for WS classification denoted by $f_{WW}\ f_{WS}\ f_{SW}\ f_{SS}$,

The table 2 present the frequencies of group mutations of the first exon of β-globin gene of eleven species based on the three symbolic sequences of DNA.

Table 2. Frequencies of group mutations of 11 species

	f_{RR}	f_{RY}	f_{YR}	f_{YY}	f_{MM}	f_{MK}	f_{KM}	f_{KK}	f_{WW}	f_{WS}	f_{SW}	f_{SS}
Human	0.3297	0.2308	0.2308	0.2088	0.1978	0.1978	0.1868	0.4176	0.1209	0.2967	0.2857	0.2967
Gallus	0.3407	0.2308	0.2308	0.1978	0.2418	0.2308	0.2198	0.3077	0.0989	0.2747	0.2637	0.3626
Lemur	0.3626	0.2198	0.2198	0.1978	0.1319	0.2418	0.2308	0.3956	0.1429	0.3187	0.3077	0.2308
Rabbit	0.3483	0.2472	0.2360	0.1685	0.1798	0.1910	0.1910	0.4382	0.1011	0.3146	0.3034	0.2809
Rat	0.3297	0.2418	0.2418	0.1868	0.1978	0.2198	0.2088	0.3736	0.1758	0.2747	0.2637	0.2857
Bovine	0.3882	0.2118	0.2118	0.1882	0.1765	0.2118	0.2000	0.4118	0.1294	0.2824	0.2706	0.3176
Opossum	0.3077	0.2308	0.2308	0.2308	0.2308	0.2198	0.2088	0.3407	0.1319	0.3407	0.3297	0.1978
Gorilla	0.3478	0.2283	0.2283	0.1957	0.1957	0.1957	0.1848	0.4239	0.0978	0.3043	0.2935	0.3043
Mouse	0.3118	0.2258	0.2258	0.2366	0.1935	0.2043	0.1935	0.4086	0.1183	0.3118	0.3011	0.2688
Goat	0.3882	0.2118	0.2118	0.1882	0.1647	0.2353	0.2235	0.3765	0.1059	0.2941	0.2824	0.3176
Chimpanzee	0.3462	0.2308	0.2308	0.1923	0.1923	0.1923	0.1827	0.4327	0.1250	0.2981	0.2885	0.2885

2.2.2 Position analyses

The position inside sequence is important information Therefore; unsufficient information in a feature vector is important reason that causes poor similarity results.

For instance, if two sequences have the same frequency of components but have two different sequencing directions, if we calculate just the frequency similarity. We get them identical, but the position of their components is completely different, there is no biological relationship between them. For this reason and to improve the effectiveness of similarity study of DNA sequence, that considered as a main challenge in the field of bioinformatics sequences. We used the concept of the position of the DNA components.

We based on the group mutations presented above for calculate the average position (average distance); we have proposed the following formula.

$$P_{UV} = \frac{(\sum_{i=1}^{k}(Position_{UV}))}{\mathrm{K}*(\mathrm{n}-1)} \quad (5)$$

K is the number of word UV.

Wherein the position of each component is defined as the average position of the word uv divide by the length of the DNA sequence n.

The Table 3 present the average position of group mutations for the first exon of β-globin gene of eleven species based on the three symbolic sequences of DNA.

Table 3. Average Position of group mutations of 11 species.

	P_{RR}	P_{RY}	P_{YR}	P_{YY}	P_{MM}	P_{MK}	P_{KM}	P_{KK}	P_{WW}	P_{WS}	P_{SW}	P_{SS}
Human	0.5502	0.4746	0.4956	0.4274	0.4621	0.4512	0.4551	0.5480	0.5325	0.4554	0.4573	0.5539
Gallus	0.5115	0.4558	0.4762	0.5317	0.5490	0.4668	0.4670	0.4922	0.5336	0.4259	0.4286	0.5837
Lemur	0.5608	0.4632	0.4841	0.4194	0.5220	0.4590	0.4636	0.5250	0.4632	0.4505	0.4505	0.6332
Rabbit	0.5814	0.4479	0.4414	0.4569	0.4789	0.4243	0.4613	0.5457	0.4969	0.4539	0.4557	0.5807
Rat	0.5326	0.4500	0.4695	0.5171	0.4847	0.4879	0.4922	0.5048	0.4457	0.4585	0.4592	0.5917
Bovine	0.5387	0.4654	0.4876	0.4419	0.5192	0.4163	0.4187	0.5600	0.4920	0.4716	0.4747	0.5316
Opossum	0.5165	0.4731	0.4950	0.4861	0.4987	0.4599	0.4610	0.5346	0.3974	0.4747	0.4751	0.6258
Gorilla	0.5635	0.4695	0.4896	0.4070	0.4571	0.4463	0.4501	0.5535	0.4855	0.4596	0.4622	0.5637
Mouse	0.5877	0.4424	0.4644	0.4506	0.5269	0.4493	0.4528	0.5218	0.4399	0.4520	0.4531	0.6146
Goat	0.5258	0.4654	0.4876	0.4684	0.5277	0.4382	0.4409	0.5460	0.5033	0.4701	0.4735	0.5316
Chimpanzee	0.5502	0.4780	0.4956	0.4163	0.4606	0.4514	0.4555	0.5468	0.5851	0.4587	0.4603	0.5288

2.2.3 Example

For the Following Sequence

ATGGTGCACCTGAC

We get the three symbolic sequences:

Φ_{RY} =**RYRRYRYRYYYRRY.**

Φ_{MK} =**MKKKKKMMMMKKMM.**

Φ_{WS} =**WWSSWSSWSSWSWS.**

From the three sequences we constructed two twelve-component vectors, the first for calculate the frequency of group mutations and the second for their average position.

For the Word RR, its frequency calculated by the formula (4) $\mathbf{F}_{(RR)} = \frac{2}{(14-1)} = \mathbf{0.15}$, and its average position based on formula (5) is $\mathbf{P}_{(RR)} = \frac{((3+12)/2)}{(14-1)} = 0.57$, the same thing for the rest group mutations,

2.3 Similarity and Dissimilarity

In order to analysis the similarity and dissimilarity between two DNA sequences, each sequence represented by two twelve-component vectors as presented above , The similarities between such vectors calculated by the Euclidian distance between their end points for both vectors frequency and average position. In the next, we calculated the average distance by the following formula.

$$Sim_{(s1,s2)} = \frac{SP(s1,s2)+SF(s1,s2)}{2} \quad (6)$$

SP Define the similarity between two sequences related to the position of their components (twelve component average position vectors)

SF Define the similarity between two sequences related to the frequency of their components (twelve component frequency vectors).

More similarity between two DNA sequences defined by smaller value of Euclidian distance.

The following tables present the similarity/dissimalirity matrix of frequencies and average position of group mutation of 11 species and average similarity between them.

Table 4. Similarity/dissimilarity matrix relative to the frequency of group mutations for 11 species.

	Human	Gallus	Lemur	Rabbit	Rat	Bovine	Opossum	Gorilla	Mouse	Goat	Chimpanzee
Human	0.0	0.1491	0.1262	0.0660	0.0879	0.0798	0.1507	0.0357	0.0508	0.1039	0.030
Gallus		0.0	0.2096	0.1853	0.1372	0.1483	0.2015	0.1527	0.1681	0.1288	0.1683
Lemur			0.0	0.1210	0.1262	0.1240	0.1424	0.1316	0.1143	0.1121	0.11215
Rabbit				0.0	0.1231	0.0964	0.1601	0.0672	0.0917	0.1112	0.0909
Rat					0.0	0.1027	0.1523	0.1125	0.1068	0.1158	0.0962
Bovine						0.0	0.1960	0.0736	0.1152	0.0576	0.0728
Opossum							0.0	0.1652	0.1160	0.1859	0.1617
Gorilla								0.0	0.0719	0.0947	0.0344
Mouse									0.0	0.1245	0.0694
Goat										0.0	0.1058
Chimpanzee											0.0

Table 5. Similarity/dissimilarity matrix relative to the average position of group mutations for 11 species.

	Human	Gallus	Lemur	Rabbit	Rat	Bovine	Opossum	Gorilla	Mouse	Goat	Chimpanzee
Human	0.0	0.1635	0.1259	0.0924	0.1540	0.0956	0.1741	0.557	0.1454	0.0945	0.096
Gallus		0.0	0.1598	0.1557	0.1261	0.1675	0.1790	0.1911	0.1644	0.1299	0.1860
Lemur			0.0	0.1064	0.1279	0.1352	0.1133	0.1056	0.1143	0.1338	0.1754
Rabbit				0.0	0.1276	0.1063	0.1520	0.0851	0.0917	0.1081	0.1345
Rat					0.0	0.1649	0.0962	0.1512	0.1153	0.1375	0.2009
Bovine						0.0	0.1586	0.0949	0.1333	0.0473	0.1275
Opossum							0.0	0.1514	0.1108	0.1499	0.1957
Gorilla								0.0	0.1207	0.1095	0.1078
Mouse									0.0	0.1331	0.1960
Goat										0.0	0.1242
Chimpanzee											0.0

Table 6. Average Similarity/dissimilarity matrix between frequency and position of group mutations for the 11 genes sequences.

	Human	Gallus	Lemur	Rabbit	Rat	Bovine	Opossum	Gorilla	Mouse	Goat	Chimpanzee
Human	0.0	0.1563	0.1260	0.0791	0.1209	0.0877	0.1624	0.0457	0.0980	0.099	0.0451
Gallus		0.0	0.1847	0.1705	0.1316	0.1578	0.1902	0.1718	0.1644	0.1293	0.1771
Lemur			0.0	0.1136	0.1270	0.1295	0.1278	0.1186	0.0873	0.1229	0.1484
Rabbit				0.0	0.1253	0.1013	0.1601	0.0672	0.0917	0.1112	0.0909
Rat					0.0	0.1338	0.1242	0.1318	0.1110	0.1266	0.1485
Bovine						0.0	0.1772	0.0842	0.1242	0.0524	0.1001
Opossum							0.0	0.1583	0.1134	0.1677	0.1957
Gorilla								0.0	0.0962	0.1017	0.0711
Mouse									0.0	0.1287	0.1326
Goat										0.0	0.1150
Chimpanzee											0.0

We have observed in (Table 6) of similarity above, that is a great similarity between the sequence of human with gorilla, human with Chimpanzee and chimpanzee with gorilla another similarity between mouse, lemur, and mouse with rat, such as mouse and rat belong to the same Muridae mammalian family. Also for the bovine and goat, they belong to the same Bovidae mammalian family.

Each of opossum and Gallus are far from the rest species, because opossum is the most remote species from the remaining mammals and the Gallus is the only non-mammalian animal among all other species of the dataset. However, the rest nine species are mammals family.

The obtained result it is not an accident, but shows the relationship in evolutionary sense between the twelve species.

The relationship between the 11 species according our own DNA analysis presented in the following dendrogram.

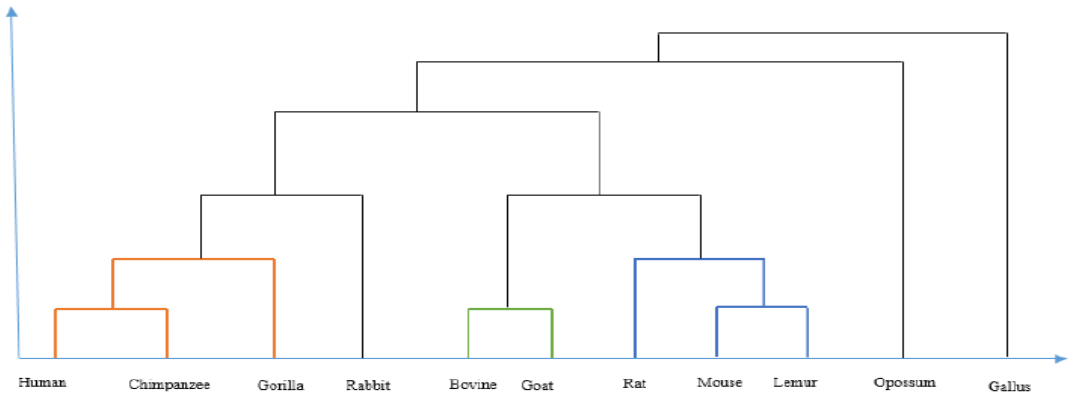

Figure 1. The dendrogram of twelve species.

3. CONCLUSION AND PERSPECTIVE

Similarity analysis of DNA sequences are still important subjects in bioinformatics, the similarity between two DNA sequences defined by the frequency and position of their components. The representation of a DNA sequence by three symbolic sequences helpful to define all possible mutations groups. We build two-dimensional vectors, the first represents the frequency of mutation groups and the second represents their average positions,

To calculate the similarity and dissimilarity of DNA sequences, Euclidean distances are applied based on the frequency and position of mutation groups

The evaluation results of 11 different species coincides with the evolutionary sense. The proposed method has a wide Range of applicability for analysis of biological sequence.

REFERENCES

[1] Gish W, Miller W, Myers E, Lipman D, AltschulS: Basic local alignment search tool. J Mol Biol , 215(3):403-410. doi:10.1016/S0022-2836(05)80360-2 (1990).

[2] Lipman DJ, Pearson WR: Rapid and sensitive protein similarity searches. Science, 227:1435-1441, (1985).

[3] Edgar RC: Search and clustering orders of magnitude faster than blast Bioinformatics, 26:2460-2461, (2010).

[4] Li WZ, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22:1658-1659, (2006).

[5] Pham TD, Zuegg J: A probabilistic measure for alignment-free sequence comparison. Bioinformatics , 20:3455-3461, (2004).

[6] Freno A: Selecting features by learning markov blankets. Lect Notes Comput Sci, 4692:69-76, (2007).

[7] Deshpande M, Karypis G: Evaluation of techniques for classifying biological sequences. Lect Notes Comput Sci, 2336:417-431, (2002).

[8] Blaisdell BE: A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci U S A, 83(14):5155-5159, (1986).

[9] Vinga S, Almeida J: Alignment-free sequence comparison–a review. Bioinformatics, 19:513-523, (2003).

[10] Wu,T.J.,Hsieh,Y.C.and Li,L. A. Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics, 57,441–448, (2001).

[11] Dai Q, Liu XQ, Yao YH, Zhao FK: Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison. J Theor Biol , 276:174-180, (2011).

[12] Zhao B, He RL, Yau SS: A new distribution vector and its application in genome clustering. Mol Phylogenet Evol , 59:438-443, (2011).

[13] Hamori, E., Ruskin, J., Curves, H.: A Novel Method of Representation of Nucleotide Series Especially Suited for Long DNA Sequences. J. Biol. Chem. 258, 1318–1327 (1983).

[14] Qi, Z., Qi, X.: Novel 2D graphical representation of DNA sequence based on dual nucleotides. Chem. Phys. Lett. 440, 139–144 (2007).

[15] Gates, M.A.: A Simple way to look at DNA. J. Theor. Biol. 119, 319–328 (1986).

[16] Guo, X.F., Randic, M., Basak, S.C.: A novel 2-D graphical representation of DNA sequences of low degeneracy. Chem. Phys. Lett. 350, 106–112 (2001).

[17] Randic, M., Vrakoc, M., Lers, N., Plsvsic, D.: Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. 368, 1–6 (2003).

[18] Liu, X.Q., Dai, Q., Xiu, Z.L., Wang, T.M.: PNN-curve: A new 2D graphical representation of DNA sequences and its application. J. Theor. Biol. 243, 555–561 (2006).

[19] i, Z., Fan, T.: PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. 442, 434–440 (2007).
[20] Shi L, Huang HL: Dna sequences analysis based on classifications of nucleotide bases. Adv Int Soft Comput , 137:379-384, (2012).
[21] Yu HJ: Similarity analysis of dna sequences based on three 2-d cumulative ratio curves. Lect Notes Comput Sci , 6840:462-469, (2012).
[22] Li C, Wang J: Similarity analysis of dna sequences based on the generalized lz complexity of (0,1)-sequences. J Math Chem , 43:26-31, (2008).
[23] Dong GZ, Pei J: Classification, clustering, features and distances of sequence data. Adv Database Syst, 33:47-65, (2007).
[24] Nandy, A., Harle, M., Basak, S.C.: Mathematical descriptors of DNA sequences development and applications. ARKIVOC ix, 211–238 (2006).
[25] Zhao, L., et al.: An S-Curve-Based Approach of Identifying Biological Sequences. Acta Biotheoretica 58(1), 1–14 (2009).
[26] Xie, G., Mo, Z.: Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. Journal of Theoretical Biology 269(1), 123–130 (2011).
[27] Wu TJ, Huang YH, and Li LA, Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. Vol. 21 no .222005, pages 4125–4132 doi: 10.1093/bioinformatics/bti658, (2005).
[28] Sierk M, Person W. Sensitivity and Selectivity in Protein Structure Comparison. Protein Sci.2004 ; 13:773–785.
[29] Krasnogor N, Pelta DA. Measuring the Similarity of Protein Structures by Means of the Universal Similarity Metric. Bioinformatics. 2004;20:1015–1021.
[30] Reinert G, Schbath S, and Waterman MS. Probabilistic and statistical properties of words: an overview. J Comput Biol. 2000;7:1–46.
[31] QI. D and Wang T , Comparison study on k-word statistical measures for protein: From sequence to 'sequence space'