MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **75** (2016) 403-416

ISSN 0340 - 6253

Quantitative Analysis of Genetic Relationship of Species Based on the Network for Protein-Coding DNA Sequences

Yan Peng [a], Yuewu Liu[b, *], Sichun Ruan[b], Yulin Wei[b]

[a] *Key Laboratory for Crop Germplasm Innovation and Utilization of Hunan Province, Hunan Agricultural University, Hunan, 410128, China*

[b] *College of Science, Hunan Agricultural University, Hunan, 410128, China*

yuewuliu@whu.edu.cn

(Received September 28, 2015)

Abstract

A novel quantitative analysis method of genetic relationship of species is presented, which is based on the network composed of 20 nodes corresponding to 20 amino acids and edges corresponding to their interactions. The degree of genetic relationship of species is determined by the distance of the vectors which consist of 40 elements corresponding to the out-degrees and in-degrees of 20 nodes. Compared with the classical methods, it avoids choosing the four reference directions corresponding to four bases of DNA sequence, whose optimal angles are very difficult to get. Moreover, our method is efficient because it only computes the distance of vectors. In contrast, much time is spent for the classical methods on computing the eigenvalue of matrix, which has large sizes for long DNA sequence. The contrast experiments show that this approach is superior to the classical methods.

1 Introduction

No one can have failed to notice the fact that there are a large collection of biological sequences in the sequence databanks, and these databanks continue to grow at an exponential rate. To this phenomenon, it is highly desired to develop effective methods and tools that can integrate these data and mine useful information to promote the development of agriculture,

* Corresponding author.

The first two authors contributed equally to this work.

biology, medicine, and drug design. Among these methods, the graphical representation of DNA sequence has become a very powerful technology for viewing, sorting and comparing various gene structures with an intuitive feel [1].

In 1983, Hamori and Ruskin [2] proposed a 3-D graphical representation for DNA sequences. They mapped a DNA sequence into a three-dimensional space function which can be displayed and manipulated conveniently. Inspired by this idea, Nandy [3] presented the 2-D graphical representation. His strategy is that the four types of bases (adenine (A), guanine (G), thymine (T) and cytosine (C)) are assigned to the four directions (-*x*), (+*x*), (-*y*) and (+*y*), respectively. With the process of a random walk along with the four directions, a curve of a DNA sequence is got. The advantage of this approach is that the curve is more intuitive than 3-D curve, but it may cross and overlap of the resulting curve by itself. To eliminate circuit formation, Randić [4] and Yau, et al. [5] presented their approaches almost simultaneously in 2003. Randić assigned four types of bases (A, G, T, and C) to four symmetric non-equivalent horizontal lines respectively, and Yau et al. assigned four types of bases to four vectors in Cartesian coordinate system, respectively. Subsequently, Liao [6], Dai [7], Wu [8], Jafarzadeh [9], Yamaguchi [10], Wąż [11], Tan [12], Das [13], Liu [14], etc. developed the graphical representations of DNA sequences. Furthermore, Randić [15], Huang [16], Qi [17], Yao [18], El-Lakkani [19], Pal [20] also considered the graphical representation of protein sequences and its application.

After the curve corresponding to the DNA sequence is obtained, three mathematical objects: the *E* matrix, *M/M* matrix, and *L/L* matrix are usually used to numerically characterize DNA sequences, which are proposed by Randić [4]. The definitions of the *E* matrix, *M/M* matrix, and *L/L* matrix are as follows:

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

$$(M/M)_{i,j} = \begin{cases} \frac{E_{i,j}}{|i-j|}, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases},$$

$$(L/L)_{i,j} = \begin{cases} \frac{E_{i,j}}{\sum_{k=i}^{j-1} E_{k,(k+1)}}, & \text{if } i \leq j \\ \frac{E_{ij}}{\sum_{k=j}^{i-1} E_{k,(k+1)}}, & \text{if } i > j \end{cases},$$

where (x_i, y_i) denotes the coordinates of the *i*-th base of DNA sequence in the graphical representation. $E_{i,j}$ is the element of the *i*-th row and *j*-th column of the matrix *E*, and the

same to $l_{i,j}$, $(M/M)_{i,j}$, $(L/L)_{i,j}$.

Finally, the degree of similarity/dissimilarity between two species depends on the distance of the maximum eigenvalues of mathematical objects.

For the above methods, the first and most important step is to choose the reference vectors, according to which a curve of a DNA sequence is obtained by a random walk. These reference vectors determine the accuracy of the degree of similarity and dissimilarity among species. However, it is very difficult to obtain the most optimal reference vectors because the exact degree of genetic relationship of any species is unclear. Another shortcoming is that the classical methods are inefficient, because they must compute the eigenvalue of matrix, whose sizes are large for a long sequence. To get rid of a strong dependency upon the reference vectors and expensive computation, we have constructed a network of protein-coding DNA sequences of species. Some properties of network are used to analyze the genetic relationship of species. The contrast experiments show the advantage of our method.

2 Materials and Methods

2.1 Constructing a network

The central dogma of molecular biology manifests that the genetic information is transmitted by way of DNA → RNA → protein. So it is reasonable to deduce that the genetic relationship of DNA sequences of species is equivalent to the genetic relationship of amino acid sequences corresponding to DNA sequences.

The first step of our method is transcription. A particular segment of DNA is copied into RNA (mRNA, tRNA or rRNA) by the enzyme RNA polymerase. If the gene transcribed encodes a protein, mRNA will be transcribed.

The second step is translation. The mRNA is "read" according to the genetic code, which relates the mRNA sequence to the amino acid sequence in proteins (Table 1). Each group of three bases in mRNA constitutes a codon, and each codon specifies a particular amino acid (hence, it is a triplet code). The mRNA sequence is thus used as a template to assemble the chain of amino acids that form a protein.

Table 1. 20 amino acids and their corresponding mRNA codons.

Amino acids	The mRNA codons	Amino acids	The mRNA codons
Ala	GCU, GCC, GCA, GCG	Leu	UUA, UUG, CUU, CUC, CUA, CUG
Arg	CGU, CGC, CGA, CGG, AGA, AGG	Lys	AAA, AAG
Asn	AAU, AAC	Met	AUG
Asp	GAU, GAC	Phe	UUU, UUC
Cys	UGU, UGC	Pro	CCU, CCC, CCA, CCG
Gln	CAA, CAG	Ser	UCU, UCC, UCA, UCG, AGU, AGC
Glu	GAA, GAG	Thr	ACU, ACC, ACA, ACG
Gly	GGU, GGC, GGA, GGG	Trp	UGG
His	CAU, CAC	Tyr	UAU, UAC
Ile	AUU, AUC, AUA	Val	GUU, GUC, GUA, GUG
Stop codon	UAG, UGA, UAA		

20 amino acids are taken as 20 nodes in a network. If two amino acids are adjacent, the two nodes corresponding to the two amino acids have an edge, and the direction of the edge corresponds to the order of the two amino acids. In addition, if two amino acids are adjacent in the amino acids sequence, the weight of the edge is 1, and if the same two adjacent amino acids are repeated for the n-th time in the sequence, the weight of the edge becomes to n. To clearly describe the method of constructing the network, an example is taken. A fragment of mRNA is “ACU UCU CCU ACC UCC”, and their corresponding amino acid sequence is “Thr Ser Pro Thr Ser”. There are three amino acids (“Thr”, “Ser”, “Pro”) in the amino acid sequence, so there are three nodes in the network. The first directed edge is from the amino acid “Thr” to “Ser”, and its weight is 1. The second directed edge is from the amino acid “Ser” to “Pro”, and its weight is 1. The third directed edge is from the amino acid “Pro” to “Thr”, and its weight is 1. The fourth directed edge is from the amino acid “Thr” to “Ser”, and its

weight becomes 2, because it is the second time that this directed edge occurs. The network is shown in Figure 1.

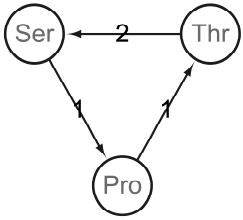

Figure 1. The network corresponds to the example mRNA sequences

2.2 A numerical characterization for genome sequence

For the k-th node in the network, its edges directing from itself to other nodes will be $EO_i^k,\ i=0,\ 1,\ \dots,\ N_k,$ and the weights of these edges will be $WO_i^k,\ i=0,\ 1,\ \dots,\ N_k$, where N_k is the total number of the edges directing from the k-th node to other nodes. If $N_k=0$, it means that there is not an edge directing from the k-th node to other nodes. In the same way, these edges directing from other nodes to itself will be $EI_i^k,\ i=0,\ 1,\ \dots,\ N_k,$ and the weights of these edges will be $WI_i^k,\ i=0,\ 1,\ \dots,\ N_k$. Let $WO^k=\sum_{i=0}^{N_k} WO_i^k$ and $WI^k=\sum_{i=0}^{N_k} WI_i^k$. Then, $WO^1,WO^2,\cdots,WO^{21}$ are formed a vector $\left(WO^1,WO^2,\cdots,WO^{21}\right)$, where the length of this vector is 21 because there are 20 kinds of amino acids and one kind of "Stop codon". Divided by the sum of its elements, this vector is normalized to

$\left(NWO^1,NWO^2,\cdots,NWO^{21}\right)$, where $NWO^k=WO^k\Big/\sum_{k=0}^{21} WO^k$.

Similarly, the normalized vector $\left(NWI^1,NWI^2,\cdots,NWI^{21}\right)$ is got. We proposed a feature vector corresponding to the k-th node that consists of NWO^k and NWI^k. Furthermore, a numerical characterization corresponding to the amino acid sequence is presented, defined by

$$v=(NWO^1,NWI^1,NWO^2,NWI^2,\cdots,NWO^{21},NWI^{21}).$$

For the above example shown in Fig. 1, for the amino acid "Thr", $\left(WO^1,WI^1\right)=(2,\ 1)$, and

for the amino acid “Ser” and “Pro” , $(WO^2, WI^2) = (1,\ 2)$ and $(WO^3, WI^3) = (1,\ 1)$, respectively. So $(WO^1, WO^2, WO^3) = (2,\ 1,\ 1)$, and $(WI^1, WI^2, WI^3) = (1,\ 2,\ 1)$. These two vectors are normalized to $(NWO^1, NWO^2, NWO^3) = (0.5,\ 0.25,\ 0.25)$, and $(NWI^1, NWI^2, NWI^3) = (0.25,\ 0.5,\ 0.25)$, respectively. Finally, the numerical characterization for this sequence is (0.5, 0.25, 0.25, 0.5, 0.25, 0.25, 0 , 0, … , 0, 0).

The similarity/dissimilarity between species is determined by the distance of the corresponding numerical characterizations. For example, the numerical characterization for the i-th sequence is $v_i = (NWO_i^1, NWI_i^1, NWO_i^2, NWI_i^2, \cdots, NWO_i^{21}, NWI_i^{21})$, and the numerical characterization for the j-th sequence is $v_j = (NWO_j^1, NWI_j^1, NWO_j^2, NWI_j^2, \cdots, NWO_j^{21}, NWI_j^{21})$. So the degree of similarity for these two sequences is defined as follows:

$$\left\| v_i - v_j \right\| = \sqrt{\sum_{k=1}^{21} \left\{ \left(NWO_i^k - NWO_j^k \right)^2 + \left(NWI_i^k - NWI_j^k \right)^2 \right\}}$$

2.3 The pseudo-code description of the algorithm

Our algorithm mainly includes two phases: constructing a network and determination of numerical characterization. In order to show the algorithm flow more clearly, the pseudo-code of the algorithm is listed in Table 2.

Table 2. The pseudo-code of the algorithm

Input: Protein-coding DNA sequences of species **Output:** Genetic relationship of species
/ / The preparatory work 1. Transform the protein-coding DNA sequences to RNA sequences. 2. Convert RNA sequences into the amino acid sequences based on the triplet codes. 3. Determine the optimal number of “Stop codons”, and the sequences preceding it will be used to analyze the genetic relationship of species. / / Construct networks 4. 20 amino acids are taken as 20 nodes. Then construct the network according to the ordering of the sequence.

5. Count the in-degree and out-degree for each node.

/ / Numerical characterization

6. Let the in-degrees of all nodes form a vector, then normalize it.

7. Let the out-degrees of all nodes form a vector, then normalize it.

8. Numerical characterization of a species is composed of the above two normalized vectors.

/ / Returning the genetic relationship of species

9. The distance of numerical characterizations for two species corresponds to the degree of similarity.

3 Results

In this section, we will illustrate the use of this method with an examination of the genetic relationship among 11 species for the coding sequences of the first three exons of *β-globin* genes, which are shown in Table 3.

In RNA splicing, introns are removed and exons are covalently joined to one another as part of generating the mature mRNA. And the mature mRNA genetic information is in the sequence of nucleotides, which are arranged into codons being made up of three bases each. Each codon encodes for a specific amino acid, except the “Stop codons”. Most “Stop codons” correspond to the addition of an amino acid to a growing polypeptide chain, which may ultimately become a protein.

Similarity/dissimilarity analysis of different species should be based on a nearly equal number of proteins, which corresponding to a nearly equal number of “Stop codons”. In Table 3, the minimum number of “Stop codons” of the first three exons of *β-globin* genes for 11 species is 2. However, for *Human*, *Opossum*, *Gallus*, *Rat*, *Gorilla* and *Chimpanzee*, the numbers of amino acids before the second “Stop codon” are between 11 and 13. They may be too short to distinguish from one another. Therefore, we used the corresponding amino acid sequence before the third “Stop codon” to analyze the genetic relationship among 11 species.

Table 3. Locations of the first three exons in the *β-globin* genes of 11 species, the total number *n* of "Stop codons", and the lengths of the corresponding amino acid sequences among the first three "Stop codons" , denoted by L_1, L_2, L_3, respectively.

No.	*Species*	*Gene ID (NCBI)*	*Locations*	L_1	L_2	L_3	*n*
1	*Human*	*U01317*	*62187..62278,62409..62631,63482..63610*	5	8	54	3
2	*Goat*	*M15387*	*279..364,493..715,1621..1749*	3	34	0	2
3	*Opossum*	*J03643*	*467..558,672..894,2360..2488*	5	7	15	10
4	*Gallus*	*V00409*	*465..556,649..871,1682..1810*	5	7	36	10
5	*Lemur*	*M15734*	*154..245,376..598,1467..1595*	2	103	0	2
6	*Mouse*	*V00722*	*275..367,484..705,1334..1462*	5	50	115	6
7	*Rabbit*	*V00878*	*129..221,348..569,1143..1271*	13	111	0	2
8	*Rat*	*X06701*	*310..401,517..739,1377..>1505*	5	6	56	5
9	*Gorilla*	*X61109*	*4538..4630,4761..4982,5833..>5881*	5	8	54	3
10	*Bovine*	*X00376*	*278..363,492..714,1613..1741*	3	34	80	3
11	*Chimpanzee*	*X02345*	*4189..4293,4412..4633,5484..>5532*	5	8	58	3

The networks of the corresponding amino acid sequences before the third "Stop codon" of 11 species are constructed by our method described in section 2, which are shown in Figure 2. The degrees of genetic relationship among 11 species are listed in Table 4. As indicated in Table 4, the smallest entry is associated with the pairs (*Human, Gorilla*), (*Human, Chimpanzee*), (*Gorilla, Chimpanzee*), (*Lemur, Rabbit*), (*Goat, Rabbit*) *and* (*Goat, Lemur*). On the other hand, the largest entry appears in the rows belonging to *Opossum* and *Gallus*. In fact, *Opossum* is the most remote species from the remaining mammals, and *Gallus* is the only non-mammalian representative.

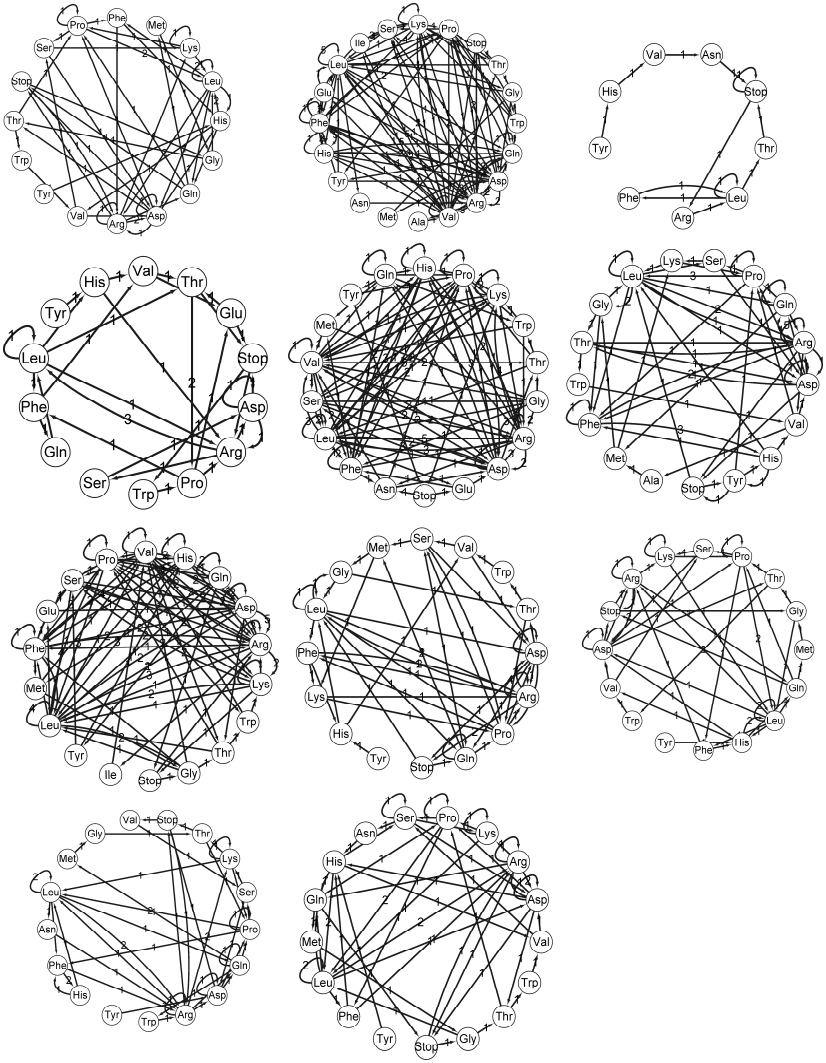

Figure 2. The networks of *Human, Goat, Opossum, Gallus, Lemur, Mouse, Rabbit, Rat, Gorilla, Bovine and Chimpanzee, respectively (from left to right, top to bottom).*

Table 4. Degrees of genetic relationship of 11 species

	Goat	***Opossum***	***Gallus***	***Lemur***	***Mouse***	***Rabbit***	***Rat***	***Gorilla***	***Bovine***	***Chimpanzee***
Human	0.29	0.92	0.41	0.29	0.31	0.29	0.29	0.00	0.25	0.14
Goat	0	0.92	0.42	0.14	0.31	0.11	0.36	0.29	0.34	0.32
Opossum		0	0.80	0.95	0.93	0.96	1.00	0.92	0.93	0.92
Gallus			0	0.43	0.38	0.43	0.41	0.41	0.44	0.44
Lemur				0	0.30	0.10	0.34	0.29	0.35	0.29
Mouse					0	0.29	0.21	0.31	0.29	0.36
Rabbit						0	0.34	0.29	0.37	0.31
Rat							0	0.29	0.25	0.33
Gorilla								0	0.25	0.14
Bovine									0	0.29
Chimpanzee										0

4 Discussion

In order to show the accuracy of the mathematical object proposed by us, we compare it with the classical three mathematical objects, the *E* matrix, *M/M* matrix and *L/L* matrix. In Table 5, we list the recently reported results of the degree of similarity/dissimilarity of 10 species: *Goat, Opossum, Gallus, Lemur, Mouse, Rabbit, Rat, Gorilla, Bovine, Chimpanzee* compared with *Human* by different methods. For an impartial comparison, all of these results were proposed by different researchers and normalized to *Human* to *Goat* ratio.

Similarity/dissimilarity of species has been studied for many years, but so far there has been no exact degree of similarity/dissimilarity of species. In order to compare our method with the classical approaches, we cite 11 literatures and treat the mean of similarity/dissimilarity of species in them as the gold standard for comparison. The reasons are as follows:

a) The 11 literatures were done by 11 different scholars.

b) The 11 different approaches were used in 11 literatures.

c) Most of the 11 literatures were cited over 30 times which indicated the high academic influence of them.

Table 5. Similarity/dissimilarity indexes among 11 species. All indexes are normalized to *Human-Goat* ratio

Refs.	Citation times	Normalized index from difference between Human and								
		Opossum	***Gallus***	***Lemur***	***Mouse***	***Rabbit***	***Rat***	***Gorilla***	***Bovine***	***Chimpanzee***
[21]	236	2.43	1.79	1.43	1.37	0.69	0.70	0.34	1.38	0.28
[22]	112	0.56	1.206	1.24	0.60	0.63	0.61	0.89	0.50	0.60
[23]	65	1.54	1.30	1.45	1.28	0.89	0.84	0.32	0.79	0.33
[24]	61	3.71	0.82	2.73	0.69	0.50	0.48	0.07	3.59	0.58
[25]	53	2.62	2.24	1.63	0.70	0.95	0.90	0.07	0.70	0.13
[26]	51	1.14	1.12	1.07	0.78	0.77	0.86	0.27	0.78	0.41
[27]	45	1.70	1.58	1.05	0.26	0.41	0.93	0.19	1.07	0.46
[28]	42	2.49	2.42	1.05	0.93	1.12	1.11	0.55	0.76	2.01
[29]	40	8.22	6.78	3.57	3.21	1.70	4.17	0.71	1.63	0.92
[30]	26	5.22	4.50	1.78	3.53	2.88	2.19	0.00	2.84	0.00
[31]	17	1.84	2.03	1.23	0.60	0.99	0.97	0.51	0.92	0.46

The average of these results in Table 2 is v^{*} = (1, 2.86, 2.34, 1.66, 1.27, 1.05, 1.26, 0.36, 1.36, 0.56). The errors of the classical three methods and our novel method are defined as the distances between the above average and the results obtained by them, respectively. The errors of the classical three methods: *E* matrix, *M/M* matrix, *L/L* matrix and our work are 3.23, 3.14, 3.14 and 1.44, respectively. It is noticeable that the error of our method is reduced by about 50% compared with the classical three methods.

5 Conclusions

Our method is developed on the basis of the classical methods: the *E* matrix, the *M/M* matrix and the *L/L* matrix. Compared with these classical methods, it has some advantages. First, our method is based on the network which is composed of 20 amino acids and their interactions. It avoids the process that four types of bases (adenine (A), guanine (G), thymine (T) and cytosine (C)) are assigned to the four reference directions. It is well-known that these reference vectors determine the curves corresponding to the DNA sequences of species, thus they decide the accuracy of the degree of similarity and dissimilarity among species. However, it is very difficult to obtain the optimal reference vectors because we do not know the accurate degree of similarity and dissimilarity of any species.

Second, in our approach, the degree of similarity and dissimilarity of species is

measured by the Euclidean distance between the numerical characterizations which consist of out-degrees and in-degrees of 20 nodes. Because the length of the numerical characterization is always 40, the calculation is very simple. The classical methods are based on the matrix invariants (e. g. the leading eigenvalue) of mathematical objects: the *E* matrix, *M/M* matrix, and *L/L* matrix. The mathematical object is a square matrix, and its sizes depend on the length of DNA sequence of species. So, the longer the length of DNA sequence is, the greater the number of calculations is required. The calculation of the matrix invariants (e. g. the leading eigenvalue) of the great size object is very expensive.

Third, our method can also be applied to the analysis of the protein sequences.

However, our method also has also some disadvantages. First, the degrees of similarity and dissimilarity of species may be not accurate for the sequences with shorter length. Generally speaking, about 40 amino acids (120 bases) contain enough information to analyze the genetic relationship with other species. If two species have a closed relationship, our method needs their sequences with longer length.

Second, our method is only applicable to the amino acids sequences and DNA sequences that code for proteins. The reason is that it is based on the network composed of 20 nodes corresponding to 20 amino acids.

Acknowledgment: This work was supported by Foundation of Hunan Educational Committee, China (No. 14C0570, 15B115) and Science and Technology Program of Hunan Province, China (No. 2015JC3099)

References

[1] Nandy, M. Harle, S. C. Basak, Mathematical descriptors of DNA sequences: development and applications, *Arkivoc* **9** (2006) 211-238.

[2] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318-1327.

[3] A. Nandy, A new graphical representation and analysis of DNA sequence structure. I: Methodology and application to globin genes, *Curr. Sci.* **66** (1994) 309-314.

[4] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **368** (2003) 1-6.

[5] S. S. T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y.K. Ho, DNA sequence representation without degeneracy, *Nucleic Acids Res.* **31** (2003) 3078-3080.

[6] B. Liao, A 2D graphical representation of DNA sequence, *Chem. Phys. Lett.* **401** (2005) 196-199.

[7] Q. Dai, X. Q. Liu, T. M. Wang, C (i, j) matrix: A better numerical characterization for graphical representations of biological sequences, *J. Theor. Biol.* **247** (2007) 103-109.

[8] R. Wu, Q. Hu, R. Li, G. Yue, A novel composition coding method of DNA sequence and its application, *MATCH Commu. Math. Comput. Chem.* **67** (2012) 269-276.

[9] N. Jafarzadeh, A. Iranmanesh, C-curve: a novel 3D graphical representation of DNA sequence based on codons, *Math. Biosci.* **241** (2013) 217-224.

[10] K. Yamaguchi, S. Mizuta, A New Graphical representation of DNA Sequences using symmetrical vector assignment, *Rev. Bioinf. Biom.* **1** (2014) 1-26.

[11] P. Wąż, D. Bielińska-Wąż, Non-standard similarity/dissimilarity analysis of DNA sequences, *Genom.* **104** (2014) 464-471.

[12] C. Tan, S. Li, P. Zhu, 4D Graphical representation research of DNA sequences, *Int. J. Biom.* **8** (2015) #1550004.

[13] S. Das, S. Palit, A. R. Mahalanabish, N. R. Choudhury, A new way to find similarity/dissimilarity of DNA sequences on the basis of dinucleotides representation, in: K. Maharatna, G. K. Dalapati, P. K. Banerjee, A. K. Mallick, M. Mukherjee (Eds.), *Computational Advancement in Communication Circuits and Systems*, Springer, 2015, pp. 151-160.

[14] Y. Liu, Y. Peng, A novel technique for analyzing the similarity and dissimilarity of DNA sequences, *Gen. Mol. Res.* **13** (2013) 570-577.

[15] M. Randić, T. Pisanski, Protein alignment: Exact versus approximate. An illustration, *J. Comput. Chem.* **36** (2015) 1069-1074.

[16] G. Huang, W. Huang, W. Xie, Y. Li, L. Xu, H. Zhou, A 2D pattern matching algorithm for comparing primary protein sequences, *Curr. Bioinf.* **9** (2014) 210-217.

[17] Y. Zhao, X. Li, Z. Qi, Novel 2D graphic representation of protein sequence and its application, *J. Fib. Bioeng. Inf.* **7** (2014) 23-33.

[18] Y. Yao, S. Yan, H. Xu, J. Han, X. Nan, P.-a. He, Q. Dai, Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation, *Evol. Bioinf.***10** (2014) 87-96.

[19] A. El-Lakkani, H. Mahran, An efficient numerical method for protein sequences similarity analysis based on a new two-dimensional graphical representation, *SAR QSAR Environ. Res.* **26** (2015) 125-137.

[20] J. Pal, A. Dey, S. Ghosh, D. Bhattacharya, T. Mukherjee, Analysis of similarity between protein sequences through the study of symbolic dynamics, in: K. Maharatna, G. K. Dalapati, P. K. Banerjee, A. K. Mallick, M. Mukherjee (Eds.), *Computational Advancement in Communication Circuits and Systems*, Springer, 2015, pp. 197-214.

[21] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, *Chem. Phys. Lett.* **371** (2003) 202-207.

[22] B. Liao, T.-m. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, *Chem. Phys. Lett.* **388** (2004) 195-200.

[23] Y. H. Yao, T. M. Wang, A class of new 2-D graphical representation of DNA sequences and their application, *Chem. Phys. Lett.* **398** (2004) 318-323.

[24] R. Chi, K. Ding, Novel 4D numerical representation of DNA sequences, *Chem. Phys. Lett.* **407** (2005) 63-67.

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

[26] 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** (2009) 459-468.

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

[28] Z. J. Zhang, DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences, *Bioinf.* **25** (2009) 1112-1117.

[29] Q. Dai, X. Liu, T. Wang, A novel 2D graphical representation of DNA sequences and its application, *J. Mol. Graph. Model.* **25** (2006) 340-344.

[30] J. Wang, Y. Zhang, Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation, *Chem. Phys. Lett.* **423** (2006) 50-53.

[31] X. Tang, P. Zhou, W. Qiu, On the similarity/dissimilarity of DNA sequences based on 4D graphical representation, *Chinese Sci. Bul.* **55** (2010) 701-704.