MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **75** (2016) 431-446

ISSN 0340 - 6253

A Novel Method of 2D Graphical Representation for Proteins and Its Application

Dandan Sun[a] **, Chunrui Xu**[a]**, Yusen Zhang**[a,1]

[a] *School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China.*

(Received July 25, 2015)

Abstract. In this paper, we propose the graph energy of 20 amino acids and the 2D graphical representation of protein sequences based on six physicochemical properties of 20 amino acids and the relationship between them. Moreover, we could get a specific vector from the graphical curve of a protein sequence, and use this vector to calculate the distance between two sequences. This approach avoids considering the differences in length of protein sequences. Finally, we research the similarities/dissimilarities of ND5 and 36PDs using our method and get better results compared with ClustalX2.

1 Introduction

As the number of biological sequences increases fast in the public databases because of the rapid development of sequencing techniques, how to infer the potential information of a large number of sequences effectively and accurately becomes a critical challenge in biological information. Therefore, many valid methods in information extraction from DNA, RNA or protein sequences are proposed [1-6]. We all know that proteins, encoded by DNA, determine the material basis of an organism's anatomy and physiology. Thus, detecting the similarity of proteins is definitely important [7-9], especially considering the structure and function of proteins.

Models of protein analysis can be divided into two classes: sequence alignment [10-12] and alignment-free sequence comparison. The former applies a score function to represent deletion, insertion and substitution among amino acids in protein sequences comparison.

[1]Corresponding author: zhangys@sdu.edu.cn

But there are some limitations in sequence alignment fixing on computation complexity and the fact that some sequences lack significant conserved domains. On the other hand, alignment-free methods like graphical representation of sequences are able to overcome these limitations [13-17]. The graphical representation of protein sequences, developing from graphical representation for DNA, is usually implemented by letter sequence representation (LSR). As protein sequence is composed of 20 amino acids, the graphic representation of protein sequences is certainly more complicated than that of DNA or RNA [18-20].

Wu et al. [9] proposed a graphical representation of protein sequences on the basis of codons encoding the amino acids and calculated the graph energy and Laplacian energy of 20 amino acids respectively. Graph energy was well applied with the unique 2-D graphical representation in this paper and the result is good comparing with ClustalW.

Instead of utilizing codons, we proposed a novel graphical representation for protein on the basis of 6 typical physicochemical properties of amino acids and obtained the graph energy of 20 amino acids via the relationship between amino acids, which is more visual and reasonable. Considering the difficulty in dealing with sequences with different lengths, the advantage of our method is obvious. Without using complicated slipping window in reference [9], we handle this problem by moment vector, which is easy and well used in 2-D graphic representation.

The main strong points of our method are as follows:

1. In our method, 6 typical properties [21-24] are considered to construct representative graph for every amino acid. The physicochemical properties of amino acids are more important than other factors in determining the rate and pattern of protein evolution [25]. Therefore, these properties have a direct and significant impact on estimation of distance between two polypeptide sequences [26].

2. The curve of a protein sequence is obtained from the application of the Gutman's graph energy [27, 35-38]. The energy of graph is meaningful for analysis of graphs, and it is fit for our unique construction of graphs for 20 amino acids.

3. The moment vector of a protein sequence was successfully applied in a few researches [28], and these authors have demonstrated that the correspondence between a protein sequence and its moment vector is one-to-one.

4. Our novel graphical representation, which could deal with sequences with different lengths without difficult calculations, has no circuit or degeneracy.

The article is structured as follows: In Sections 2, we describe the construction of novel curves for protein sequences. The construction of phylogenetic trees of several typical protein data and comparison between our results and others' results are described in section 3.

2 Materials and methods

2.1 Graphic representation of protein sequences

2.1.1 The energy of graph

Let $G = [V, E]$ be a finite and undirected graph, with vertex set $V = v_1, v_2, ..., v_n$ and edge set $E = e_1, e_2, ..., e_m$. The adjacency matrix $A = (a_{ij})$ of G is a square matrix of order n, and a_{ij} is defined as:

$$a_{ij} = \begin{cases} 1, \text{ if } (v_i, v_j) \in E \\ 0, \text{ if } (v_i, v_j) \notin E \end{cases} \tag{1}$$

In our paper, we weight the (0,1) matrix with the relationship between amino acids. The eigenvalues of this weighted adjacency matrix are $\lambda_1, \lambda_2, ..., \lambda_n$. The graph energy $\mathbf{E}(G)$ is defined as:

$$\mathbf{E}(G) = \sum_{i=1}^{n} |\lambda_i| \tag{2}$$

2.1.2 Physicochemical properties and graph energies of amino acids

Since the protein sequence is composed of 20 amino acids by different physicochemical properties, each amino acid has the specific properties. Therefore, recognizing the properties of every amino acid is very essential to classify proteins and study its structures and functions. The physicochemical properties of amino acids are found to have strong effects on the pattern of protein evolution [25]. Here, we consider six typical physicochemical properties: relative molecular weight, volume, surface area, specific volume, pKa (-COOH) and pKa ($-NH_3^+$) (six physicochemical properties of 20 amino acids are shown in Supplementary materials Table 1).

For each amino acid, we know the numerical characteristic of its physicochemical

properties. To describe the relationship between two amino acids, we define a threshold $\mathbf{T}_k$ for each property:

$$\mathbf{T}_k = \frac{t_{max}^{(k)} - t_{min}^{(k)}}{19} \qquad k = 1, 2, ..., 6 \tag{3}$$

Where $t_{max}^{(k)}$ and $t_{min}^{(k)}$ indicate maximal value and minimum value of the kth property respectively for 20 amino acids. That is to say, $\mathbf{T}_1$ is used for relative molecular weight, $\mathbf{T}_2$ for volume, $\mathbf{T}_3$ for surface area, $\mathbf{T}_4$ for specific volume, $\mathbf{T}_5$ for pKa (-COOH) and $\mathbf{T}_6$ for pKa (-NH_3^+). Considering the kth property, if the absolute value of difference between amino acid i and amino acid j is less than $\mathbf{T}_k$, we admit the two amino acids are relational. It is in accord with the fact that if amino acid i has relationship(s) with j, then amimo acid j also has relationship(s) with i. The relationship(s) is always symmetric. For example, in relative molecular weight (the numerical values of A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V are 71.08, 156.2, 114.11, 115.09, 103.13, 128.14, 129.12, 57.06, 137.15, 113.17, 113.17, 128.18, 131.21, 147.18, 97.12, 89.08, 101.11, 186.2, 163.18 and 99.14, respectively), $t_{max}^{(1)}$ is 186.20 and $t_{min}^{(2)}$ is 57.06, so the threshold $\mathbf{T}_1 = (186.20 - 57.06)/19 = 6.80$. And then we take amino acid A and W as examples.

The relative molecular weight of amino acids A is 71.08 and absolute difference between A and other amino acids is $|A-R| = 85.12, |A-N| = 43.03, |A-D| = 44.01, |A-C| = 32.05, |A-Q| = 57.06, |A-E| = 58.04, |A-G| = 14.02, |A-H| = 66.07, |A-I| = 42.09, |A-L| = 42.09, |A-K| = 57.1, |A-M| = 60.04, |A-F| = 76.1, |A-P| = 26.04, |A-S| = 18, |A-T| = 30.03, |A-W| = 115.12, |A-Y| = 92.1, |A-V| = 68.92$. Therefore, based on relative molecular weight, amino acid A has no relationship with other amino acids since all absolute differences mentioned above are more than $\mathbf{T}_1$ (=6.8). Similarly, we can obtain following results:

Based on volume, A has relationship with S, no relationship with other 18 amino acids;

Based on surface area, A has relationship with S, no relationship with other 18 amino acids;

Based on specific volume, A has relationship with M, P and W, no relationship with other 16 amino acids;

Based on pKa (-COOH), A has relationship with G and W, no relationship with other 17 amino acids;

Based on pKa (-NH_3^+), A has relationship with D, G and I, no relationship with

other 16 amino acids.

To sum up, A has 1 relationship with D, I, M and P respectively and has 2 relaitonships with G, S and W respectively.

Assuming that 20 amino acids are corresponding to 20 different points in a graph G_i, after considering 6 physicochemical properties, we can get a unique graph for amino acid i. And the rules for drawing the 20-vertex graph for amino acid A is as follows:

rule 1

There will be m edge(s) between A and amino acid X in the 20-vertex graph if A has m relationship(s) with X.

rule 2

Assuming that both Y and Z have relationship(s) with A, there should be n edge(s) between Y and Z if Y has n relationships(s) with Z. On the other hand, if amino acid Y has relationship(s) with A, Z has relationship(s) with Y, but not with A, then the relationship(s) between Y and Z is ignored.

Now we have known that A has 1 relationship with D, I, M and P, respectively; has 2 relationships with G, S and W, respectively. By rule 1, we get a graph for A in Figure 1; adding the rule 2, we get the final 20-vertex graph for amino acid A in Figure 2. In Figure 2, we can find W has 2 relationships with M (both have relation with A), W has 1 relationship with G, and so on. Similarly, we can obtain the 20-vertex graph for W in Figure 3, which shows that W has 1 relationship with G, E and H, respectively and has 2 relationships with M and A, respectively.

Then we can get a weighted adjacency matrix of the amino acid A, which corresponds to the multigraph induced by the vertex A and its first neighbors. According to that, we obtain 20 different weighted adjacency matrices of 20 amino acids (which are shown in Supplementary Materials) and calculate the eigenvalues of weighted adjacency matrix for every amino acid. Finally, their graph energies are computed using Eq.(2). The graph energies of 20 amino acids are shown in Table 1.

Table 1: The graph energies of 20 amino acids

AA	W	M	S	R	K	H	Y	Q	F	N
E(G)	16.01	51.43	37.37	29.36	40.76	27.96	24.06	38.03	31.77	40.05
AA	C	E	D	P	T	I	A	L	G	V
E(G)	17.29	28.26	33.81	20.88	46.74	35.69	18.20	25.84	27.36	53.66

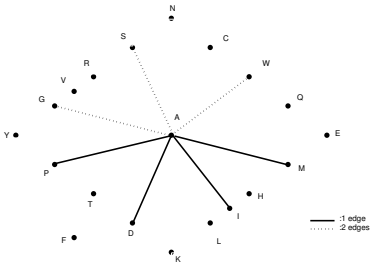

Figure 1: The original graph for amino acid A.

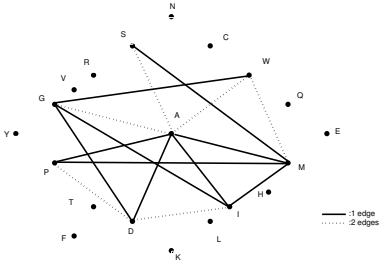

Figure 2: The final 20-vertex graph for amino acid A.

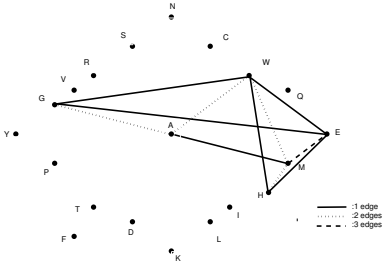

Figure 3: the 20-vertex graph for amino acid W.

2.1.3 The 2D graphical representation of protein sequences basing on graph energy

Proteins consist of twenty kinds of natural amino acids. Since the earliest protein sequences and structures were determined, it has been clear that the positioning and properties of amino acids are key to understand many biological processes. In this paper, the 2D graphical representation of protein sequences is constructed basing on graph energy of each amino acid as follows:

Given a protein sequence with N amino acids $S = s_1, s_2, ..., s_N$, we inspect it by stepping one amino acid at a time. For the step $i(i = 1, 2, ..., N)$, the point $P_i(x_i, y_i)$ can be defined as :

$$\begin{cases} x_i = i \\ y_i = \mathbf{E}(s_i) \end{cases} \tag{4}$$

Where $\mathbf{E}(s_i)$ is the graph energy of amino acid s_i. So we get a graphical curve of the protein sequence.

We take a short segment of a protein of yeast Saccharomyces cerevisiae as an example to show the graphical representation of protein sequences. Our 2-D graphical representation of protein I is illustrated in Figure 4. In Figure 5, we illustrate the 2D graphical representation of ND5 proteins for nine different species.

Protein I: WTFESRNDPAKDPVILWLNGGPGCSSLTGL

2.2 Moment vector

We could get a specific vector from the graphical curve of a protein sequence, which is obtained by the method aforementioned, and use this vector to calculate the distance between two sequences.

2.2.1 Introduction of the moment vector and its application

Given a curve of a protein sequence, we could represent this curve with a series of points, which are actually the coordinates of amino acids, like $(1, y_1), (2, y_2), ..., (n, y_n)$. Then we compute the moment vector by Eq.(5).

$$M_j = \sum_{i=1}^{n} \frac{(x_i - y_i)^j}{n^j} \qquad j = 1, 2, ..., n \tag{5}$$

Where n is the number of amino acids included in a protein sequence, and (x_i, y_i) is the coordinates of ith amino acid of the sequence. Finally, we get a n-dimensional

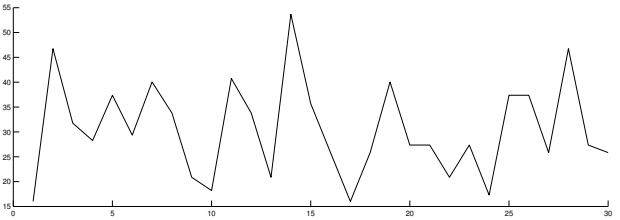

Figure 4: The 2D graphical representation of Protein I.

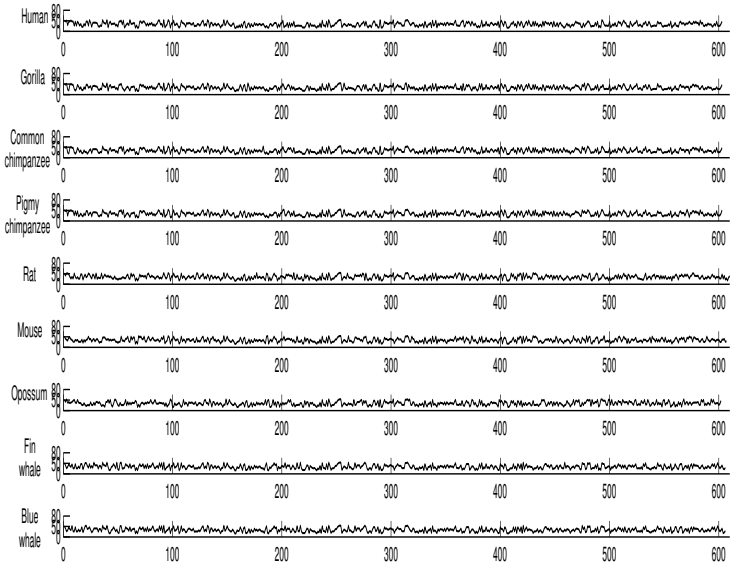

Figure 5: the 2D graphical representation of ND5 proteins .

moment vector $(M_1, M_2, ..., M_n)$ for each protein sequence and this vector is one-to-one with its corresponding sequence [25].

2.2.2 Modification for dimension of moment vector

To avoid complicated computation and problem about different lengths of sequences, a further research is carried out. We translate beta-globin sequences of human, gorilla, cod, cairina moschata, gallus and chelonoidis nigra (their versions in NCBI are AAA16334.1, P02024.2, O13077.2, CAA33756.1, CAA23700.1 and P83123.3 respectively) into 147-dimensional moment vectors and then calculate the Euclidean distances between human and the other five species when the dimensions of moment vectors are changed. The result is showed in Table 2.

It is clearly showed in Table 2 that the distances between beta-globin sequence of

Table 2: The Euclidean distances between human and the other five species

	Gorilla	Cod	Cairina moschata	Gallus	Chelonoidis nigra
2-dim	0.1071	0.9543	0.5953	0.6027	0.4461
3-dim	0.1194	1.2088	0.6033	0.6176	0.591
4-dim	0.124	1.4275	0.6086	0.63	0.7166
5-dim	0.1256	1.5715	0.6143	0.6421	0.8217
10-dim	0.1264	1.8532	0.622	0.6584	1.0419
20-dim	0.1264	1.9119	0.6223	0.659	1.081
25-dim	0.1264	1.9135	0.6223	0.659	1.0815
30-dim	0.1264	1.9138	0.6223	0.659	1.0816
50-dim	0.1264	1.9138	0.6223	0.659	1.0816
100-dim	0.1264	1.9138	0.6223	0.659	1.0816
147-dim	0.1264	1.9138	0.6223	0.659	1.0816

human and the other five species are stable when the dimension increases to 30. At the same time, we also research another two datasets (Cytochrome C sequences of 8 species and ND5) to search the regularity by the same method. The results are shown in Table 2-3 in supplementary materials. From Table 2 in supplementary materials, we can find that the distances of human and the other seven species are also stable when the dimension increases to 30; although human is much far from fish, plant and bacterium in evolution, the differences just appear in 0.0001 or 0.00001, which can be ignored. From Table 3 in supplementary materials, we can find that the distances of human and the other eight species are almost stable when the dimension grows to 30, but the differences appear in 0.1, which is not as perfect as short sequences. After constructing phylogenetic trees of

ND5 using 30-dim moment vectors and 600-dim moment vectors, we discover that these two trees are the same. In summary, 30-dim moment vectors are representative if protein sequences are not too long, and we can find a relatively small number of dimensions for long protein sequences (like thousands of or ten thousands of base pairs) via the same method. Therefore, we extract 30-dimensional moment vectors of protein sequences to replace n-dimensional moment vectors and then calculate distances between sequences to construct phylogenetic trees.

3 The similarities/dissimilarities analysis

To test our novel method, we apply it to the real protein sequence data to analyze the similarities or dissimilarities of different species. Next, we will compare the results of our method with those of ClustalX2 [29-30].

3.1 Outline of the similarities/dissimilarities model

In order to construct a phylogenetic tree for different species, we translate protein sequences into graphical curves and then extract their 30-dimension moment vectors. Finally, we can construct the phylogenetic tree with these moment vectors by proper distance formula and rules.

3.2 The similarities/dissimilarities of nine ND5 protein sequences

The protein sequence similarity can be measured from distance between these multi-dimensional vectors, such as Euclidean distance, Manhattan distance, City Block distance. Here, we take Euclidean distance as the similarity measure between two vectors.

We research the similarity of the nine ND5 proteins using our method. Given two protein sequences P_1 and P_2, 30-dimensional moment vectors of them are $V = (v_1, v_2, ..., v_{30})$ and $W = (w_1, w_2, ..., w_{30})$, and the distance between them is defined as:

$$d(P_1, P_2) = \sqrt{\sum_{i=1}^{30} (v_i - w_i)^2} \tag{6}$$

The smaller is the Euclidean distance d, the more similar are two protein sequences. We calculate the Euclidean distance between two vectors of ND5 proteins from nine different species, as shown in Table 3, and then we can find some results as follows:

1. The distance between Common chimpanzee and Pigmy chimpanzee is the smallest, and it shows that proteins of them are very similar with each other, which is consistent with the biology fact.

2. The values of d between Human, Gorilla, Common chimpanzee and Pigmy chimpanzee are relatively small, which means that they are also similar, and these results are consistent with the evolution relationship.

3. Considering the overall situation, Opossum is the farthest from other eight species, which is accordant to the evolution theory.

ClustalX2 is one of the most popular multiple alignment of sequence programs for DNA or proteins. In order to highlight the effectiveness of our approach, we construct the phylogenetic tree using our approach shown in Figure 6. And we can see that the result of our method is almost consistent with that of ClstalX2.

Table 3: The Euclidean distances matrix for nine ND5 proteins sequences

	Human	Gorilla	C.chim	P.chim	Rat	Mouse	Opossum	F.whale	B.whale
Human	0	1.9533	0.5863	0.7773	3.708	1.9673	6.1867	2.21	2.3105
Gorilla		0	1.8602	2.17	3.1928	3.6482	7.898	1.7569	1.6112
C.chim			0	0.3501	3.956	2.2025	6.0645	2.4786	2.5092
P.chim				0	4.2143	2.1067	5.7371	2.7756	2.8044
Rat					0	3.9999	9.2647	1.6453	1.6353
Mouse						0	5.3188	3.1213	3.2736
Opossum							0	8.1876	8.2828
F.whale								0	0.3601
B.whale									0

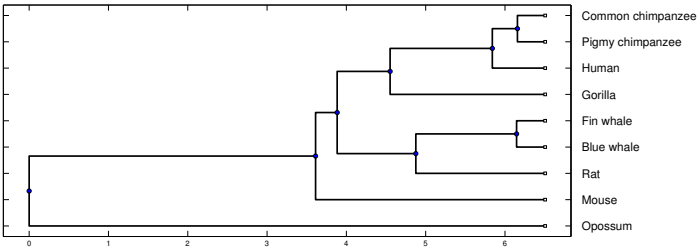

Figure 6: Phylogenetic tree of the nine ND5 proteins constructed by our method.

3.3 The similarities/dissimilarities of 36 protein sequences

We have applied our method to analyzing a database of 36 proteins domains classified into 5 different families (globin, alpha-beta, tim-barrel, all-alpha and all-beta).

Globin: 1eca, 5mbn, 1hlb, 1hlm, 1babA, 1babB, 1ithA, 1mba, 2hbg, 2lhb, 3sdhA, 1ash, 1flp, 1myt, 1lh2, 2vhbA, 2vhb.

Alpha-beta: 1aa9, 1gnp, 6q21A, 1ct9A, 1qraA, 5p21.

Tim-barrel: 6xia, 2mnr, 1chrA, 4enl.

Beta: 1cd8, 1ci5, 1qa9, 1cdb, 1neu, 1qfoA, 1hnf.

Alpha: 1cnp, 1jhg.

We use UPGMA and Eq.(6) to construct a phylogenetic tree via our novel method. The result is shown in Figure 7.

Since 36 sequences have been classified into five families [31-33], it is easy to estimate

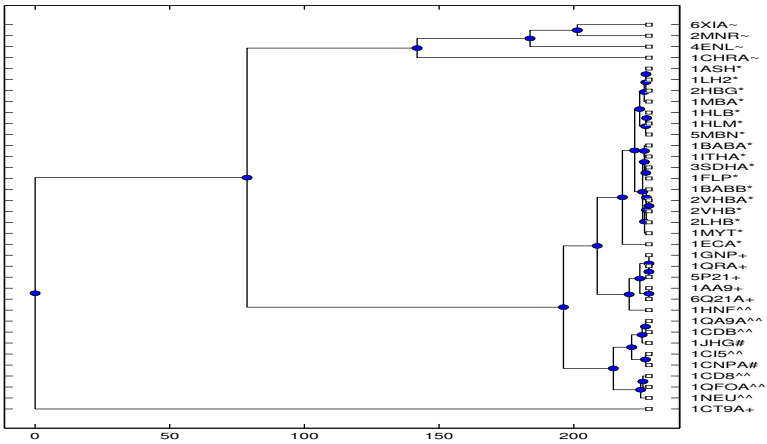

Figure 7: Phylogenetic tree of the 36 proteins constructed by our method.

the result of applying our approach to this dataset. In Figure 7, we can see that only 1CT9A is misplaced. Additionally, family Alpha is not completely separated from Beta. Although there are some shortages, our result is consistent with those references. We also construct the phylogenetic tree of the dataset by ClustalX2 and the result is shown in Figure 8. Compared with ClustalX2, our method is much better than this multiple-alignment method for both short sequences and long sequences.

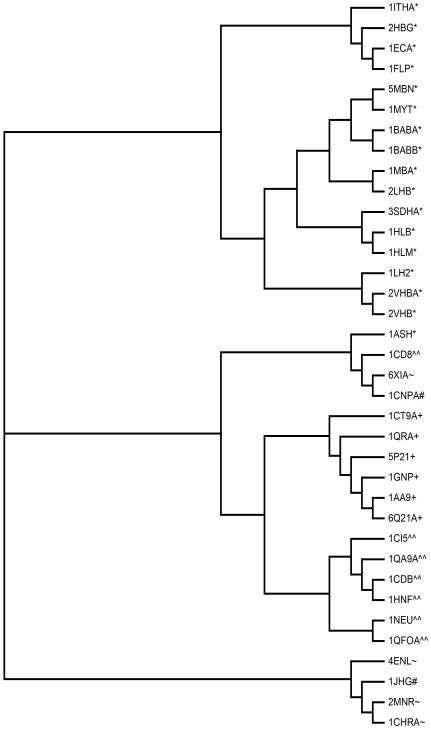

Figure 8: Phylogenetic tree of the 36 proteins constructed by ClustalX2.

4 Conclusion

As an application of graph energy and moment vector, we transform protein sequences into 2-D curve graphs and then analyze their evolutionary relationships. In this paper, the unique weighted adjacency matrix of every amino acid is not only innovative but also meaningful, especially considering the six typical physicochemical properties of amino acids and the nexus among them. As a good graph-representing method which has no circuit or degeneracy, our approach meets all the requirements mentioned in reference [34]. To add to this, our method is valid and efficient in dealing with several typical datasets.

Acknowledgments: The authors wish to thank Prof. Ivan Gutman for his constructive comments and the anonymous referees for their corrections and valuable comments., which were very helpful for improving the presentation of this paper. This work was supported in part by the Shandong Natural Science Foundation (2015ZRE27216).

References

[1] Y. S. Zhang, W. Chen, A new measure for similarity searching in DNA sequences, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 477–488.

[2] Y. S. Zhang, W. Chen, Comparisons of RNA secondary structures based on LZ complexity, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 513–528.

[3] P. P. Qian, Y. S. Zhang, G. Q. Jian, A novel representation of protein sequences and its application, *J. Converg. Inf. Tech.* **6** (2011) 227–235.

[4] Z. C. Wu, X. Xiao, K. C. Chou, 2D–MH: a web–server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids, *J. Theor. Biol.* **267** (2010) 29–34.

[5] Z. C. Mu, J. Wu, Y. S. Zhang, A novel method for similarity/dissimilarity analysis of protein sequences, *Physica A* **392** (2013) 6361–6366.

[6] Y. S. Zhang, X. T. Yu, Analysis of protein sequence similarity, *The Fifth International Conference on Bio–Inspired Computing: Theories and Applications*, 2010, pp. 1255–1258.

[7] V. Brendel, P. Bucher, I. R. Nourbakhsh, B. E. Blaisdell, S. Karlin, Methods and algorithms for statistical analysis of protein sequences, *Proc. Natl. Acad. Sci. USA* **89** (1992) 2002–2006.

[8] D. M. Fowler, C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany, D. Baker ,S. Fields, High–resolution mapping of protein sequence–function relationships, *Nature Meth.* **7** (2010) 741–746.

[9] H. Y. Wu, Y. S. Zhang, W. Chen, Z. C. Mu, Comparative analysis of protein primary sequences with graph energy, *Physica A* **437** (2015) 249–262.

[10] H. Li, N. Homer, A survey of sequence alignment algorithms for next–generation sequencing, *Brief Bioinf.* **11** (2010) 473–483.

[11] A. Chakraborty , S. Bandyopadhyay, FOGSAA: Fast optimal global sequence alignment algorithm, *Sci Rep.* **3** (2013) #1746 (pages 9).

[12] D. F. Feng, R. F. Doolittle, Progressive sequence alignment as a prerequisite to correct phylogenetic trees, *J. Mol. Evol.* **25** (1987) 351–360.

[13] S. Vinga, J. Almeida, Alignment–free sequence comparison – review, *Bioinf.* **19** (2003) 513–523.

[14] T. D. Pham, J. Zuegg, A probabilistic measure for alignment–free sequence comparison, *Bioinf.* **20** (2004) 3455–3461.

[15] G. Reinert, D. Chew, F. Z. Sun, M. S. Waterman, Alignment–free sequence comparison (I): Statistics and power, *J. Comput. Biol.* **16** (2009) 1615–1634.

[16] M. R. Kantorovitz, G. E. Robinson, S. Sinha, A statistical method for alignment–free comparison of regulatory sequences, *Bioinf.* **23** (2007) i249–i255.

[17] I. Borozan, S. Watt, V. Ferretti, Integrating alignment–based and alignment–free sequence similarity measures for biological sequence classification, *Bioinf.* **31** (2015) 1396–1404.

[18] 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.

[19] P. A. He, J. Z. Wei, Y. H. Yao, Z. X. Tie, A novel graphical representation of proteins and its application, *Physica A* **391** (2012) 93–99.

[20] Y. H. Yao, Q. Dai, L. Li, X. Y. Nan, P. A. HE, Y. Z. Zhang, Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation, *J. Comput. Chem.* **31** (2010) 1045–1052.

[21] A. A. Zamyatin, Protein volume in solution, *Prog. Biophys. Mol. Biol.* **24** (1972) 107–123.

[22] C. Chotia, The nature of the accessible and buried surfaces in proteins, *J. Mol. Biol.* **105** (1975) 1–14.

[23] C. Tandford, The interpretation of hydrogen ion titration curves of proteins, *Adv. Protein Chem.* **17** (1962) 70–165.

[24] E. J. Cohn, J. T. Edsall, *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions*, Reinhold Pub. Corp., New York, 1943.

[25] X. H. Xia, W. H. Li, What amino acid properties affect protein evolution? *J. Mol. Evol.* **47** (1998) 557–564.

[26] C. L. Yu, S. Y. Cheng, R. L. He, S. S. T. Yau, Protein map: An alignment–free sequence comparison method based on various properties of amino acids, *Gene* **486** (2011) 110–118.

[27] I. Gutman, The energy of a graph, *Ber. Math. Stat. Sekt.* **103** (1978) 1–22.

[28] S. S. T. Yau, C. L. Yu ,R. He, A protein map and its application, *DNA. Cell. Biol.* **27** (2008) 241–250.

[29] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson ,D. G. Higgins, Clustal W and Clustal X version 2.0, *Bioinf.* **23** (2007) 2947–2948.

[30] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, J. D. Thompson, Multiple sequence alignment with the Clustal series of programs, *Nucleic Acids Res.* **31** (2003) 3497–3500.

[31] Z. C. Mu, J. Wu , Y. S. Zhang, A novel method for similarity/dissimilarity analysis of protein sequences, *Physica A* **392** (2013) 6361–6366.

[32] S. L. Zhang , L. P. Yang , T. M. Wang, Use of information discrepancy measure to compare protein secondary structures, *J. Mol. Struct. (Theochem)* **909** (2009) 102–106.

[33] Y. Wang, L. Y. Wu, J. H. Zhang, Z. W. Zhan, X. S. Zhang, L. Chen, Evaluating protein similarity from coarse structures, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **6** (2009) 583–593.

[34] M. Randić, J. Zupan, A. T. Balaban, D. Vikić–Topić, D. Plavšić, Graphical representation of proteins, *Chem. Rev.* **111** (2011) 790–862.

[35] I. Gutman, D. Vidović, N. Cmiljanović, S. Milosavljević, S. Radenković, Graph energy – A useful molecular structure–descriptor, *Indian J. Chem.* **42A** (2003) 1309–1311.

[36] I. Gutman, J. Y. Shao, The energy change of weighted graphs, *Lin. Algebra Appl.* **435** (2011) 2425–2431.

[37] I. Gutman, Laplacian energy of a graph, *Lin. Algebra Appl.* **414** (2006) 29–37.

[38] B. Zhou, I. Gutman, On Laplacian energy of graphs, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 211–220.