International Journal of Combinatorial GraphTheory and Applications
Vol. 6, No. 1, (January-June, 2021)

Reviewed by: Hua Wang

Received: 10th March 2021 *Revised: 15th April 2021* *Accepted: 05th May 2021*

A New 2-D Graphical Representation of Protein Sequences

Ying Guo, Tian-ming Wang

Department of Applied Mathematics,
Dalian University of Technology, Dalian 116024, P. R. China
e-mail: guoying@163.com

Abstract. In this paper, we propose a new approach to represent protein sequences through a 2-D cartesian coordinate system. By the 2-D graphical representation, we make a comparison of sequences from eight different species. Then we construct phylogenetic tree based on the approach.

1. Introduction

With the completion of more and more genome projects, a number of genomic sequences are available in public databases. Mathematical analysis of the large volume genomic DNA sequences, RNA secondary structure and protein sequence data is one of the challenges for bio-scientists. Now there are many methods of phylogenetic analysis biological sequences. Some algorithms are to calculate a matrix representing the distance between each pair of sequences and then transform this matrix into a tree, other type of approaches, instead of building a tree, the tree is found by evaluating the fitness of different topologies, such as parsimony and maximum likelihood method. Graphical representation provides a simple way of viewing, sorting and comparing various gene structures. As a tool, the graph which is representing a large number of genomic sequences can be studied in a perceivable form. We outlined several recent novel graphical representations of DNA sequences and RNA sequences in the literature[1-10], which have led to novel quantitative characterization and mathematical characterization of proteome maps[11-13]. In contract to DNA and RNA sequences, the approaches of graphical representations of protein sequences are developed slowly. Recently highly condensed graphical

(

[0] MR(2000) subject catalogue: 92D20, 92C40.

[0] Keywords: protein sequence, graphic representation, phylogenetic tree.

representation of proteins was considered[14-15]. Randic[16] presented a novel graphical representation of proteins. The approach is based on a graphical representation of triplets of DNA in which the interior of a square or the interior of a tetrahedron is used to accommodate 64 sets for the 64 codons. Then an 8×8 tabular representation of amino acids can construct the associated zigzag curves. Fenglan Bai[17] presented a 3-D graphical representation protein sequences. She assigned the 20 amino acids to 20 vertices of the regular dodecahedron the center of which is located at origin of the Cartesian coordinate system.

In this letter, we introduce a new 2-D graphical representation approach to represent protein sequences. We assign the the 20 amino acids to 20 different vertices in the 2-D cartesian coordinate system. The graph of the 20 amino acids is shown in Figure 1. Using the approach, we make a comparison of similarity/dismilarity of the sequences of eight different species through E matrix, M/M matrix and L/L matrix. We come to a good conclusion of the eight species of phylogenetic tree.

2. A 2-D Representation of Proteins

In table 1, the full names and symbols of the 20 amino acids are outlined, according to alphabetical order.

Table 1 the full names and the symbols of the 20 Amino acids

Full name	Three symbol	One symbol	Full name	Three symbol	One symbol
Alanine	ala	A	Leucine	leu	L
Arginine	arg	R	Lysine	lys	K
Asparagine	arg	N	Mathionine	met	M
Aspartic	asp	D	Phenylabanine	phe	F
Cysteine	cys	C	Proline	pro	P
Histidine	his	H	Serine	ser	S
Glutamine	gln	Q	Threonine	thr	T
Glutamic	glu	E	Triptophan	trp	W
Glycine	gly	G	Tyrosine	tyr	Y
Isoleucine	ile	I	Valine	val	V

We divide an unite circle into 20 equal sections. The 20 amino acids are represented by the 20 unit vectors which are in the unite circle. The assignment of twenty amino acids in Figure 1 is of course arbitrary, leading to 20! possible different coding schemes. The unite vectors representing 20 amino acids are as follows:

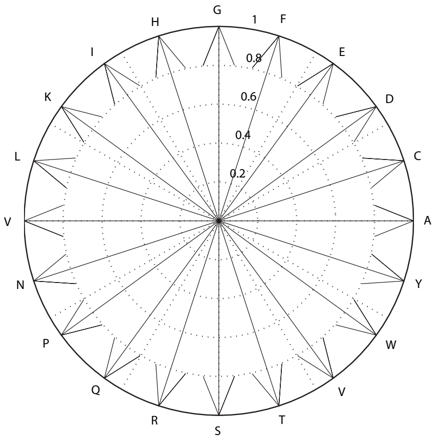

Figure 1

Then the unit vectors representing 20 amino acids are as follows (here θ=$\Pi/10$):

$A \rightarrow v_1 = (cos0,\ sin0)$

$C \rightarrow v_2 = (cos\theta,\ sin\theta)$

$D \rightarrow v_3 = (cos2\theta,\ sin2\theta)$

$E \rightarrow v_4 = (cos3\theta,\ sin3\theta)$

$F \rightarrow v_5 = (cos4\theta,\ sin4\theta)$

$G \rightarrow v_6 = (cos5\theta,\ sin5\theta)$

$H \rightarrow v_7 = (cos6\theta,\ sin6\theta)$

$I \rightarrow v_8 = (cos7\theta,\ sin7\theta)$

$K \rightarrow v_9 = (cos8\theta,\ sin8\theta)$

$L \rightarrow v_{10} = (cos9\theta,\ sin9\theta)$

$V \rightarrow v_{11} = (cos10,\ sin10)$

$N \rightarrow v_{12} = (cos11\theta,\ sin11\theta)$

$P \rightarrow v_{13} = (cos12\theta,\ sin12\theta)$

$Q \rightarrow v_{14} = (cos13\theta,\ sin13\theta)$

$R \rightarrow v_{15} = (cos13\theta,\ sin13\theta)$

$S \rightarrow v_{16} = (cos15\theta,\ sin15\theta)$

$T \rightarrow v_{17} = (cos16\theta,\ sin16\theta)$

$V \rightarrow v_{18} = (cos17\theta,\ sin17\theta)$

$W \rightarrow v_{19} = (cos18\theta,\ sin18\theta)$

$Y \rightarrow v_{20} = (cos19\theta,\ sin19\theta)$

Like DNA sequences, protein sequences can also be seen as a string on the alphabet set $\sum = \{A,\ R,\ N,\ D,\ C,\ H,\ Q,\ E,\ G,\ I,\ L,\ K,\ M,\ F,\ P,\ S,\ T,\ W,\ W,\ Y,\ V\}$ of 20 amino acids. So a protein sequence can become a numerical sequence:

$$p(n) = p(0) + \sum_{i=1}^{n} s(i), \quad p(0) = (0, 0), \tag{2.1}$$

where s(i) satisfies the following conditions:

$$s(i) = \begin{cases} (\cos 0, \sin 0) & if \quad i = A, \\ (\cos\theta, \sin\theta) & if \quad i = C, \\ \quad\vdots & \\ (\cos 18\theta, \sin 18\theta) & if \quad i = W, \\ (\cos 19\theta, \sin 19\theta) & if \quad i = Y. \end{cases} \tag{2.2}$$

(i=0,1,2,$\cdots$,n. where n is the length of the protein sequence being studied.) So point $p_i(x_i,y_i)$(i=0,1,2,$\cdots$,n) can represent the following form:

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} \cos 0, \ \cos\theta, \ \cos 2\theta \ \cdots \ \cos 18\theta, \ \cos 19\theta \\ \sin 0, \ \sin\theta, \ \cos 2\theta \ \cdots \ \cos 18\theta, \ \cos 19\theta \end{pmatrix} \begin{pmatrix} A_i \\ C_i \\ D_i \\ \vdots \\ W_i \\ Y_i \end{pmatrix} \tag{2.3}$$

where A_i, C_i, D_i, $\cdots$, W_i, Y_i are the cumulative occurrence numbers of A, C, D, $\cdots$, W, Y in the subsequence from the 1^{st} amino acid to the i^{th} amino acid of the sequence, we define $A_0 = C_0 \cdots = W_0 = Y_0$. With i running from 1 to n, we have the points P_0, P_1, $\cdots$, P_n. Connecting points, P_0, P_1, $\cdots$, P_n one by one, we can get the appropriate curve which is called the 2-D curve of the protein sequence. The points P_0, P_1, $\cdots$, P_n are referred to as the nodes of the 2-D curve.

For example, in Figure 2, by using the formula(2.3) we present the 2-D curve of the ND6 protein sequence of human:

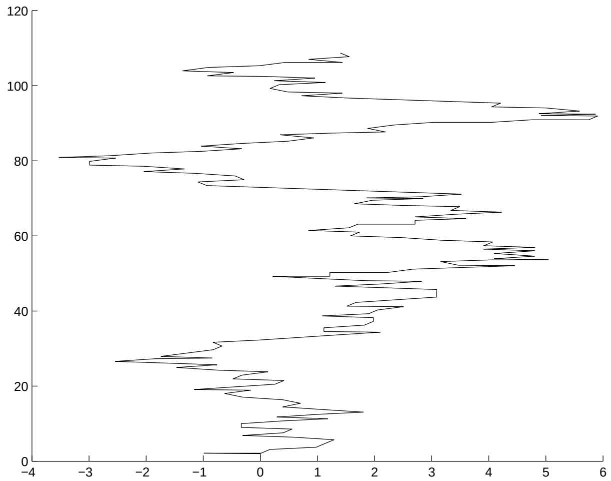

Figure 2

3. Similarity

In this section, we apply the 2-D curves to analyzing the similarity/dissimilarity of the ND6 protein sequences of eight species in order to examine the validity of our new approach. Using the similarity distance matrix we got, we draw the phylogenetic tree of the eight species by suiting clustering method. The results fix realities quite well.

The sequences are listed in Table 2.

Table 2 The protein sequences and their lengths of ND6 of 8 species

Species	Coding sequence
human	MMYALFLLSVGLVMGFVGFSSKPSPIYGGLVLIVSGVVGCVIILNFGGGYMGLMVFLIYLGGMMVVFGYTTAMAIEEYPEAWGSGVEVLVSVLVGLAMEVGFVLWVKEYDGVVVVVNFNSVGSWMIYEGEGSGFIREDPIGAGALYDYGRWLVVVTGWPLFVGVYIVIEIARGN
Opossum	MKMMTIYIISLLLMIGFVAFASKPSPIYGGLSLVVSGGLGCGMVVSLEDVFLGLVVFLVYLGGMLVVFGYTTAMATEEYPETWVGNVVAFIMLLFVLLLQVGWYFMSKLVYIIMAIKLFDFVETSLVGQDYNGVSQLYYCGGWALALLGWILFMTIYVVLEVVRERSY
H.seal	MMTYIVFILSIIFVVSFVGFSSKPSPIYGGLVLIISGAVGCGIVLSFGGSFLGLMVFLIYLGGMLVVFGYTTAMAIEQYPEVWVSNKAVLGAFVMGLLSELLLACYILKDDEVDVVFEFNGMGDWVIYDTGDSGFFSEEAMGIAALYSYGTWLVIVTGWSLLTGVLVIMEVTRGN
G.seal	MMTYIVFILSIIFVISFVGFSSKPSPIYGGLVLIISGAVGCGIVLSFGGSFLGLMVFLIYLGGMLVVFGYTTAMATEQYPEVWVSNKAVLGAFVMGLLSELLLACYILKDDEVDAVFEFNGMGDWVIYDTGDSGFFSEEAMGIAALYSYGTWLVIVTGWSLFIGVLVIMEVTRGN
Rat	MTNYMFILSLLFLTGCLGLALKPSPIYGGFGLIVSGCIGCLMVLGFGGSFLGLMVFLIYLGGMLVVFGYTTAMATEEYPETWGSNWFIFSFFVLGLFMELVVFYLFSLNNKVELVDFDSLGDWLMYEIDDVGVMLEGGIGVAAIYSCATWMMVVAGWSLFAGIFIIIEITRD
Mouse	MNNYIFVLSSLFLVGCLGLALKPSPIYGGLGLIVSGFVGCLMVLGFGGSFLGLMVFLIYLGGMLVVFGYTTAMATEEYPETWGSNWLILGFLVLGVIMEVFLICVLNYYDEVGVINLDGLGDWLMYEVDDVGVMLEGGIGVAAMYSCATWMMVVAGWSLFAGIFIIIEITRD
Gorilla	MTYVLFLLSVGLVMGFVGFSSKPSPIYGGLVLIVSGVVGCAIILNCGGGYMGLMVFLIYLGGMMVVFGYTTAMAIGEYPEAWGSGVEVLVSVLVGLAMEVGLVLWVKEYDGVVVVVNFNNVGSWMIYEGEGSGLIREDPIGAGALYDYGRWLVVVTGWTLFVGVYIVIEIARGN
Chimpanzee	MTYALFLLSVSLVMGFVGFSSKPSPIYGGLVLIVSGVVGCAIILNYGGGYMGLMVFLIYLGGMMVVFGYTTAMAIEEYPEAWGSGVEVLVSVLVGLAMEVGLVLWVKGYDGMVVVVNFNSVGSWMIYEGEGPGLIREDPIGAGALYDYGRWLVVVTGWTLFVGVYIVIEIARGN

We can get the 2-D curve by using the formula(2.1). In Table 3, we give the similarities and dissimilarities distance matrix for the eight species based on the first 10 leading eigenvalues of E matrix of each curve.

Table 3

The relative similarity/dissimilarity matrix for the coding sequences of Table 1 based on the formula (3)

Species	Human	Opossum	H.seal	G.seal	Rat	Mouse	Gorilla	Chimpanzee
Human	0	0.2621	0.2254	0.2038	0.2395	0.1568	0.0378	0.0646
Opossum		0	0.4875	0.4659	0.5016	0.4189	0.2999	0.3267
H.seal			0	0.0216	0.0141	0.0686	0.1876	0.1608
G.seal				0	0.0357	0.0470	0.1660	0.1392
Rat					0	0.0827	0.2017	0.1749
Mouse						0	0.1190	0.0922
Gorilla							0	0.0268
Chimpanzee								0

In Table 4, the similarities matrix is given based on the first 10 leading eigenvalues of L/L matrix. In this paper, L/L matrix is equal to the M/M matrix.

Table 4

The relative similarity/dissimilarity distance matrix for the coding sequences of Table 1 based on the formula (3)

Species	Human	Opossum	H.seal	G.seal	Rat	Mouse	Gorilla	Chimpanzee
Human	0	1.7072	3.2896	3.0119	4.8497	3.4514	0.8973	1.1403
Opossum		0	3.2896	4.3543	6.1000	4.7067	2.1915	2.3687
H.seal			0	0.7081	2.0682	0.8077	2.5448	2.3247
G.seal				0	2.0145	1.0200	2.3333	2.0934
Rat					0	1.9655	4.1956	3.8823
Mouse						0	2.7024	2.5389
Gorilla							0	0.5345
Chimpanzee								0

In order to see the similarity and dissimilarity among the eight protein sequences intuitively, we transform the above data into phylogenetic tree by UPGMA method.

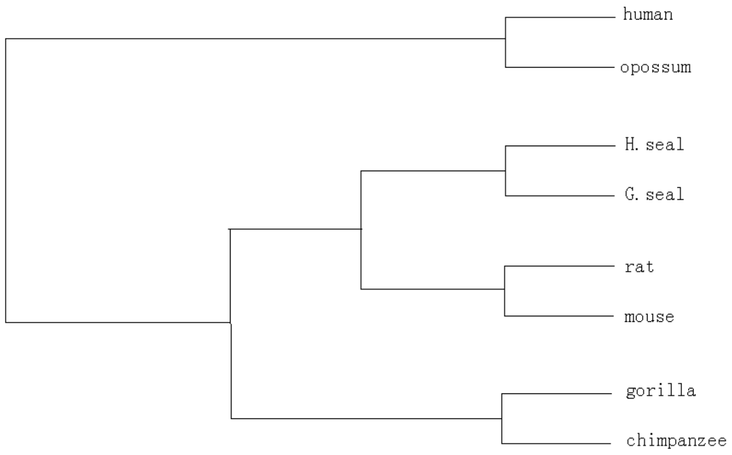

Figure 3: Table 3 corresponding sequence phylogenetic graph

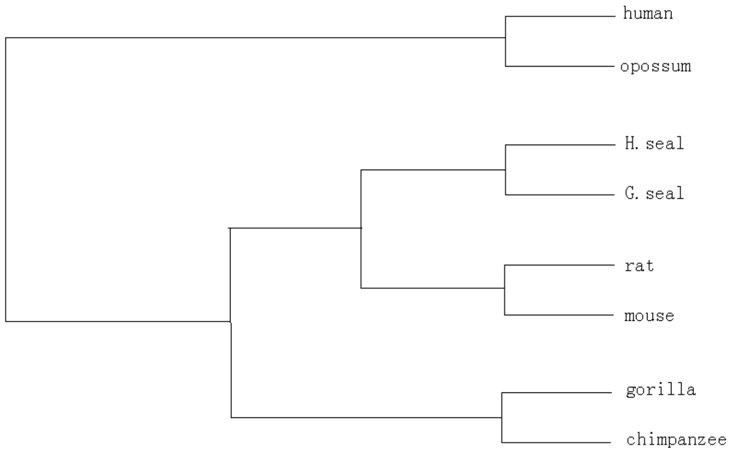

Figure 4: Table 4 corresponding sequence phylogenetic graph

We got a consistent result according to the table 3 and table 4. From the phylogenetic graph, we can see that H.seal and G.seal, rat and mouse, gorilla and chimpanzee are the most close. Then human and opossums to them are more close. From the figures, we find our results about similarity fix basically the reality. Our approach gives numerical characterization of protein sequences by graphic representation and used to similarity analysis of protein sequences. The method is simpler, more convenient, and faster.

4. Conclusions

With the completion of more and more genome projects, a number of genomic sequences are available. It's of importance to analyze these sequences and find regions that have known structure and functions. We present a 2-D representation of protein sequences. The advantage of our approach is that it allows visual inspection of data, helping in recognizing major similarities among different protein sequences. Furthermore, we can construct phylogenetic tree of species.

Acknowledgments

We appreciate the financial support of this work that was provided by China National Natural Science Foundation with the contract No. 10571019.

References

[1] 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..

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

[3] X. Guo, M. Randic, and S. C. Basak, A novel 2D graphical representation of DNA sequences of low degenracy, Chem. Phys. Lett., 350 (2001), 106-112.

[4] M. Randic, and S. C. Basak, Characterization of DNA primary. sequences based on the average distance between bases, J. Chem. Inf. Comput. Sci., 41 (2001), 561-568.

[5] A. Nandy, and P. Nandy, On the Uniqueness of Quantitative DNA Difference Descriptors in 2D Graphical. Representation Models, Chem. Phys. Lett., 368 (2003), 102-107.

[6] M. Randic, and M. Vracko, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem. Phys. Lett., 371 (2003), 202-207.

[7] C. X. Yuan, B. Liao, and T. M. Wang, New 3D graphical representation of DNA sequence and their numerical characterization, Chem. Phys. Lett., 379 (2003) 412-417.

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

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

[10] F. L. Bai, W. Zhu, and T. M. Wang, Analysis of similarity between RNA secondary structures, Chem. Phys. Lett., 408 (2005), 258-263.

[11] M. Randic, Quantitative Charaterization of proteomics Maps by Matrix Invariants. In Handlbook of Proteomic Methods (eds., Conn, P. M.), Humana Press Inc., Totowa, NJ 2003, 429-450.

[12] M. Randic, M. Vracko, and M. Novic, On characterisation of dose variations of 2-D proteomic maps by matrix invariants, J. Proteome. Res., 1 (2002), 217-226.

[13] M. Randic, On the graphical representation of proteomics and their numerical characterization, J. Chem. Inf. Comput. Sci., 41 (2001), 1330-1338.

[14] M. Randic, 2-D graphical representation of proteins based on virtual genetic code, SAR. QSAR. Environ. Res., 15 (2004), 147-157.

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

[16] M. Randic, and J. Zupan, Unique graphical representation of protein sequences based on nucleotide triplet codons, Chem. Phys. Lett., 397 (2004), 247-252.

[17] F. L. Bai, and T.M. Wang, On graphical and numerical representation of protein sequences, J. Biomol. Str. Dyn., 23 (2006), 537-545.