MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **73** (2015) 749-764

ISSN 0340 - 6253

F-Curve, a Graphical Representation of Protein Sequences for Similarity Analysis Based on Physicochemical Properties of Amino Acids

Hailong Hu[1,2]

[1]*College of Science, Zhejiang A&F University, Hangzhou, China, 311300*
[2]*Key Laboratory of Chemical Utilization of Forestry Biomass of Zhejiang Province, Hangzhou, China, 311300*
huhailonga@163.com

(Received December 8, 2014)

Abstract

We introduce a new graphical representation of protein sequences based on two selected physicochemical properties of amino acids. Using the two physicochemical properties of amino acids, we firstly obtain a 2D discrete point set for amino acids of a protein sequence. Then we use fractal interpolation method to interpolate these discrete points to represent protein sequences. Next we employ the fractal dimension of protein curves images to analyze the similarity of protein sequences by comparing the distance of vectors. The ND5 protein sequence of nine species is an example to demonstrate our method. We finally use a linear correlation and significance analysis to compare our results and some other graphical representation results with the ClustalW results. The compared results show that our proposed method is effective for the similarity analysis of proteins.

1. Introduction

Protein sequence is a string of letters, which correspond to amino acid residues of the polypeptide chain from the *N*-terminus to *C*-terminus. Protein sequence analysis has become a routine task in biological information and other related fields [1-5]. Since it is difficult to directly extract the useful information from protein sequences by experimentation, many researchers choose to consider the graphical representation and numerical description to analyze the protein sequences and infer biological information from protein sequences [6,7].

Due to the difficulty in obtaining information and comparing different sequences by letter strings of proteins, many methods were proposed to translate these letter strings of proteins sequences into mathematical forms and after that analyzed the similarity. The graphical representation of protein is a way of researching protein sequences. Many researchers constructed a graphical representation of protein sequences based on triplet codons of DNA. For example, Randić [8] outlined a new graphical representation of proteins, which is highly compacted 2D representation of a DNA or RNA sequence and then selected a unique virtual genetic code to represent protein sequences. Bai et al. [9] proposed a graphical representation of protein sequences based on nucleotide triplet codons of DNA and then transform a DNA triplet character string into complex numerical sequence. He et al. [10] defined a transformation which can transform each triplet codon into a 6-bit binary and then constructed a graphical representation of protein sequences. Maaty et al. [11] outlined a new graphical representation of protein sequence method based on physicochemical properties of amino acids side chains. Maaty et al. [12] represented protein sequences by placing twenty amino acids on the surface area of a unit sphere which is divided into twenty latitude-like circles and n longitude-like semi-circles for representing all residues. Abo-Elkhier [13] presented a method of graphical representation of protein sequences by mapping twenty amino acids to twenty circles and proteins residues to n lines on the surface of a right cone which is unit base and unit height. In addition, Wen et al. [14], Randić [15], Wu et al. [16] and Yao et al. [17] proposed a graphical representation of proteins by mapping two or three selected physicochemical properties of amino acids to 2D or 3D Cartesian coordinate bases which corresponding to twenty amino acids, in order to reflect the physicochemical properties of proteins fully. However, aforementioned methods generated zigzag curves which geometric property cannot be fully applied for the graphical representation of protein sequences. Liu et al. [18] represented protein sequences and analyzed the similarity by using P-H curve and Deng et al. [19] by dual-vector curve (DV-curve), Qi et al. [20] proposed a graphical representation of protein sequences based on the Huffman tree method. Li et al. [21] represented protein sequences by cubic Bezier spline interpolating discrete points and analyzed the similarity by curvature frequency vectors.

In this paper, we propose a novel graphical representation for protein sequences. We firstly construct a 2D space discrete point set for amino acids of each protein sequence based on two physicochemical properties: $pK_a(COOH)$ and $pK_a(NH_3^+)$. Then, we use fractal method to interpolate these discrete points to obtain the protein sequence curve: F-curve. We finally analyze the similarity of proteins sequences by their curve image's fractal dimension. We compared between our results and other authors' results with the ClustalW results through the correlation and significance analysis. The results show that our method is applicable.

2. 2D graphical representation of protein primary sequences

Proteins consist of twenty amino acids; each amino acid can be represented by one letter of twenty different letters:A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y. Listed in table 1. In this study, based on two physicochemical properties' indexes of twenty amino acids, $pK_a(COOH)$ and $pK_a(NH_3^+)$, the graphical representation of proteins is constructed as follows.

We firstly construct the 2D Cartesian coordinates of each amino acid according to their values of $pK_a(COOH)$ and $pK_a(NH_3^+)$. The values of these two parameters are listed in Table 1. If the values are directly considered as coordinates of the points representing the twenty amino acids in a Cartesian (x, y) coordinate system, then those points will all lie in the first quadrant due to all values of $pK_a(COOH)$ and $pK_a(NH_3^+)$ are positive, which is not convenient for analyzing and comparing different structures and functions of proteins. So we obtain the X, Y coordinates by the transformation (1)

$$\begin{cases} x_i = pK_a(COOH)_i - \overline{pK_a(COOH)} \\ y_i = pK_a(NH_3^+)_i - \overline{pK_a(NH_3^+)} \end{cases} \quad (1)$$

Where $\overline{pK_\alpha(COOH)}$ and $\overline{pK_\alpha(NH_3^+)}$ are the averages of matching $pK_a(COOH)$ and $pK_a(NH_3^+)$ values of all amino acids. Using the transformation (1), we obtain the new coordinates matching the points which are also listed in the last two columns in Table 1.

After mapping the protein sequence's amino acids to these characterized vectors,our graphical representation is obtained by summing these vectors. For any given protein sequence $S=S_1S_2...S_N$, we have N points in the graph. Each point P_i has two components (X_i,

Y_i), which are obtained as follows.

$$\begin{cases} X_i = \sum_{k=1}^{i} S_k{}^1, \\ Y_i = \sum_{k=1}^{i} S_k{}^2. \end{cases}$$

Where S_k^j $(j = 1, 2)$ represents the jth component of the vector corresponding to S_k.

Table 1. Two parameters of the twenty amino acids and their coordinates in 2D Cartesian coordinates

Amino acid	Abbreviation	Symbol	$pK_a(COOH)$	$pK_a(NH_3^+)$	x	y
Alanine	Ala	A	2.35	9.87	0.163	0.389
Cysteine	Cys	C	1.71	10.78	-0.477	1.299
Aspartic acid	Asp	D	1.88	9.60	-0.307	0.119
Glutamic acid	Glu	E	2.19	9.67	0.003	0.189
Phenylalanine	Phe	F	2.58	9.24	0.393	-0.241
Glycine	Gly	G	2.34	9.60	0.153	0.119
Histidine	His	H	1.78	8.97	-0.407	-0.511
Isoleucine	Ile	I	2.32	9.76	0.133	0.279
Lysine	Lys	K	2.20	8.90	0.013	-0.581
Leucine	Leu	L	2.36	9.60	0.173	0.119
Methionine	Met	M	2.28	9.21	0.093	-0.271
Asparagine	Asn	N	2.18	9.09	-0.007	-0.391
Proline	Pro	P	1.99	10.60	-0.197	1.119
Glutamine	Gln	Q	2.17	9.13	-0.091	-0.351
Arginine	Arg	R	2.18	9.09	-0.007	-0.391
Serine	Ser	S	2.21	9.15	0.023	-0.331
Threonine	Thr	T	2.15	9.12	-0.037	-0.361
Valine	Val	V	2.29	9.74	0.103	0.259
Tryptophan	Trp	W	2.38	9.39	0.193	-0.091
Tyrosine	Tyr	Y	2.20	9.11	0.013	-0.371

Then we interpolate these 2D discrete points to gain a curve for the graphical representation of protein. As we know, a traditional mathematical interpolation function or a curve fitting function is a linear combination of a set of basis functions, which are typically one of the polynomial, rational and triangle functions. Fractal objects have several important features: fine structures (i.e. detailed in arbitrarily small scales); self-similarity (perhaps exact, approximate or statistical); too irregular to be described by a traditional geometrical method (both locally and globally); construction by some simple way (perhaps recursively). So fractal

geometry of fractal function theory has developed beyond its mathematical framework and grow up to be a useful tool for the shape analysis in sciences and engineering applications [22,23]. In this study, we utilize the fractal interpolating function to construct the curve of a protein sequence by Iterated Function System (IFS).

For a set of given points (for instance, 2D discrete points corresponding to amino acids in a protein sequence), we interpolate these points by fractal method. So we have to call the interpolating curve as F-curve of the protein. The constructing detailed description is in [24,25]. Considering IFS $\{R^2, w_i: 1,2,...,N\}$, where w_i is an affine transformation:

$$w_i\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_i & 0 \\ c_i & d_i \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix}, i = 1, 2, \cdots, N$$

Constrained to satisfy

$$w_i\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} x_{i-1} \\ y_{i-1} \end{pmatrix} \text{ and } w_i\begin{pmatrix} x_N \\ y_N \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix}, i = 1, 2, \cdots, N$$

Assuming that the vertical scaling coefficient d_i is a free parameter, we can obtain a unique solution:

$$a_i = \frac{x_i - x_{i-1}}{x_N - x_0}$$

$$e_i = \frac{x_N x_{i-1} - x_0 x_i}{x_N - x_0}$$

$$c_i = \frac{y_i - y_{i-1}}{x_N - x_0} - d_i \frac{y_N - y_0}{x_N - x_0}$$

$$f_i = \frac{x_N y_{i-1} - x_0 y_i}{x_N - x_0} - d_i \frac{x_N y_0 - x_0 y_N}{x_N - x_0}$$

We apply our method on two shorter segments of the protein of the yeast Saccharomyces cerevisiae, called as protein I and protein II.

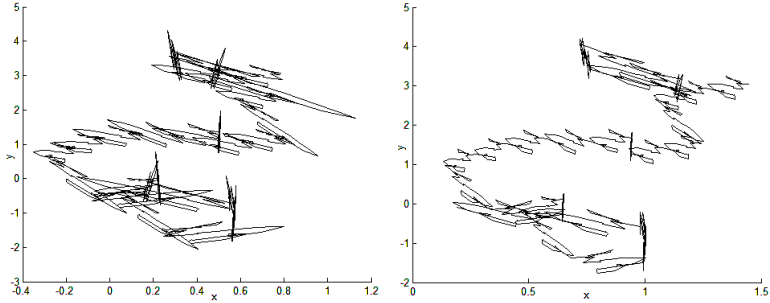
Protein I: W T F E S R N D P A K D P V I L W L N G G P G C S S L T G L

Protein II: W F F E S R N D P A N D P I I L W L N G G P G C S S F T G L

The corresponding 2D space curves of protein I and II are presented in Fig. 1(a) and (b). By looking at Fig. 1(a) and (b), we can have concluded that the two F-curves are similar all in

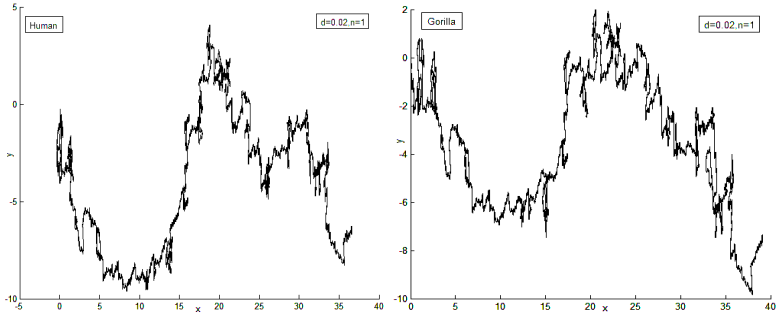
all shape, and easily observe the difference between them.

In addition, graphical representations of nine different species of ND5 are constructed and illustrated in Fig. 2. These show that our F-curve method can provide a good visualization for analyzing protein sequences. Comparing these curves, we can see that the curves of the proteins of Human, Gorilla, P. Chimp and C. Chimp are more similar with each other although Gorilla is a little different, and F. Whale and B. Whale, Rat and Mouse are also more similar, respectively. In addition, we find that the curve of Opossum is distinct from others. The results are in agreement with the known fact of evolution [26-29].

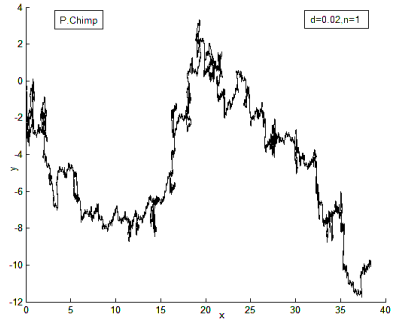


(a) protein I (b) proteinII

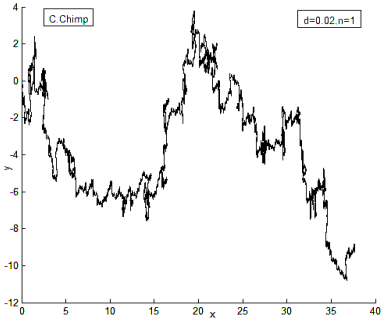
Fig. 1. The F-curves of protein I and II
(the vertical scaling coefficient d_i=0.1, iterated number n=1)



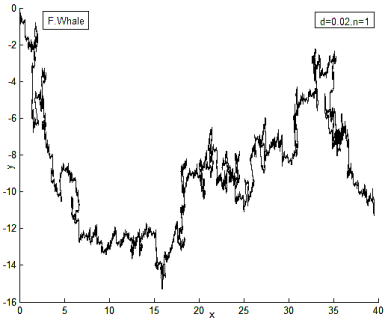
Human Gorilla



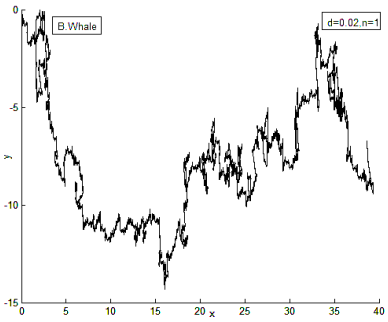
P. Chimp



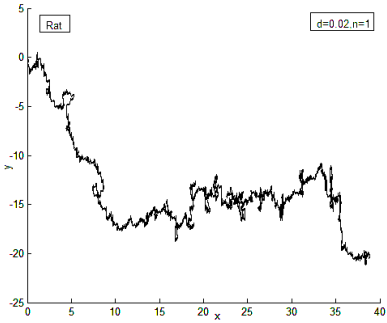
C. Chimp



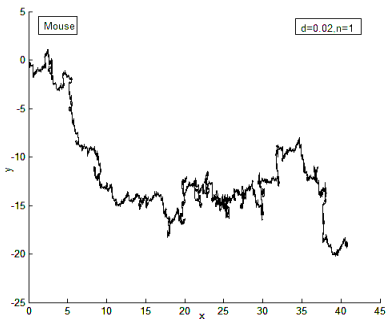
F. Whale



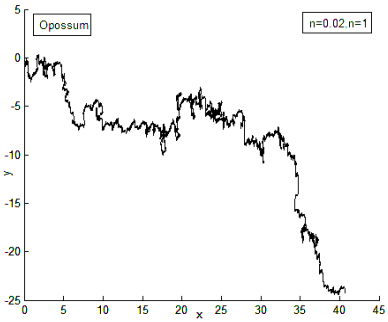
B. Whale



Rat



Mouse



Opossum

Fig. 2. The F-curves of nine proteins sequences of ND5
(the vertical scaling coefficient d_i=0.02, iterated number n=1)

3. Numerical characterization and similarity analysis of proteins by 2D fractal image

We transform the graphical representation of the characteristic curve into another mathematical object in order to facilitate analysis of proteins. As we know, the fractal dimension can describe the fractal interpolation curve. However, it is related to the vertical scaling coefficient d_i but not to the interpolating points (x_i, y_i). If we directly use the fractal dimensions to describe the curves corresponding to the protein sequences, all the dimensions will be equal when all d_i is same. So we firstly transform fractal curves corresponding to the protein sequences into JPEG images, and then we use the box-counting algorithm to calculate the image fractal dimensions [30,31]. The box-counting dimension formula is

$$D = \frac{\log N(\varepsilon)}{\log(1/\varepsilon)},$$

where $N(\varepsilon)$ is the number of boxes with the size ε which can fill the entire area of an image. Changing the size of ε means that the number of boxes $N(\varepsilon)$ also changes. These changes cause D to change. For example, the protein I and protein II image fractal dimensions are 2.0317 and 2.0318, respectively, when iterated number n=1, vertical scaling coefficient d_i=0.1; the dimensions are 2.0308 and 2.0318,respectively, when n=1, d_i=0.2.

We construct a set of F-curves of each protein sequence according to different d_i in order

to describe the protein sequence precisely. So we calculate a set of fractal dimensions from these F-curves images. And then calculate the average of these dimensions, denoted as $\overline{D}$.

We construct a two-dimensional vector (u_x, u_y) to describe the protein sequence, which $u_x = \overline{D}, u_y = length\ of\ sequence$. So we call this vector as a dimension-length vector. The similarity analysis can be measured by calculating the distance between these two-dimensional vectors. Here, we utilize the Euclidean distance as the similarity measure between two vectors. Given two vectors

$$X = (X_1, X_2, \cdots, X_N),$$

$$Y = (Y_1, Y_2, \cdots, Y_N),$$

the Euclidean distance between the two vectors is

$$d(X, Y) = \left[\sum_{i=1}^{N} (X_i - Y_i)^2 \right]^{\frac{1}{2}}.$$

We compare the similarities among nine ND5 proteins sequences (Table 2) to illustrate this measurement method. We calculate the protein sequences image fractal dimensions are Table 3 corresponding to d_i=0.01, 0.02, ..., 0.05. The dimension-length vectors belonging to each species are given in Table 4.

Depending on Table 4, we generate a protein map of nine different species in Fig.3. Observing Fig.3, we can find that the species of Human, Gorilla, P. Chimp and C. Chimp are more similar with each other, and F. Whale and B. Whale, Rat and Mouse are also similar with each other, respectively.

Then, we calculate the Euclidean distance between two vectors of ND5 proteins from nine different species, listed in Table 5. The smaller the distance is, the more similar the two proteins are. From Table 5, we can find that:

The distances among Human, Gorilla, C. Chimap and P. Chimap are all quite small, that is to say, they are more similar with each other; of course, the distance between F. Whale and B. Whale is small, so they are more similar with each other; the previously described methods [11,17] results imply (Mouse, B.Whale),(Opossum, Mouse) are very similar, but this is not inconsistent with the known fact of evolution[26-29] and our method can distinguish them

efficiently.

Table 2. The ND5 proteins of nine different species

Species	Sequence code（NCBI）	Sequence length
Human	CAA24036	603
Gorilla	BAA07306	603
P.Chimp	BAA07315	603
C.Chimp	BAA07302	603
F.Whale	CAA43449	606
B.Whale	CAA51005	606
Rat	CAA32964	610
Mouse	CAA24088	607
Opossum	CAA82687	602

Table 3. Five dimensional vectors based on image fractal dimension

Human	Gorilla	P.Chimp	C.Chimp	F.Whale	B.Whale	Rat	Mouse	Opossum
2.0318	2.0317	2.0319	2.0319	2.0319	2.0318	2.0322	2.0321	2.0322
2.0314	2.0313	2.0316	2.0316	2.0315	2.0315	2.032	2.032	2.0321
2.0309	2.0312	2.0312	2.0312	2.0309	2.0313	2.0318	2.0318	2.0314
2.0303	2.0307	2.0307	2.0308	2.0303	2.0308	2.0316	2.0316	2.0317
2.0297	2.0302	2.0302	2.0306	2.0297	2.0303	2.0313	2.0313	2.0315

Table 4. Dimension-length vectors (u_x,u_y) corresponding to the nine species

	Human	Gorilla	P.Chimp	C.Chimp	F.Whale	B.Whale	Rat	Mouse	Opossum
u_x	2.03082	2.03102	2.03112	2.03122	2.03086	2.03114	2.03178	2.03176	2.03178
u_y	603	603	603	603	606	606	610	607	602

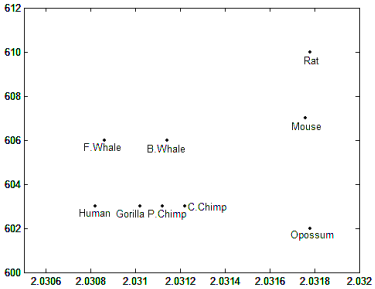


Fig. 3. The protein map of nine different species

Table 5. The distances of dimension-length vectors between the nine different species

	Human	Gorilla	P.Chimp	C.Chimp	F.Whale	B.Whale	Rat	Mouse	Opossum
Human	0	0.0002	0.0003	0.0004	3.0000	3.0000	7.0000	4.0000	1.0000
Gorilla		0	0.0001	0.0002	3.0000	3.0000	7.0000	4.0000	1.0000
P.Chimp			0	0.0001	3.0000	3.0000	7.0000	4.0000	1.0000
C.Chimp				0	3.0000	3.0000	7.0000	4.0000	1.0000
F.Whale					0	0.0003	4.0000	1.0000	4.0000
B.Whale						0	4.0000	1.0000	4.0000
Rat							0	3.0000	8.0000
Mouse								0	5.0000
Opossum									0

4. Our method compared with other methods

As we know that the ClustalW program is one of the most popular multiple sequence alignment approaches for DNA or proteins [32,33], so we run the ClustalW program, and the results are included as a distance matrix in Table 6. To compare our method with ClustalW, the correlation coefficients and significance analysis is provided.

The correlation coefficient r between our method and ClustalW method can be calculated by corresponding rows of Table 5 and Table 6. The correlation coefficient of the first row in both matrices is 0.7819, which belongs to Human protein, the second to Gorilla and so on. In the first column of Table 7, the correlation coefficients for the rows belonging to all nine species are enumerated. And we similarly calculate the correlation coefficients between the methods of Refs. [11,14,17] and ClustalW method. These correlation coefficients are also included in other columns of Table 7. According to the statistics knowledge, if the correlation coefficient r of two variables U and V satisfies $r_{0.05}(n\text{-}2) < |r| \leq r_{0.01}(n\text{-}2)$, here, n=9, that is $0.666 < |r| \leq 0.798$, then we say that U and V are linear correlation; if $|r| > r_{0.01}(n\text{-}2)$, that is $|r| > 0.798$, U and V are said to be strong linear correlation. From Table 7, we find that the linear correlation and strong linear correlation groups between our method and other methods (Refs. [11,14,17]) referring to the ClustalW method are 7, 3, 5, 6, 7, respectively. We have concluded that our method has more linear correlation groups with the ClustalW than other methods.

Table 6. The similarity distance for nine different species of ND5 proteins calculated by the ClustalW.

	Human	Gorilla	P.Chimp	C.Chimp	F.Whale	B.Whale	Rat	Mouse	Opossum
Human	0	10.7	7.1	6.9	41.0	41.3	50.2	48.9	50.4
Gorilla		0	9.7	9.9	42.7	42.4	51.4	49.9	54.0
P.Chimp			0	5.1	40.1	40.1	50.2	48.9	50.1
C.Chimp				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	25.9	54.0
Mouse								0	50.8
Opossum									0

Table 7. The correlation coefficients for nine ND5 proteins of our method and the methods in Refs [11,14,17], as compared with ClustalW method.

	Our method Table 7	Ref. [14] with Table 4	Ref. [11] with Table 3	Ref. [11] with Table 4	Ref. [17] with Table 4
Human	0.7819	0.6090	0.9419	0.7470	0.9349
Gorilla	0.7630	0.8278	0.9363	0.7902	0.9399
P.Chimp	0.7856	0.8332	0.8755	0.7887	0.9605
C.Chimp	0.7808	0.8148	0.9448	0.8110	0.9679
F.Whale	0.8360	0.2729	0.8146	0.6310	0.8884
B.Whale	0.8430	0.5964	0.6593	0.6523	0.8489
Rat	0.9213	0.5777	0.6479	0.8970	0.7003
Mouse	0.6391	0.6027	0.6308	0.6229	0.6558
Opossum	0.4299	0.4899	0.4772	0.8109	0.5244

Because we have a small sample data (n=9) which can produce high correlation coefficients, we make the significance analysis for the correlation coefficients to check whether the correlation of two sets of data is sufficiently strong or likely occurred by chance. We compute the statistical significance for correlation coefficient values that are greater than 0.7 through t-test.

Our sample size is 9, so the degree of freedom is 7. A t-value of greater than 2.365 indicates that significance of less than 0.05 chance of having occurred by coincidence. In Table 8, we list the t-values corresponding to the r-values which are higher than 0.7. All computed t-values are higher than 2.365. This states that the r-values in Table 7 are not

occurred by chance. There are 7 *t*-values of greater than 2.365 in our method, while there are only 3, 5, 6, 7 *t*-values in additional methods (Refs. [11, 14, 17]), respectively.

Table 8. The *t*-values computed for the correlation coefficients $|r|>0.7$, based on them the significance is determined.

	Our method Table 7	Ref. [14] with Table 4	Ref. [11] with Table 3	Ref. [11] with Table 4	Ref. [17] with Table 4
Human	3.3181	-	7.4169	2.9731	6.9689
Gorilla	3.1229	3.9035	7.0524	3.4116	7.2800
P.Chimp	3.3588	3.9871	4.7944	3.3944	9.1294
C.Chimp	3.3069	3.7181	7.6311	3.6679	10.1859
F.Whale	4.0314	-	3.7160	-	5.1202
B.Whale	4.1463	-	-	-	4.2490
Rat	6.2663	-	-	5.3700	2.5955
Mouse	-	-	-	-	-
Opossum	-	-	-	3.6664	-

5. Conclusion and future work

On the basis of physicochemical properties of amino acids, we propose a novel method for graphical representation of protein sequences by fractal interpolating method, called F-curve, which provide good insight for protein sequences visualization. And then we use image fractal dimension to analyze the similarity of protein sequences. This method is more simpler, convenient, and fast. Furthermore, the method will be expanded to multifractal and combine other physicochemical properties of amino acids to study the protein structures and functions.

Acknowledgment: This work is supported by the Science Research Project of the Educational Department of Zhejiang Province of China under Grant No. Y201329332 and FX2013075.

References

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

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

[3] B. Liao, B. Y. Liao, X. G. Lu, Z. Cao, A novel graphical representation of protein sequences and its application, *J. Comput. Chem.* **32** (2011) 2539 – 2544.

[4] M. Randić, M. Novič, D. Plavšić, Milestones in graphical bioinformatics, *Int. J. Quantum Chem*. **113** (2013) 2413–2448.

[5] M. Randić, Very efficient search for protein alignment – VESPA, *J. Comput. Chem*. **33** (2012) 702–707.

[6] P. Echenique, Introduction to protein folding for physicists, *Contemp. Phys.* **48** (2007) 81 – 108.

[7] J. Song, H. W. Tang, A new 2-D graphical representation of DNA sequences and their numerical characterization, *J. Biol. Methods* **63** (2005) 228–239.

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

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

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

[11] M. I. Abo el Maaty, M. M. Abo–Elkhier, M. A. Abd Elwahaab, 3D graphical representation of protein sequences and their statistical characterization, *Physica A* **389** (2010) 4668–4676.

[12] M. I. Abo el Maaty, M. M. Abo–Elkhier, M. A. Abd Elwahaab, Representation of protein sequences on latitude-like circles and longitude–like semi–circles, *Chem. Phys. Lett.* **493** (2010) 386–391.

[13] M. M. Abo–Elkhier, Similarity/dissimilarity analysis of protein sequences using the spatial median as a descriptor, *J. Biophys. Chem.* **3** (2012) 142–148.

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

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

[16] 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.

[17] 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.

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

[19] W. Deng, Y. H. Luan, DV-Curve representation of protein sequences and its application, *Comput. Math. Methods Med.* **2014** (2014) #203871 (pp. 1-8).

[20] Z. H. Qi, J. Feng, X. Q. Qi, L. Li, Application of 2D graphic representation of protein sequence based on Huffman tree method, *Comput. Biol. Med.* **42** (2012) 556–563.

[21] Z. Li, C. C. Geng, P. A. He, Y. H. Yao, A novel method of 3D graphical representation and similarities analysis for protein sequences, *MATCH Commun. Math. Comput. Chem.* **71** (2014) 213–226.

[22] P. R. Massopust, Fractal functions and their applications, *Chaos Solitons Fractals* **2** (1997) 171–190.

[23] M. A. Vyzantiadou, A. V. Avdelas, S. Zafiropoulos, The application of fractal geometry to the design of grid or reticulated shell structures, *Computer Aided Des.* **39** (2007) 51–59.

[24] P. Bouboulis, Modelling discrete sequences with fractal interpolation functions of higher order, *J. Math. Sci.* **3** (2010) 1–10.

[25] D. S. Mazel, M. H. Hayes, Using iterated function systems to model discrete sequences, *IEEE Trans. Signal Process.* **40** (1992) 1724–1734.

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

[27] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, H. Zhang, An information–based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics* **17** (2001) 149–154.

[28] H. H. Otu, K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinformatics* **19** (2003) 2122–2130.

[29] V. Makarenkov, F. Lapointe, A weighted least–squares approach for inferring phylogenies from incomplete distance matrices, *Bioinformatics* **20** (2004) 2113–2121.

[30] N. Sarkar, B. B. Chaudhuri, An efficient approach to estimate fractal dimension of textural images, *Pattern Recog.* **25** (1992) 1035–1041.

[31] O. M. Bruno, R. O. Plotze, M. Falvo, M. Castro, Fractal dimension applied to plant identification, *Inf. Sci.* **178** (2008) 2722–2733.

[32] J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.* **22** (1994) 4673–4680.

[33] C. Z. Guo, M. Q. Sun, ClustalW–A software for multiple sequence alignment of protein and nucleic acid sequence, *Biotech. Lett*. **11 (**2000) 146–149.