MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **74** (2015) 563-574

ISSN 0340 - 6253

A New Measure for Pairwise Comparison of Protein Sequences

Nafiseh Jafarzadeh[a], Ali Iranmanesh[a,b,*]

[a]*Department of Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran*

[b]*School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*

(Received April 21, 2015)

Abstract

In this paper, we first introduce a new mathematical approach for comparing DNA primary sequences based on differential expression. The core of our method is a new measure for pairwise comparison of DNA sequences. Next, according to this approach, we give an analogous measure to analysis of protein sequences based on differential amino acids. Our method does not require complex calculations and will be convenient for a fast comparison of biological sequences. Finally, to illustrate its utility, we construct phylogenetic tree for ND5 and ND6 protein sequences of nine species.

1 Introduction

DNA is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. In the recent years, a rapid growth of sequence data in DNA databases has been observed. We know that it is difficult to obtain information directly from the DNA primary sequences, and then mathematical analysis of the large volume of sequences data becomes one of the challenges for bio-scientists.

* Corresponding author. *Email address*: iranmanesh@modares.ac.ir

DNA sequence similarity, the degree of similarity among finite sets of strings of nucleic bases, is a basic problem in bioinformatics, and the resulting information can be used to deduce structures, functions and evolutionary relationships of genes.

Therefore, research in this realm has become an important topic in the field of bioinformatics [1]. Over the last few years, several mathematicians have presented various methods to assign mathematical descriptors to DNA sequences in order to quantitatively compare the sequences and determine similarities and dissimilarities between them [2-29].

As we know, the genetic code consists of 64 triplets of nucleotides. These triplets are called codons. Each codon encodes for one of the 20 amino acids used in the synthesis of proteins. The translation of information encoded in a gene into protein or RNA structures called gene expression. Expressed genes include genes that are transcribed into messenger RNA (mRNA) and then translated into protein.

In this paper,a new method for the similarity analysis of DNA sequences based on differential protein expression is proposed. The basis of our method is a new measure, which we call it "Differential Expression Measure" (DEM). This measure is constructed from codons, which exist in a sequence and have different expressions comparing with another sequence in the same position. The main advantage of this method is that it does not require a graphical representation and it makes a simple and quick comparison for both DNA and protein sequences.

2 Construction of Differential Expression Measure (DEM)

In this section, at first we give some definitions. Let S and Q be two DNA sequences with N(S) and N(Q) codon number respectively.

Definition 2.1. $x_i(S\backslash Q)$ is called a discrimination codon that distinguishes S from Q in the *i*th position based on protein expression. If the *i*th codon in S and the *i*th codon in Q have similar protein expression, then $x_i(S\backslash Q) = \emptyset$. Otherwise, $x_i(S\backslash Q) = D_i$,where D_i is the *i*th codon in S.

Definition 2.2. We denote the set of $x_i(S\backslash Q)$ for $i\epsilon\{1,\ldots,N(S)\}$ by $X(S\backslash Q)$. Therefore, $X(S\backslash Q)$ is the set of all codons that distinguishes S from Q based on protein expression. Similarly, we can define $x_i(Q\backslash S)$ and $X(Q\backslash S)$.

Definition 2.3. We denote the differential expression measure that distinguishes S from Q based on protein expression by DEM(S\Q) and the following formula:

$$DEM(S\backslash Q) = \sum_{\omega \epsilon X(S\backslash Q)} \left[\frac{N(\omega)}{N(S) - N(\omega)}\right]$$

which, N(x) is the number of appearances of the codon x in X(S\Q). We can define the similar formula for DEM(Q\S).

Definition 2.4. The differential expression measure of two sequences S and Q is denoted by the following formula:

$$DEM(S,Q) = \sqrt{DEM(S\backslash Q)^2 + DEM(Q\backslash S)^2}$$

In continues, we intend to show that the set of all DNA sequences (or any subset of that), is a metric space with function DEM as a metric and use it for analyzing and comparing DNA sequences.

In mathematics, a metric space is a set which a notion of distance (called a metric) between elements of the set is defined. A metric or distance is a function with special properties that defines a distance between elements of a set.

In other words, we can say a metric space is an ordered pair (M, d) where M is a set and d is a metric on M, i.e., a function

$$d: M \times M \rightarrow \mathbf{R}$$

such that for any x, y, z ϵ M, the following holds:

1. $d(x,y) \geq 0$,
2. $d(x,y) = 0$, iff $x = y$,
3. $d(x,y) = d(y,x)$ (*symmetry*) and
4. $d(x,z) \leq d(x,y) + d(y,z)$ (*triangle inequality*) .

The following proposition responses our intention.

Proposition 2.5. Let Λ be a set of some arbitrary DNA sequences, then (Λ, *DEM*) is a metric space.

Proof. For the function *DEM* to be a metric, it must satisfy above conditions (I up to IV). Clearly, according to mentioned definitions, DEM satisfies conditions I,II and III. Then it's

efficient to prove just condition IV. Let Q,S and P be three arbitrary DNA sequences, we want to show that DEM(Q,S) ≤DEM(Q,P) + DEM(P,S).

Suppose *x* is an arbitrary codon in X(Q\S), if *x* is also contained in X(Q\P), then we can obtain the triangle inequality DEM(Q\S) ≤DEM(Q\P) + DEM(P\S) .

Otherwise, if there is a codon x in X(Q\S) which is not contained in X(Q\P), then x must be contained in X(P\S). Therefore, we have DEM(Q\S) ≤ DEM(Q\P) + DEM(P\S).

Similarly, we can show that, DEM(S\Q) ≤DEM(P\Q) + DEM(S\P).

Now we need to prove the following inequality.

$$\sqrt{DEM(S\backslash Q)^2 + DEM(Q\backslash S)^2} \leq \sqrt{DEM(P\backslash Q)^2 + DEM(Q\backslash P)^2} + \sqrt{DEM(S\backslash P)^2 + DEM(P\backslash S)^2}$$

From the previous step, we can get

$$\sqrt{DEM(S\backslash Q)^2 + DEM(Q\backslash S)^2} \leq \sqrt{(DEM(P\backslash Q) + DEM(S\backslash P))^2 + (DEM(Q\backslash P) + DEM(P\backslash S))^2}$$

Then it's sufficient to prove the following inequality

$$\sqrt{\left(DEM(P\backslash Q) + DEM(S\backslash P)\right)^2 + \left(DEM(Q\backslash P) + DEM(P\backslash S)\right)^2} \leq$$
$$\sqrt{DEM(P\backslash Q)^2 + DEM(Q\backslash P)^2} + \sqrt{DEM(S\backslash P)^2 + DEM(P\backslash S)^2}$$

By squaring both sides twice, this is equivalent to the following inequality $(DEM(P\backslash Q)\ \ DEM(S\backslash P) + DEM(Q\backslash P)\ \ DEM(P\backslash S))^2 \leq [DEM(P\backslash Q)^2 + DEM(Q\backslash P)^2][DEM(S\backslash P)^2 + DEM(P\backslash S)^2]$

i.e., $2\,DEM(P\backslash Q)\ \ DEM(S\backslash P)DEM(Q\backslash P)\ \ DEM(P\backslash S) \leq DEM(P\backslash Q)^2 DEM(P\backslash S)^2 + DEM(Q\backslash P)^2 DEM(S\backslash P)^2$

Obviously, this inequality is true evermore (Since always we have $(a - b)^2 \geq 0$, i.e. $2ab \leq a^2 + b^2$). Therefore, we prove that DEM(Q,S) ≤ DEM(Q,P) + DEM(P,S).Hence, *DEM* is a metric and (*Λ*, *DEM*) is a metric space. ■

Example 2.6. According to above, consider a set of the coding sequences of the first exon of β-globin gene of Human and some more different species in Table1as a metric space. Then using DEM as a metric, we will have a pairwise comparison for human gene with other species in Table 1.

Table 1. The coding sequences of the first exon of β-globin gene of Human and seven different species

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT GGGGCAAG GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGG TGAACGTGGATGAAGTTGGTGGTGAGGGCCCTGGGCAGGTTGGTATCAAGG
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGTGGGGAAAGG TGAACTCCGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGG TGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGC
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGG TCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCC
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT GAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTG AATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTG CAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG

Table 2. Comparison of Human and other species in Table1 based on DEM

Species	X(Human\---)	X(---\Human)	DEM(Human\---)	DEM(---\Human)	DEM
Chimpanzee	{GAG,TAA}	{GAT, GAA, AGG, TTG, GTA, TCA, AGG}	0.0690	0.0588	0.0906
Gorilla	{}	{AGG}	0	0.0333	0.0333
Mouse	{CCT,GAG,TCT,ACT,GCC,GGC,GTG,GAG, TAA}	{GAT, GCT, GCT, TCT, TGC, GGA, TCC, GAT, GAA}	0.2069	0.2118	0.2961
Rabbit	{ACT,CCT,TAA}	{TCC,AGT,GAA}	0.1023	0.1034	0.1455
Gallus	{CTG, CCT, TCC, GCC, GTT, GCC, GAT, GTT, GGT, GGC}	{TGG, GCT, CAG, CTC, ATC, GGC, GCC, TGT, GCC, GCC}	0.3607	0.3153	0.4791
Opossum	{CTG, CCT, TCT, GCC, GTT, GCC, CTG, CTG, GGC, AAC, GTG, GAT, GAA, GTT}	{TTG, TCT, AAC, TGC, ATC, ACC, ATC, TCT, CAG, GTT, GAC, CAG, ACT, CTT}	0.3498	0.3547	0.4981

In Table 2, we show the processes of pairwise comparison of Human and other species based our new measure. Observing this Table, we note that the most similar species to Human are Gorilla and Chimpanzee, which is expected as their evolutionary relationships. Also, we find that Gallus and Opossum are the most remote from the Human, which coincides with the fact that Gallus is no mammalian and opossum is the most remote species from the remaining mammals.

In continues, we discuss on this question: Can we use this information for the similarity/dissimilarity analysis of protein sequences directly?

3 Construction of Differential Amino acid Measure (DAM)

In continues, according to the above definitions, we give analogous definitions for protein sequences.

Let S and Q be two protein sequences with L(S) and L(Q) sequence lengthrespectively.

Definition 3.1. $\dot{x}_i(S\backslash Q)$ is called a discrimination amino acid that distinguishes S from Q in the *i*th position. If S and Q have the same amino acid in the *i*th position, then $\dot{x}_i(S\backslash Q) = \emptyset$. Otherwise, $\dot{x}_i(S\backslash Q) = A_i$,where A_i is the *i*th amino acid in S. We denote the set of $\dot{x}_i(S\backslash Q)$ for $i\epsilon\{1,\ldots,L(S)\}$ by $\dot{X}(S\backslash Q)$. Therefore, $\dot{X}(S\backslash Q)$ is the set of all amino acids that distinguishes S from Q based on the position of amino acids. Similarly, we can define $\dot{x}_i(Q\backslash S)$ and $\dot{X}(Q\backslash S)$.

Definition 3.2. We denote the Differential Amino acid Measure (DAM) that distinguishes S from Q by DAM(S\Q) and the following formula:

$$DAM(S\backslash Q) = \sum_{\dot{x}\epsilon\dot{X}(S\backslash Q)} [N_{\dot{X}}(\dot{x}) / (L(S) - N_{\dot{X}}(\dot{x}))]$$

which, $N_{\dot{X}}(\dot{x})$ is the number of appearances of the amino acid$\dot{x}$ in $\dot{X}(S\backslash Q)$. We can define the similar formula for DAM(Q\S).

Definition 3.3. The differential amino acid measure of two protein sequences S and Q is denoted by the following formula:

$$DAM(S,Q) = \sqrt{DAM(S\backslash Q)^2 + DAM(Q\backslash S)^2}$$

Clearly, analogous to proposition 1, we can show that the set of all protein sequences (or any subset of that), is a metric space with function DAM as a metric and use it for analyzing and comparing protein sequences.

The question then arises: Does this method work for sequences whose lengths are much different with each other? According to the Definition 3.2., the formula of DAM(S\Q) is independent on L(Q) and the formula of DAM(Q\S) is independent on L(S) then having the same length for two sequences (S and Q) is not necessary. However, when we want to compare a lot of sequence using this method, if the lengths be almost similar, the comparison will be more reliable.

4 Results and Discussion

In this section to illustrate the utility of our new method, we apply it to compare ND5 and ND6 proteins of nine different species from NCBI website, which are shown in Table 3 and Table 4.

Table3. The ND5 proteins of nine different species

Species	Accession	Length
Human	AP_000649.1	603
Gorilla	NP_008222.1	603
Pygmy Chimpanzee		
Common Chimpanzee	NP_008196.1	603
Fin Whale	NP_006899.1	606
Blue Whale	NP_007066.1	606
Rat	AP_004902.1	610
Mouse	NP_904338.1	607
Opossum	NP_007105.1	602

Table 4. The ND6 proteins of nine different species

Species	Accession	Length
Human	CAA24037.1	174
Gorilla	BAA07307.1	174
Pygmy Chimpanzee	BAA85301.1	
Common Chimpanzee	BAA85275.1	
Fin Whale	CAA43450.1	175
Blue Whale	CAA51006.1	175
Rat	CAA32965.1	172
Mouse	CAA24089.1	172
Opossum	CAA82688.1	168

In Tables 5 and 6, we present the similarity/dissimilarity matrices for species listed in Table 3 and 4 based on DAM. Observing Table 5 and 6, we note that the most similar species pairs

are (F. Whale, B. Whale), (P. Chimpanzee, C. Chimpanzee), (Human, C. Chimpanzee), (Human, P. Chimpanzee), (Gorilla, P. Chimpanzee), (Gorilla, C. Chimpanzee) and (Human, Gorilla), which is expected as their evolutionary relationship. At the same time, we find that Opossum is the most remote from the other species, which indicate to the fact that Opossum is the most remote specie from the remaining mammals. By further study of the values in the table, we can gain more information about their similarity.

Table 5. The similarity/dissimilarity matrix for the nine ND5 proteins based onDAM

	Human	*Gorilla*	*P.Chimpanzee*	*C.Chimpanzee*	*F.Whale*	*B.Whale*	*Rat*	*Mouse*	*Opossum*
Human	0	0.1422	0.0969	0.0945	0.4608	0.4638	0.5375	0.5291	0.8757
Gorilla		0	0.1303	0.1326	0.4754	0.4734	0.5465	0.5357	0.8981
P.Chimpanzee			0	0.0707	0.4533	0.4537	0.5375	0.5291	0.8811
C.Chimpanzee				0	0.4560	0.4565	0.5425	0.5343	0.8864
F.Whale					0	0.0492	0.5034	0.5165	0.894
B.Whale						0	0.5010	0.5092	0.8968
Rat							0	0.3172	0.8894
Mouse								0	0.8981
Opossum									0

Table 6. The similarity/dissimilarity matrix for the nine ND6 proteins based on DAM

	Human	*Gorilla*	*P. Chimpanzee*	*C. Chimpanzee*	*F. Whale*	*B. Whale*	*Rat*	*Mouse*	*Opossum*
Human	0	0.0522	0.0522	0.0581	0.9632	0.9710	0.6510	0.6202	0.9347
Gorilla		0	0.0347	0.0464	0.9651	0.9651	0.6579	0.6236	0.9129
P. Chimpanzee			0	0.232	0.9633	0.9633	0.6507	0.6163	0.9113
C. Chimpanzee				0	0.9614	0.9614	0.6560	0.6217	0.9093
F. Whale					0	0.0638	0.9582	0.9235	0.6409
B. Whale						0	0.9681	0.9397	0.6487
Rat							0	0.2029	0.9570
Mouse								0	0.9685
Opossum									0

Another usage of the similarity/dissimilarity matrix is that it can be used to construct phylogenetic tree. The quality of the constructed tree may show whether the matrix is good and therefore whether the method of abstracting information from DNA sequences is efficient. In Figure1 and Figure 2, we show the phylogenetic tree of the nine ND5 and ND6

proteins based on the distance matrix in Table 5 and 6, using the UPGMA method in the MATLAB software. From this figures, we observe that Opossum is clearly separated from the rest and this coincides with real biological phenomenon and the result of this tree is incomplete agreement with [30].

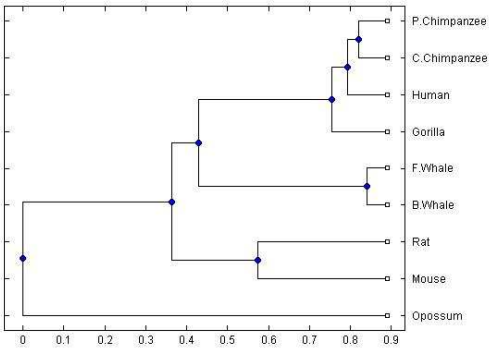

Figure1. The phylogenetic tree for ND5 proteins of nine species

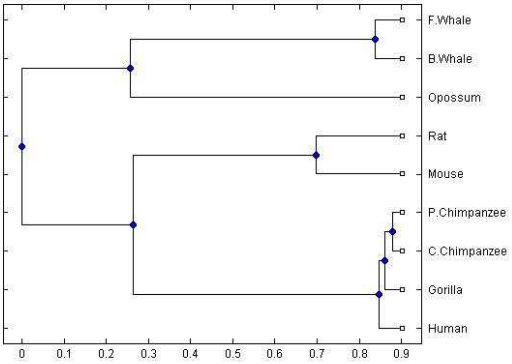

Figure 2. The phylogenetic tree for ND6 proteins of nine species

In Table 7, the protein similarity distance of ClustalW approach [27] for nine ND5 protein sequences is shown. Then in Table 8, we calculate the correlation coefficients and do the

significance analysis to compare ClustalW approach with our method and other current methods.

Table 7. The distances for the ND5 protein sequences of nine species based on ClustalW

	Human	*Gorilla*	*P.Chimpanzee*	*C. Chimpanzee*	*F.Whale*	*B.Whale*	*Rat*	*Mouse*	*Opossum*
Human	0			6.9				48.9	
Gorilla		0	9.7	9.9	42.7	42.4	51.4	49.9	54.0
P. Chimpanzee			0	5.1	40.1	40.1	50.2	48.9	50.1
C.Chimpanzee				0	40.4	40.4	50.8	49.6	51.4
F.Whale					0	3.5	45.3	46.8	52.7
B.Whale						0	45.0	45.9	52.7
Rat							0		
Mouse								0	50.8
Opossum									0

Table 8. The coefficients of correlation for the nine ND5 proteins of our approach and the approaches in Refs [8, 4, 10,12] compared with ClustalW results

	Our method	*Ref [8] method*	*Ref [4] method*	*Ref [10] method*	*Ref [12] method*
Human	0.9371	0.9113	0.9282	0.9405	0.8985
Gorilla	0.9250	0.9199	0.7784	0.9374	0.7942
P.Chimpanzee	0.9192	0.9092	0.9341	0.9431	0.8993
C.Chimpanzee	0.9247	0.9710	0.9404	0.8778	0.9089
F.Whale	0.8948	0.8666	0.7412	0.6496	0.7895
B.Whale	0.8948		0.8054		
Rat	0.7460	0.8412	0.7376	0.6450	0.8013
Mouse	0.6425	0.4288	0.7145	0.6236	0.7787
Opossum	0.6009	0.5259	0.6146	0. 4728	0.6850

As we see in Table 8, comparing with other methods, our method givemore proper results.

5 Conclusion

In this paper, we have proposed a novel measure for analyzing DNA sequences based on differential expression, which we called it DEM. Then using DEM as a metric, we proved that a set of some arbitrary DNA sequences is a metric space. Therefore, we have extracted a new mathematical method to comparing and analyzing genomes. This method is independent of graphical representation and complex calculations, and then it makes a simple and quick

comparison. Finally, according to this approach, we have introduced a new measure for analyzing protein sequences based on amino acids discrimination which we called it DAM and finally using the UPGMA method, we have presented the phylogenetic tree of the nine ND5 and ND6 proteins based onour new approach and discussed about the results.

Acknowledgement. The authors would like to thank the referee for the valuable comments. This work is supported in part by a grant (**BS-1393-1-01**) from the Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.

References

[1] T. Jiang, Y. Xu, M. Q. Zhang (Eds.), *Current Topics Computation in Molecular Biology*, Tsinghua Univ. Press, Beijing, 2002.

[2] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol.***119** (1986) 319–328.

[3] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, *Curr. Sci.* **66** (1994) 309–313.

[4] J. Wen, Y. Y. Zhang, A 2D graphical representation of protein sequence and its numerical characterization, *Chem. Phys. Lett.* **476** (2009) 281–286.

[5] C. T. Zhang, R. Zhang, H. Y. Ou, The Z–curve databases: a graphic representation of genome sequence, *Bioinformatics* **19** (2003) 593–599.

[6] R. Zhang, C. T. Zhang, Z curve, an intuitive tool for visualizing and analyzing the DNA sequences, *J. Biomol. Struct. Dyn.* **11** (1994) 767–782.

[7] A. Nandy, M. Harle, S. C. Basak, Mathematical descriptors of DNA sequences: development and applications, *ARKIVOC* **9** (2006) 211–238.

[8] B. Liao, T. M. Wang, New 2D graphical representation of DNA sequences, *J. Comput. Chem.* **25** (2004) 1364–1368.

[9] B. Liao, W. Zhu, Y. Liu, 3D Graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, *MATCH Commun. Math. Comput. Chem.* **56** (2006) 209–216.

[10] I. Moheb, M. Mervat, A. Marwa, 3D graphical representation of protein sequences and their statistical characterization, *Physica A* **389** (2010) 4668–4676.

[11] M. Randić, M. Vračko, A. Nandy, S. C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, *J. Chem. Inf. Comput. Sci.* **40** (2000) 1235–1244.

[12] Y. H. Yao, Q. Dai, C. Li, P. A. He, X. Y. Nan, Y. Z. Zhang, Analysis of similarity/dissimilarity of protein sequences, *Proteins* **73** (2008) 864–871.

[13] M. Randić, A. T. Balaban, On a four–dimensional representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* **43** (2003) 532–539.

[14] 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.

[15] P. He, J. Wang, Characteristic sequences for DNA primary sequence, *J. Chem. Inf. Comput. Sci.* **42** (2002)1080–1085.

[16] J. Feng, Y. Hu, P. Wan, A. Zhang, W. Zhao, New method for comparing DNA primary sequences based on a discrimination measure, *J. Theor. Biol*. **266** (2010) 703–707.

[17] X. Guo, X. Nandy, Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy, *Chem. Phys. Lett.* **369** (2003) 361–366.

[18] M. Randić, M. Vračko, On the similarity of DNA primary sequences, *J. Chem. Inf. Comput. Sci*. **40** (2000) 599–606.

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

[20] X. Zhou, K. Li, M. Goodman, A. Sallam, A novel approach for the classical Ramsey number problem on DNA-based supercomputing, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 347–370.

[21] Q. Zhang, B. Wang, On the bounds of DNA coding with H-distance, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 371–380.

[22] C. Li, N. Tang, J. Wang, Directed graphs of DNA sequences and their numerical characterization, *J. Theor. Biol*. **241** (2006) 173–177.

[23] Y. Zhang, W. Chen, Invariants of DNA sequences based on 2DD-curves, *J. Theor. Biol*. **242** (2006) 382–388.

[24] J. Feng, Y. Hu, P. Wan, A. Zhang, W. Zhao, New method for comparing DNA primary sequences based on a discrimination measure, *J. Theor. Biol*. **266** (2010) 703–707.

[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) 55–56.

[26] P. He, D, Li, Y. Zhang, X. Wang, Y. Yao, A 3D graphical representation of protein sequences based on the Gray code, *J. Theor. Biol*. **304** (2012) 81–87.

[27] J. F. Yu, X. Sun, J. H. Wang, TN curve: A novel 3D graphical representation of DNA sequence based on tri nucleotides and its applications, *J. Theor. Biol.* **261** (2009) 459–468.

[28] N. Jafarzadeh, A. Iranmanesh, A novel graphical and numerical representation for analyzing DNA sequences based on codons, *MATCH Commun. Math. Comput. Chem*. **68** (2012) 611–620.

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

[30] Y. Liu, D. Li, K. Lu, Y. Jiao, P. He, P-H Curve, a graphical representation of protein sequences for similarities analysis, *MATCH Commun. Math. Comput. Chem*. **70** (2013) 451–466.