MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **61** (2009) 519-532

ISSN 0340 - 6253

H-L curve: A Novel 2D Graphical Representation of Protein Sequences

Yongfan Li [a], Guohua Huang [b *], Bo Liao [b], Zanbo Liu [b]

[a] Hunan First Normal College, Changsha, Hunan 410002, China

[b] School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China

(Received February 25, 2008)

Abstract

According to the partial order constructed on a selected pair of physico-chemical properties of amino acids, we presented a novel 2D graphical representation of protein sequences which is called an H-L curve. The representations are mathematically proven to be no circuit (i. e., without any degeneracy), and associated with protein sequences in a one-to-one manner. In addition, our graphical curves allow more conveniently a visual inspection of protein sequences alignment. We illustrated our approach on two examples.

* Corresponding author: Guohua Huang
Present address: School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China
Fax: +86 731 8228212
Email address: guohuahhn@163.com (Guohua Huang).

1. Introduction

Huge quantities of biology data such as DNA, RNA and protein sequences are rapidly generated everyday with development of the biology technologies. It is absolutely impossible for bio-scientists to deal with amounts of raw data only by conventional experimental approaches. Graphical representation offers a single visual way of analyzing the similarities and dissimilarities among various primary sequences, initiated by Hamori and Ruskin[1,2], and by Gates[3,4] over 20 years ago. Following their pioneering work, many authors have presented different representations of DNA sequences based on 2D, 3D, even higher dimension space or others in recent years. Some examples are available in the Refs. [5-15]. Most of the representations are applied to similarity analysis [5-12], and sequences alignment [13-15]. In contrast to DNA, the graphical representations of protein sequences emerged very lately, whose first arrival [16] was 20 years later than that of DNA. The reason for leading to this is the increased complexity of biological strings built on a 20-letter alphabet (representing the 20 natural amino acids) in comparison with strings build from only four letters (representing DNA or RNA)[17]. Hence, we have 20! different combinations, other than 4! possibilities, which at most mean 24 alternatives, or half in the case of symmetry. So far, most of the existing graphical representations of protein sequences are divided into four classifications---the first [16] based on the underlying codons, the second [18-22] based on 8×8 tables of codons, the third [23] based on an arbitrary alphabetic, and the fourth [24] which assigned 20 amino acids to the 20 directions. In this contribution we will describe a dynamic construction of 2D graphical representation of protein sequences which is not sensitive to arbitrary assignments of amino acid used to represent protein sequences, but based on physico-chemical properties of amino acids.

2. 2D graphical representation of protein sequences

2.1 The assignment of 20 kinds of amino acid to 20 vectors

Randić[25] utilized the pK_a value listed in Table 1 (taken from the Ref. [26]) for terminal amino acid groups NH_3^+ and $COOH$ which determines the activity of enzymes and are

therefore of major importance in biochemistry to construct the 2-D map of amino acids. In the contribution, according to pK_a values, as shown in Fig. 1, we assign 20 kinds of amino acid to 20 different two-component vectors respectively, in which first components are equal to 1, and second components are different from one another, ranging from -2 to 2 at 0.2 unit distance intervals. The ordering of assignment of amino acids to vectors is determined mainly by the pK_a (NH_3^+) values and secondarily by the pK_a ($COOH$) value. For example, the pK_a (NH_3^+) value of the amino acid, Cyteine is the largest among the sets of pK_a (NH_3^+) values of 20 amino acids, so a vector (1, 2.0) is assigned to the amino acids, C. Similarly, that of Aspargine ,N is the smallest ,and then a vector (1,-2.0) to N. pK_a (NH_3^+) values of Leucine, L and Glycine, G are equal ,but the pK_a ($COOH$)value of L is bigger than that of G, so a vector(1,0.4) corresponds to L and (1,0.2) to G. The detailed assignment is shown in Fig. 1. Obviously, the assignment is only associated with the pK_a (NH_3^+) values and the pK_a ($COOH$) values, and avoids the arbitrariness in the Ref. [23]. Therefore, the graphical representations based on the assignment help in differentiating the physico-chemical properties from DNA sequences, and then understanding different biochemical functions of DNA sequences.

2.2 Method to construct the H-L graph

Each amino acid of a protein sequence has a walk sequentially in 2D-space as one of the above twenty vectors, so that we obtain a curve uniquely corresponding to the protein sequence, which we call an H-L curve. We describe the process below using a model example. Let us consider an amino acid sequence s=CYDGNWVVFIKPEQT. The method for constructing its curve is shown in Fig. 2. The first vector of the sequence corresponding to the first amino acid, C starts from the origin of the coordinate system(0,0) denoted as ‘start’ in the Fig. 2. Then, next vector in turn starts at the end of previous vector up to the last vector of the protein sequence. Finally, we obtain an H-L curve uniquely in Fig. 2. The strategy of walking for vectors is similar with the original method of plot DNA sequence as a walk in

2D-space using four orthogonal directions that represent the four bases was introduced in Refs. [4,27,28].

Table 1. The acid dissociation constants of side chains of amino acid (taken from the Refs. [25-26]).

Amino acid	3-Letter code	1-Letter code	$pK_a(COOH)$	$pK_a(NH_3^+)$
Small hydrophilic				
Glycine	Gly	G	2.34	9.60
Alanine	Ala	A	2.34	9.69
Threonine	Thr	T	2.63	10.43
Serine	Ser	S	2.21	9.15
Proline	Pro	P	1.99	10.60
Small hydrophibic				
Valine	Val	V	2.32	9.62
Leucine	Leu	L	2.36	9.60
Isoleucine	Ile	I	2.36	9.68
Methionine	Met	M	2.28	9.21
Aromatic				
Phenylalanine	Phe	F	1.83	9.13
Tyrosine	Try	Y	2.20	9.11
Tryptophan	Trp	W	2.38	9.39
Acid and their amides				
Aspartic	Asp	D	2.09	9.82
Glutamic acid	Glu	E	2.19	9.67
Aspargine	Asn	N	2.02	8.80
Glutamine	Gln	Q	2.17	9.13
Bass				
Lysine	Lys	K	2.18	8.95
Arginine	Arg	R	2.17	9.04
Histidine	His	H	1.82	9.17
Sulphydryl				
Cyteine	Cys	C	1.71	10.78

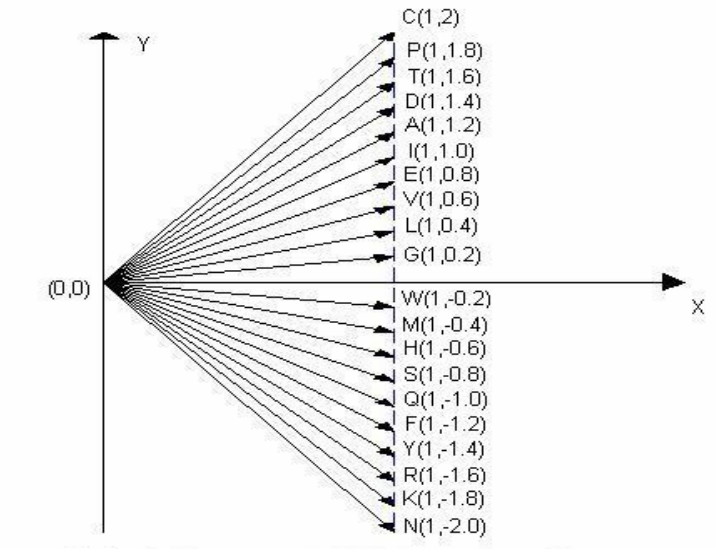

Fig. 1.The assignment of 20 different vectors to 20 kinds of amino acid

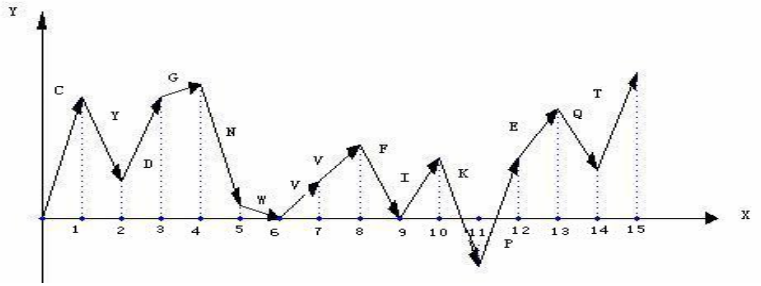

Fig. 2.The method for constructing the H-L curve of the amino acid sequence s=CYDGNWVVFIKPEQT.

In Fig. 3, we illustrated the H-L curves of Human and Orangutan HLA class I histocompatibility antigens, A-1 alpha chains, whose Primary accession numbers in Swiss-Prot are P30443 and P16211 respectively and whose Entry name are 1A01_HUMAN and 1A01_PONPY respectively. These long-range distinct patterns can be recognized visually, which reflect the differences between sequences.

2.3 The properties of H-L curves

Property 1. There is not any circuit in our 2-D graphical representation.

Proof. We assume that there exists a circuit in Fig. 4 with n vectors $\overrightarrow{a_1}$, $\overrightarrow{a_2},\cdots,\overrightarrow{a_n}$ corresponding to one of 20 kinds of amino acid respectively. It stands to reason that the equation as follows holds:

$$\overrightarrow{a_1}+\overrightarrow{a_2}+\cdots+\overrightarrow{a_n}=\vec{0}=(0,0) \quad . \tag{1}$$

then

$$\begin{cases} 1+1+\cdots+1=n=0 \\ y_1+y_2+\cdots+y_n=0 \end{cases} , \tag{2}$$

where $\overrightarrow{a_i}=(1,y_i), i=1,2,\cdots,n$. Therefore, n=0. That is to say, it is sure that there is no circuit in our representation.

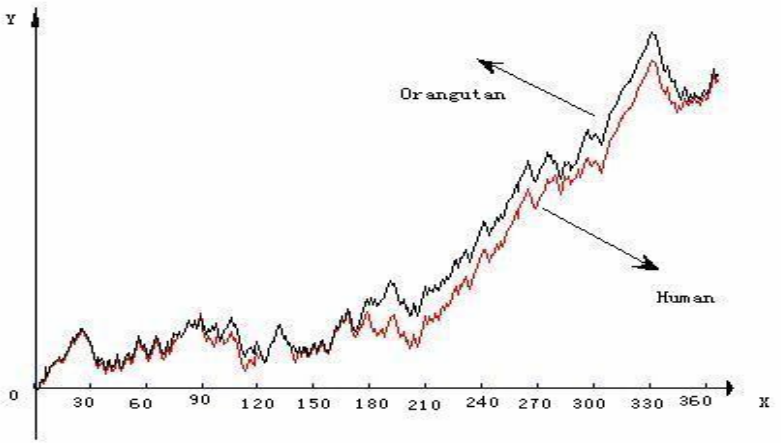

Fig. 3.Two H-L curves of amino acid sequences from Human and Orangutan respectively.

Property 2. The protein sequences correspond to the graphical H-L curves in a one to one manner.

Proof. It is easily proven that a protein sequence allow us to construct a unique graphical curve by the method introduced in section 2.2. On the contrary, given a graphical H-L curve, we deduce doubtless a unique sequence of amino acid from the ordering vector sequence from

the H-L curve in terms of the assignment listed in Table 1. Therefore, the H-L curves are associated with the protein sequences one to one.

The property 2 show us that there is no loss of information accompanying the H-L curve of graphical representation of the primary structure of a protein sequence and consequently the 2-D graphical representation of a protein sequence allow the sequence reconstruction .So each of the H-L curve can be taken as a graphical 'signature' of the corresponding sequence.

Property 3. The starting coordinate and the ending coordinate of the kth vector corresponding to the kth amino acid of a protein sequence are computed respectively by the formula as follows:

$$S(k) = (0,0) + \sum_{i=1}^{k-1} \overrightarrow{a_i} \tag{3}$$

and

$$E(k) = \sum_{i=1}^{k} \overrightarrow{a_i}, \tag{4}$$

namely,

$$\begin{cases} S_x(k) = 0 + \sum_{i=1}^{k-1} 1 \\ S_y(k) = 0 + \sum_{i=1}^{k-1} y_i \end{cases} \tag{5}$$

and

$$\begin{cases} E_x(k) = \sum_{i=1}^{k} 1 = k \\ E_y(k) = \sum_{i=1}^{k} y_i \end{cases}, \tag{6}$$

where $\overrightarrow{a_i} = (1, y_i)$ and $S(k) = (S_x(k), S_y(k))$, $E(k) = (E_x(k), E_y(k))$, for an arbitrary integer k, ,$1 \le k \le n$ (n is the length of the protein sequence).

Proof. (Mathematical Induction) It goes without saying that $S(1) = (0,0)$ and $E(1) = \overrightarrow{a_1}$, so S(1) and E(1) are true. Assume that, for an arbitrary k, S(k) and E(k) are also true, i.e. $S(k) = (0,0) + \sum_{i=1}^{k-1} \overrightarrow{a_i}$ and $E(k) = \sum_{i=1}^{k} \overrightarrow{a_i}$. Let's derive S(k+1) and E(k+1) from this

assumption. According to the method of constructing the curve, we have S(k+1)=E(k) and E(k+1)=S(k+1)+$\overrightarrow{a_{k+1}}$, namely,

and

$$
\begin{aligned}
S(k+1) &= E(k) \\
&= \sum_{i=1}^{k} \overrightarrow{a_i} \\
&= (0,0) + \sum_{i=1}^{(k+1)-1} \overrightarrow{a_i} \\
E(k+1) &= S(k+1) + \overrightarrow{a_{k+1}} \\
&= \sum_{i=1}^{k} \overrightarrow{a_i} + \overrightarrow{a_{k+1}} \\
&= \sum_{i=1}^{k+1} \overrightarrow{a_i}
\end{aligned}
$$

which exactly mean that S(k+1) and E(k+1) hold. Therefore, S(k) and E(k) are true for all integers k equal to or greater than 1 and equal to or lesser than n.

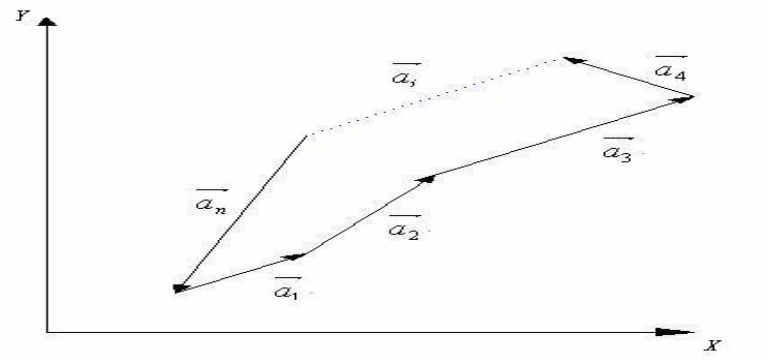

Fig. 4. A circuit with n vectors.

3. Graphical alignment

3．1 graphical alignment

Sequence alignment is one of the crucial sequence-comparison operations in bioinformatics and genetic research [13]. An alignment of two sequences s1 and s2 is obtained by insertion of chosen gaps, either into or at the ends of s1 and s2, and then placing the two resulting sequences one over the other so that every letter in either sequence is opposite a letter or a gap in the other sequence [13]. The longer aligning sequences are, the more difficultly it performs. What's more, the particular alignment program doesn't produce the best alignment result, as seen in Ref. [13]. However, even a superficial view of the graphical representation shown in Fig. 4 clearly reveals on one hand a considerable overall similarity among sequences and on the other hand the existence of local differences in their primary structure. Obviously, graphical representation enables more easily visual inspection of sequences than their representation by strings over the DNA alphabet{A, T, G, C} or the amino acid alphabet {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. So in recent years, the novel graphical alignments of sequences have been presented in Refs. [13,23,25]. Randić[13] proposed graphical alignment of DNA ,which is performed by simple subtracting the corresponding numerical sequences and looking at spots having zero difference. The alignments based on graphical representations were adopted for protein sequences in Ref. [23,25]. The advantage is that it allows visual inspection of alignment, performs very easily. In the following, we introduced the theory of graphical alignment, and then illustrated our approach on two examples.

3.2 Theories of graphical alignment

Theorem 1 The necessary and sufficient conditions of matching for the kth amino acids of two protein sequences are that the two kth vectors respectively from the corresponding graphs have the same directions and lengths.

Proof. (Sufficiency) Assume that $\vec{a}_k$ and $\overrightarrow{b_k}$ are respectively the two kth vectors of graphical representations of the kth amino acids a_k and b_k of two protein sequences a and

b. If $\vec{a}_k$ runs parallel to $\overrightarrow{b_k}$ and the lengths of $\vec{a}_k$ is equal to that of $\overrightarrow{b_k}$, according to the assignment of 20 kinds of amino acid to 20 vectors listed in Table 1, we draw a conclusion that the kth amino acids of two protein sequences match.

(Necessary) Assume also that $\vec{a}_k$ and $\overrightarrow{b_k}$ are respectively the two kth vectors of graphical representations of the kth amino acids a_k and b_k of two protein sequences a and b. If $\vec{a}_k$ doesn't run parallel to $\overrightarrow{b_k}$ or the lengths of $\vec{a}_k$ isn't equal to that of $\overrightarrow{b_k}$, no matter which kinds of amino acids two vectors are assigned to ,we always conclude that the kth amino acids between two sequences don't match.

Corollary 1 Suppose that A and B are respectively starting and end dots of the kth vector of graphical curves of the protein sequence a, and that C and D the protein sequence b. The quadrangle ACDB being a parallelogram is a sufficient and necessary condition for the kth amino acids of two protein sequences matching .

Corollary 2 Suppose that Ai and Bi are respectively starting and end dots of the kth vector of graphical representations of the protein sequence Si (i=1,2,…,N) ,where N is the number of sequences. All quadrangles AiAi+1Bi+1Bi being parallelograms is a sufficient and necessary condition for the kth amino acids of N protein sequences matching.

3.3 Illustration

Now, according to Corollary 1 and Corollary 2, we can quickly observe similarities among the protein sequences, which amino acids match and which amino acids mismatch using graphical H-L curves of protein sequences. Take two amino acid sequences taken from Ref. [25] for example:

Protein I: WTFESRNDPAKDPVILWLNGGPGCSSLTGL

Protein II:WFFESRNDPANDPIILWLNGGPGCSSFTGL. Their corresponding graphical curves are shown respectively in Fig. 5. Obviously, Fig. 5 directly shows more information of alignment without further processing, compared with Fig. 2 in Ref. [25]. On seeing Fig. 5, one can observe the facts that amino acids at the sites 2,11,14 and 27 mismatch, and that those

at the other sites match. The result is uniform with that in Ref. [25]. What' more, it is more convenient to make multiple sequences alignment by our method than by that in Ref. [25]. For example, three amino acid sequences are respectively as follows:

Protein I: PNTAHHRNCCWWTYSKIFFC,

Protein II:PNTHHRRCCCWWTYLKIFFC,

Protein III:PNTNHRVCCCWWYYNKIDFC.

Their corresponding curves are shown in Fig. 6. Utilizing our graphical alignment theory, one can obviously judge whether the amino acids at corresponding sites match or not. In Fig. 6, we can conclude by parallelogram that amino acids at the sites 4,6,7,8,13,15 and 18 among three amino acid sequences match ,and those at the other sites mismatch.

There are significant differences between our approach described here and those of Refs. [13,23,25], even though the alignment result may appear similar. First, the graphical alignments in the Refs.[13,23,25] were performed by the coordinate of dots assigned to each amino acid , while in this work we use vectors assigned to each amino acid to make graphical alignment for the first time. Second, Fig. 2 in Ref. [25] appeared uninformative and cannot be used to make alignment without further processing, while our graphical representations directly show the alignment information, and hence is more simple and efficient. Third, our approach is in particular suitable for making multiple sequences alignment as shown in Fig. 6. That is to say, the more sequences there are, the more efficient alignment there are.

4. Conclusion

In the paper, we assign 20 different vectors to 20 kinds of amino acid, according to the pK_a (NH_3^+) values and the pK_a ($COOH$) value of amino acids. We construct the 2-D graphical representations of protein sequences by the method introduced in Refs. [4,27]. Our graphical representations have no circuit and correspond to protein sequences one to one, which advantage allows visual inspection of sequence data, helping in recognizing major similarities among different protein sequences. In addition, it also facilitates the quantitative comparison among different protein sequences, which is being studied.

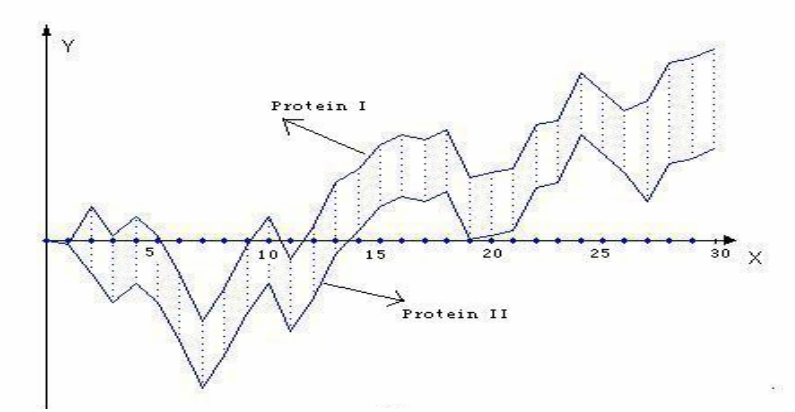

Fig.5. H-L curves of Protein I and Protein II in the first example.

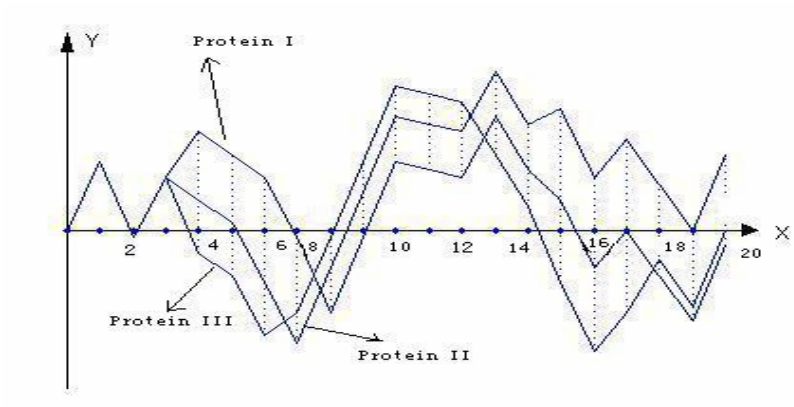

Fig. 6. H-L curves of Protein I, Protein II and Protein III in the second example.

5. **Acknowledgement**

This work is supported in part by the National Nature Science Foundation of China (Grant 10571019). The authors thank the anonymous referees for many valuable suggestions that have improved this manuscript.

References

[1] E. Hamori, J. Rukin, A novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. **258** (1983) 1318-1327.

[2] E. Hamori, Novel DNA sequence representation, Nature **314** (1985) 585-586.

[3] M. A. Gates, Simpler DNA sequence representations, Nature **316** (1985) 219-219.

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

[5] M. Randić, A. F. Kleiner, L. M. DeAlba, Distance/distance matrices, J. Chem. Inf. Comput. Sci. **34** (1994) 277-286.

[6] 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.

[7] 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.

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

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

[10] B. Liao, R. Li, W. Zhu, On the similarity of DNA primary sequences based on 5-D representation, J. Math. Chem. **42** (2007) 47-57.

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

[12] B.H.Zhang, H.S.Wang, L.Xu, Comparison for DNA primar ysequence, MATCH Commun. Math. Comput. Chem. 58(2007) 559-568.

[13] M. Randić, J. Zupan, D. Vikić-Topić, D. Plavšić, A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequences, Chem. Phys. Lett. **431** (2006) 375-379.

[14] G. Huang, B. Liao,W. Zhang, F. Gong, A Method for Sequences Alignment and Mutation Analysis, MATCH Commun. Math. Comput. Chem. **59** (2008) 635-645.

[15] W.Chen, B.Liao, Y.Liu, W.Zhu, Z.Su, A numerical representation of DNA sequences and its applications, MATCH Commun. Math. Comput.Chem.**60** (2008) in press

[16] M. Randić, J. Zupan, Highly compact 2-D graphical representation of DNA sequences.SAR QSAR Environ Res.**15** (2004) 191-205.

[17] M. Randić, J. Zupan, Vikić-Topić, On representation of proteins by star-like graphs J. Mol. Graphics & Model. **26** (2007) 290-305.

[18] F. Bai, T. Wang, A 2-D graphical representation of protein sequences based on nucleotide triplet codons, Chem. Phys. Lett. **413** (2005)458- 462

[19] M. Randić, J. Zupan, A.T. Balaban, Unique graphical representation of protein sequences based on nucleotide triplet codons, Chem. Phys. Lett.**397** (2004) 247–252.

[20] M. Randić, A.T. Balaban, M. Novič, A. Založnik, T. Pisanski, A novel graphical representation of proteins, Periodicum Biologorum **107** (2005) 403– 414.

[21] M. Randić, M. Novič, D. Vikić-Topić, D. Plavšić, Novel numerical and graphical representation of DNA sequences and proteins, SAR QSAR Environ. Res. **17** (2006) 1–13.

[22] M. Randić, Spectrum-like graphical representation of DNA based on codons, Acta Chim. Slovenica **53** (2006) 477–485.

[23] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins, Chem. Phys. Lett. **419** (2006) 528–532.

[24] F. Bai, T. Wang, On graphical and numerical representation of protein sequences, J. Biomol. Struct. Dyn. **23** (2006) 537–545.

[25] M. Randić, 2-D Graphical representation of proteins based on physico- chemical properties of amino acids, Chem. Phys. Lett. **440** (2007) 291-295

[26] http://cds.unina.it/petrilli/aminoac/amiac.htm?htm.

[27] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. methodology and application to globin genes, Current Science **66** (1994) 309-314.

[28] D. Bielińska-Waz ,T. Clark , P. Waz , W. Nowak , A. Nandy, 2D-dynamic representation of DNA sequences. Chem .Phys .Lett . **442** (2007) 140–144.