MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **71** (2014) 213-226

ISSN 0340 - 6253

A Novel Method of 3D Graphical Representation and Similarity Analysis for Proteins

Zhong Li[1], Changchun Geng, Pingan He, Yuhua Yao

College of Science, Zhejiang Sci-Tech University, Hangzhou, China, 310018

(Received January 29, 2013)

Abstract

A new graphical representation for protein sequences is introduced in this paper. We firstly construct a 3D space discrete point set for amino acids of protein sequences based on three physicochemical properties of amino acids. Then, we use a cubic Bezier spline curve to interpolate these discrete points to represent protein sequences. Different from many traditional and current graphical representations of protein sequences, our curve is smooth, continuous and parametric, we can easily apply the geometric property of the curve for the similarity analysis. We finally use the curvature frequencies of protein curves to analyze the similarity of protein sequences by comparing the distance of vectors. As an example, we take ND5 proteins from nine different species to illustrate our method. The experimental results show that our proposed method is effective for the similarity analysis of proteins.

1. Introduction

With the rapid development of sequencing techniques, the number of biological sequences is increasing rapidly in various kinds of biological databases. Information extraction, comparative analysis and discovery from DNA, RNA and protein sequences are one of the very important tasks in molecular biology and bioinformatics. Since it is more difficult to get useful information directly from the primary sequences, many researchers transform the original sequences into some mathematical forms and then make the similarity analysis. As we know, the graphical representation of biological sequences is a valid method for the mathematical description and numerical analysis. Therefore, it has become a hot research topic in biological information and other related fields [1-3].

[1] Corresponding author. E-mail: lizhong@zstu.edu.cn.

The problem of graphical representation for protein sequences is generalized from graphical representation of DNA. It is more complicated because of the substitution from four bases to twenty amino acids [4-6]. Recently, many researchers have put forward various methods of 2D and 3D graphical representation for protein sequences [7-16]. For example, Randic [10] and Liao [17] *et al* propose some methods of graphical representation for proteins based on the genetic code. Randic [11], He [4], Yao [5], Feng [19] *et al* provide the different methods of graphical representation according to the physicochemical properties of 20 amino acids. Some researchers also make improvements on the existing methods of graphical representation for DNA and then propose some new methods. For example, Randic [12] proposes a 2D method of graphical representation for protein sequences based on the graphical representation of DNA, which is introduced by Jerry [20]. He uniformly puts 20 amino acids to a unit circle and presents the graphical representation based on the CGR method according to the dictionary sort of the amino acids' three letters logogram form. Furthermore, Bai [21] presents a method of 3D graphical representation of protein sequences by mapping the 20 amino acids to the 20 vertexes of the regular dodecahedron. Randic *et al* [22] and He *et al* [23] place the 20 amino acids on the periphery of the unit circle to replace a square with a 20-side polygon. However, above-mentioned 3D methods only make a protein sequence become a set of polylines formed from three-dimensional discrete points (amino acids). Geometric property of such zigzag curves cannot be fully applied for the graphical representation of protein sequences.

Based on above analysis, we propose a novel graphical representation for protein sequences and apply this graphical curve's geometric property for the similarity of proteins. The main contributions of this paper are as follows:

(1). We fully consider the hydrophobic property of amino acids and pK_a values of amino acids to determine the space discrete point position of amino acids in protein sequences.

(2). We use the cubic Bezier spline curve to interpolate these amino acid points of protein sequences. Compared to traditional zigzag protein curve, the new graphical curve is smooth, continuous and parametric, we can apply the curve's geometric property to compare similarities of proteins.

(3). We use the curvature property of protein curves to define ten dimensional frequency vector and make the similarity analysis for different protein sequences by the vector distance.

2. 3D space curve representation of protein sequences

From the view of biology information, a protein sequence can be regarded as a character string in the alphabet of 20 amino acids Ω, where

$$\Omega = \{\text{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}\}$$

In order to obtain the graphical representation of protein sequences, we firstly need to determine three dimensional coordinates of each element in Ω, then each protein sequence can be transformed into the unique graphical description. In bioinformatics, the physicochemical properties of amino acids are the main basis form folding the space structure and performing their function. Many researchers use the equipotential point and pK_a values of the amino acids to determine corresponding 3D space coordinates, and then orderly connect them to form a polyline for the graphical representation of protein sequences.

As we know, hydrophobicity, denoted as *Hy*, is the physical property of the mutual repulsion between a molecular and the water. It is also an important property of amino acids which reflects the amino acid' overall function and structure. The pK_a value of amino acids reflects the relative ease degree that the amino acids release dissociable proton. Two functional groups *a-COOH* and *a-*NH_3^+ are weak acid groups and can release protons in aqueous solution. The ability to release and accept proton, denoted as $pK_a(COOH)$ and $pK_a(NH_3^+)$, is the basic chemical property of the proteins and often involves in the composition of protein structures and the catalysis of enzyme. Based on these analysis, we use *Hy*, $pK_a(COOH)$ and $pK_a(NH_3^+)$ of each amino acid for graphical description of protein sequences.

Table 1 gives the values of *Hy*, $pK_a(COOH)$ and $pK_a(NH_3^+)$ of 20 amino acids. However, all values of $pK_a(COOH)$ and $pK_a(NH_3^+)$ are positive. If these two values are directly corresponded to the *X* coordinate and *Y* coordinate in the 3D space, 20 amino acids are all located in the first octant of the 3D space, which is not convenient for analyzing and comparing different shapes. Here, we also apply the transformation method from [23] for solving this problem. Namely, we first compute the average of *Hy*, $pK_a(COOH)$ and $pK_a(NH_3^+)$ for all 20 amino acids and then construct the coordinate for each amino acid by

$$\begin{cases} x_m = pK_a(COOH)_m - averge(pK_a(COOH)) \\ y_m = pK_a(NH_3^+)_m - averge(pK_a(NH_3^+)) \\ z_m = Hy_m - averge(Hy) \end{cases}$$

where $averge(pK_a(COOH))$, $averge(NH_3^+)$, $averge(Hy)$ are the average of corresponding

$pK_a(COOH)$, $pK_a(NH_3^+)$ and Hy, values of all 20 amino acids. Then each amino acid is determined in different octants of the 3D space. Table 1 shows x, y, z coordinates of 20 amino acids after the transformation.

Table 1 The physicochemical properties information of the 20 amino acids

Amino acids	$pK_a(COOH)$	$pK_a(NH_3^+)$	Hy	x	y	z
A	2.35	9.87	1.8	0.163	0.389	2.29
C	1.71	10.78	2.5	-0.477	1.299	2.99
D	1.88	9.60	-3.5	-0.307	0.119	-3.01
E	2.19	9.67	-3.5	0.003	0.189	-3.01
F	2.58	9.24	2.8	0.393	-0.241	3.29
G	2.34	9.60	-0.4	0.153	0.119	0.09
H	1.78	8.97	-3.2	-0.407	-0.511	-2.71
I	2.32	9.76	4.5	0.133	0.279	4.99
K	2.20	8.90	-3.9	0.013	-0.581	-3.41
L	2.36	9.60	3.8	0.173	0.119	4.29
M	2.28	9.21	1.9	0.093	-0.271	2.39
N	2.18	9.09	-3.5	-0.007	-0.391	-3.01
P	1.99	10.60	-1.6	-0.197	1.119	-1.11
Q	2.17	9.13	-3.5	-0.091	-0.351	-3.01
R	2.18	9.09	-4.5	-0.007	-0.391	-4.01
S	2.21	9.15	-0.8	0.023	-0.331	-0.31
T	2.15	9.12	-0.7	-0.037	-0.361	-0.21
V	2.29	9.74	4.2	0.103	0.259	4.69
W	2.38	9.39	-0.9	0.193	-0.091	-0.41
Y	2.20	9.11	-1.3	0.013	-0.371	-0.81

For any given protein sequence with a length of n（$S = S_1 S_2 \cdots S_n$）, each residue can be mapped to a point P_i=(X_i, Y_i, Z_i) in the 3D space starting from the first amino acid. ${S_k}^j$ (j =1,2,3) represents the corresponding component value of the residue for amino acids. So we obtain the coordinates of P_i (i =1,2,…, n) by

$$\phi:\begin{cases} X_i = \sum_{k=1}^{i} {S_k}^1, \\ Y_i = \sum_{k=1}^{i} {S_k}^2, \\ Z_i = \sum_{k=1}^{i} {S_k}^3. \end{cases}$$

Then we convert these space discrete points into continuous parameter curve for the graphical protein description. From the computer graphics knowledge, we can use the spline curve [24] to construct the smooth curve for protein sequences. Here, we provide a relatively

simple cubic Bezier spline curve to construct the protein sequence interpolating space discrete points of amino acids. The cubic Bezier spline curve is smooth with G2 continuity (keep same tangential direction and curvature at the endpoint in the spline curve) and we can use parametric curve's geometric property to analyze the protein sequence's shape similarity.

For a set of given points (for example, 3D discrete points corresponding to amino acids in the protein sequence), we can approach these points by the cubic Bezier spline curve. The detailed construction method is in [25]. If we want to interpolate these points, we need to determine the corresponding control points of each cubic Bezier curves. Namely, given a set of 3D space points Q_i (i=0,1,2,…, n), we look for a corresponding set of d_j (j=0,1,2,…, n) which corresponds to a whole cubic Bezier spline curve. These points d_j (j=0,1,2,…, n) determine four control points B_{3i}, B_{3i+1}, B_{3i+2}, B_{3i+3} (i=0,1,…, n-1) of each cubic Bezier curve and the endpoints of each cubic Bezier curve interpolate points Q_i (i=0,1,2,…, n). For control points of each cubic Bezier curve B_{3i}, B_{3i+1}, B_{3i+2}, B_{3i+3} (i=0,1,…, n-1), we guarantee that $B_0(u)=d_0$, $B_{3n}(u)=d_n$ and

$$B_{3i+1}(u) = (1-\frac{u}{2})d_i + \frac{u}{2}d_{i+1},$$

$$B_{3i+2}(u) = \frac{u}{2}d_i + (1-\frac{u}{2})d_{i+1},\ \ i = 0,1,\cdots n-1.$$

In order to satisfy with G2 continuity at the endpoints in the spline curve, we find

$$B_{3i}(u) = \frac{u}{2}B_{3i-1}(u) + \frac{u}{2}B_{3i+1}(u) = \frac{u}{4}d_{i-1} + (1-\frac{u}{2})d_i + \frac{u}{4}d_{i+1}.$$

Thus, we obtain n-2 equations based on the position relationship between interpolation points and control points as follows

$$\begin{bmatrix} 4-2u & u & & & & \\ u & 4-2u & u & & & \\ & u & \ddots & \ddots & & \\ & & \ddots & 4-2u & u & \\ & & & u & 4-2u & \end{bmatrix}\begin{bmatrix} d_2 \\ d_3 \\ \vdots \\ d_{n-2} \\ d_{n-1} \end{bmatrix} = \begin{bmatrix} 4Q_2 - uQ_1 \\ 4Q_2 \\ \vdots \\ 4Q_{n-2} \\ 4Q_{n-1} - uQ_n \end{bmatrix}.$$

where u is a parameter in [0,1] which adjusts the spline curve shape. Normally we set u=0.5.

We can get points d_j (j=1,2,…,n) by solving a system with n-2 equations and 2 boundary conditions with $d_0 = Q_0$, $d_n = Q_n$. So all control points of cubic Bezier curve B_{3i}, B_{3i+1}, B_{3i+2}, B_{3i+3} (i=0,1,…, n-1) can be determined, then we use de Casteljzu algorithm to draw the cubic Bezier curves, as shown in Fig 1.

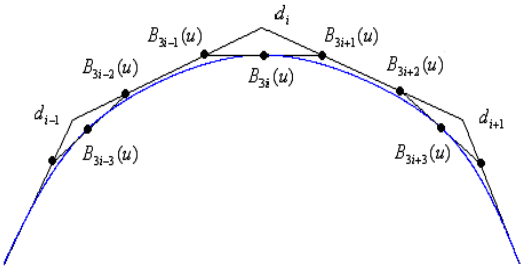

Fig 1. The cubic Bezier spline curve

The whole spline curve interpolating points Q_i (i=0,1,2,…, n) can describe a protein sequence's shape. We use two protein sequences of the yeast Saccharomyces cerevisiae [12] as examples to illustrate the new graphical representation of protein sequences.

Protein1: 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

Protein2: 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 3D space curves of two proteins are shown in Fig 2 and Fig 3. We find the spline curve can more precisely describe the protein sequence's shape and easily observe the difference of protein sequences. Furthermore, we can use the continuous curve's geometric property to analyze the similarity between different protein sequences.

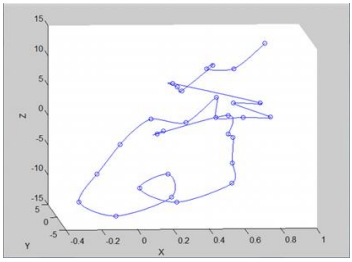

Fig 2. The 3D continuous curve of protein 1

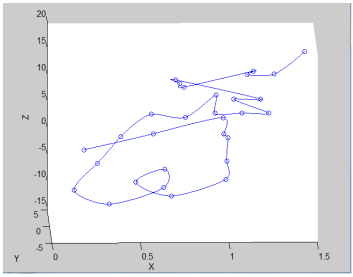

Fig 3. The 3D continuous curve of protein 2

3. Numerical description and similarity analysis of proteins by 3D continuous parameter curve

In order to describe and analyze the similarity of protein sequences, we take some mathematical methods and numerical descriptions for the shape of protein sequences. The graphical representation of protein sequences is a valid method to describe the biological geometric information and its numerical description is a quantitative measure for the biological data research. Recently, some researchers used the graphical representation of protein sequences to construct the measure matrix, such as *ED*, *GD*, *PD*, *D*/*D* and *L*/*L* for the mathematical description, and then propose some invariants related in the matrix to analyze the similarity of biological sequences [4, 5, 13].

In this paper, we fully consider the geometric property of the space continuous curve representing for protein sequences. As we know, the curvature of a point on the curve is one important geometric property which reflects the bending extent of the curve at this point. For a cubic Bezier spline curve of each protein sequence, it can be denoted as $S=S_1S_2S_3\ldots S_n$, where each cubic Bezier curve S_i is denoted as a parameter curve $P=P(t)$, $t\in[0,1]$. The curvature on the point of the curve is computed by [24]

$$k=\frac{\left|\frac{dP(t)}{dt}\times\frac{d^2P(t)}{dt^2}\right|}{\left|\frac{dP(t)}{dt}\right|^3}.$$

Then, the curvature frequency vector of the protein sequence curve is introduced to analyze the similarity between different protein sequences. We firstly take an inclusive interval by setting the minimum curvature on the protein curve as the left endpoint and the maximum curvature as the right endpoint. Secondly we divide it into several subintervals, put curvatures of these interpolating points into divided subintervals and compute the occurrence frequency. Finally we map the divided interval into the standard [0, 1]. So we make a multidimensional vector using the curvature frequency to describe protein sequences. The curvature frequency figures of above protein 1 and protein 2 are shown in Fig 4 and Fig 5.

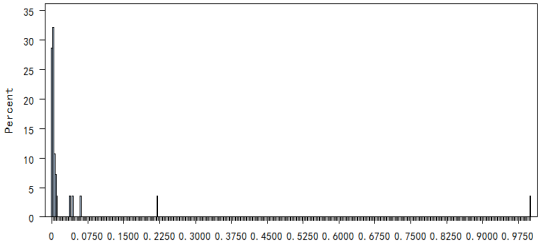

Fig 4. Curvature frequency of protein1

(Horizontal axis is the curvature interval mapped into standard $[0,1]$. Vertical axis is the percentage of the curvature frequency in the given subinterval)

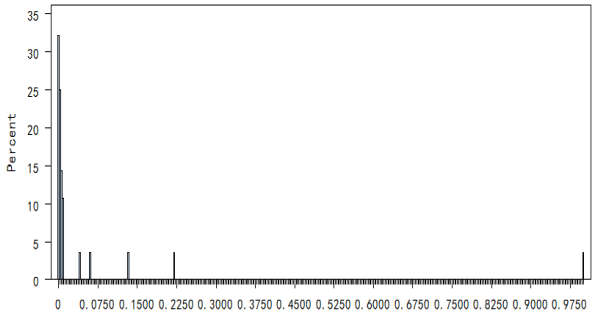

Fig5. Curvature frequency of protein2

The protein sequence similarity can be measured from the distance between these multidimensional vectors. There are several computing methods for measuring the distance between multidimensional vectors, such as Euclidean distance, Elastic-matching distance, Manhattan distance, Minkowski L_n norm. These distance-computing methods have little effect on the similarity analysis and comparison of proteins [26]. Here, we take the simple L_1 distance as the similarity measure between two vectors. Suppose the curvature frequency vectors of two protein sequences as X, Y, the L_1 distance between two vectors is defined as follow

$$d(X,Y)=\sum_{i=1}^{N}\left|X_i-Y_i\right|$$

where N is the number of the curvature frequency vector components, and X_i and Y_i are corresponding components of two vectors X and Y. The smaller is the L_1 distance, the more similar are two protein sequences.

We apply this measure method to compare the curvature frequency vector for different species. We choose ND5 proteins of nine different species from NCBI website, which are shown in Table2. We compute all the curvatures of interpolating points from nine different species' protein space curve, divide it into ten subintervals and create the curvature frequency figures. So each species is described by a ten dimensional vector based on the curvature frequency, as shown in Table 3. Then we calculate the L_1 distance between two vectors of ND5 proteins from nine different species, as shown in Table 4. Observing Table 4, we can find some results as follows:

(1). The distance of F.Whale and B.Whale's vectors are minimum, this shows they are very similar which is consistent with the biology fact.

(2). The distance of ND5 proteins of Human、Gorilla、P.chimpanzee and C.chimpanzee are also small, this shows they are also similar, which is consistent with the evolution relationship.

(3). Opossum is similar to rat and mouse from the curvature frequency analysis probably because opossum is also a kind of mouse, and opossum is regarded as rat or mouse in some regions. Furthermore, from the curvature frequency vector analysis, we find opossum is furthest different from other species, which is accordant to the evolution theory.

Table 2. NADH dehydrogenase information for nine different species' mitochondrion

Species	Sequence code（NCBI）	Sequence length
Human	AP_000649	603
Gorilla	NP_008222	603
Pygmy chimpanzee	NP_008209	603
Common chimpanzee	NP_008196	603
F.Whale	NP_006899	606
B.Whale	NP_007066	606
Rat	AP_004902	610
Mouse	NP_904338	607
Opossum	NP_007105	602

Table 3. 10 dimensional vectors based on the curvature frequency
H(Human); G(Gorilla); P(Pygmy chimpanzee); C(Common chimpanzee); F(F. Whale); B(B. Whale); R(Rat); M(Mouse); O(Opossum).

H	G	P	C	F	B	R	M	O
0.9517	0.9517	0.9501	0.9517	0.9371	0.9371	0.9391	0.9322	0.9266
0.0349	0.0352	0.0365	0.0349	0.0365	0.0364	0.0461	0.0495	0.0500
0.0067	0.0067	0.0067	0.0067	0.0148	0.0149	0.0066	0.0099	0.0117
0.0050	0.0030	0.0033	0.0033	0.0050	0.0050	0.0066	0.0050	0.0050
0	0	0	0	0	0	0	0	0
0.0017	0.0017	0.0017	0.0017	0.0033	0.0033	0	0	0.0017
0	0	0	0	0	0	0	0.0017	0.0017
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0.0017	0.0017	0.0017	0.0033	0.0033	0.0016	0.0017	0.0033

Table 4. L_1 distances of two ten-dimensional vectors for nine different species of ND5 proteins
H(Human); G(Gorilla); P(Pygmy chimpanzee); C(Common chimpanzee); F(F. Whale); B(B. Whale); R(Rat); M(Mouse); O(Opossum).

	H	G	P	C	F	B	R	M	O
H	0	0.0040	0.0066	0.0034	0.0292	0.0292	0.0288	0.0424	0.0502
G		0	0.0032	0.0006	0.0292	0.0292	0.0290	0.0424	0.0502
P			0	0.0032	0.0260	0.0262	0.0258	0.0392	0.0470
C				0	0.0292	0.0292	0.0290	0.0424	0.0502
F					0	0.0002	0.0264	0.0294	0.0304
B						0	0.0266	0.0296	0.0306
R							0	0.0170	0.0282
M								0	0.0112
O									0

4. Comparison

To show the advantage of our method, we employ the alignment-based method to compare different graphical representations of protein sequences [13]. The alignment-free method does not consider biological factors that influence the evolutionary process, but it is computationally convenient with large biological databases. We first show the protein similarity distance of ClustalW approach [27] for nine protein sequences in Table 5. We then calculate the correlation coefficients and do the significance analysis to compare ClustalW approach with our method and other current methods.

The correlation coefficients (r) of our method can be calculated through corresponding rows of Table 4 and Table 5. We first calculate the correlation coefficient of the first row in both matrices, which is relative to human protein. Then we calculate the correlation coefficients for following rows relative to other species. And we similarly calculate the correlation coefficients between Ref [9,16,18]'s method and ClustalW approach. Therefore, we can compare our method with methods in [9,16,18]. The correlation coefficients are shown in Table 6. We find our method has higher correlation coefficients with ClustalW approach than other methods.

Table 5. The similarity distance for nine different species of ND5 proteins based on the ClustalW.

	H	G	P	C	F	B	R	M	O
H	0	10.7	7.1	6.9	41.0	41.3	50.2	48.9	50.4
G		0	9.7	9.9	42.7	42.4	51.4	49.9	54.0
P			0	5.1	40.1	40.1	50.2	48.9	50.1
C				0	40.4	40.4	50.8	49.6	51.4
F					0	3.5	45.3	46.8	52.7
B						0	45.0	45.9	52.7
R							0	25.9	54.0
M								0	50.8
O									0

Table 6. The correlation coefficients for nine ND5 proteins of our method and the approaches in Ref [9,16,18] , as compared with ClustalW method.

	Our method	Ref [9] method	Ref[16] method	Ref [18] method
H	0.9485	0.9282	0.9405	0.8985
G	0.9579	0.7784	0.9374	0.7942
P	0.9405	0.9341	0.9431	0.8993
C	0.9540	0.9404	0.8778	0.9089
F	0.9819	0.7412	0.6496	0.7895
B	0.9838	0.8054	0.8123	0.8140
R	0.9879	0.7376	0.6450	0.8013
M	0.7605	0.7145	0.6236	0.7787
O	0.6688	0.6146	0.4728	0.6850

We also make the significance analysis for the correlation coefficients since it may get the higher correlation coefficients for a small data set. We similarly compute the significance for the correlation coefficients that are greater than 0.7 through *t*-test [4,13,16]. Our sample size is 9, so the degree of freedom is 7. The *t*-values of the correlation coefficients are shown in Table 7. A *t*-value of more than 2.365 indicates that a significance of less than 0.05 chance of having occurred by coincidence. We realize all computed t-values are greater than 2.365 from Table7.

Table 7. The *t*-values computed for the correlation coefficients $|r|>0.7$, based on them the significance is determined.

	Our method	Ref [9] method	Ref [16] method	Ref [18] method
H	7.9220	6.6001	7.3231	5.4154
G	8.8274	3.2806	7.1216	3.4580
P	7.3231	6.9225	7.5042	5.4405
C	8.4189	7.3163	4.8482	5.7665
F	13.7163	2.9213	_	3.4034
B	14.5194	3.5950	3.6848	3.7076
R	16.8528	2.8901	_	3.5437
M	3.0987	2.7020	_	3.2838
O	_	_	_	_

5. Conclusion and future work

In this paper, we propose a novel method for graphical representation of protein sequences based on the physicochemical properties of amino acids. Because we construct the continuous cubic Bezier spline space curve to describe the protein shape, we can use curve's geometric property such as curvature to analyze the similarity of protein sequences.

From current experimental results, we find our proposed method provides an effective way for the protein's whole similarity analysis. In our future work, we will extend the geometric property of protein curves and combine more physicochemical properties of amino acids to analyze local structure and functions of protein sequences.

Acknowledgments. This research was supported from the National Natural Science Foundation of China (No. 60903143 and 51075421); Natural Science Foundation of Zhejiang Province of China (No. Y1110504); Science and Technology Project of Zhejiang Province of China (No. 2012C21035); Young Researchers Foundation of Zhejiang Provincial Top Key Academic Discipline of Mechanical Design and Theory and Zhejiang Sci-Tech University Key Laboratory (ZSTUMD2011B004).

References

[1] 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.

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

[3] B. Liao, C. Zeng, F. Q. Li, Y. Tang, Analysis of similarity/dissimilarity of DNA sequences based on dual nucleotides, *MATCH Commun. Math. Comput. Chem.* **59** (2008) 647–652.

[4] 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.

[5] 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.

[6] Z. B. Liu, B. Liao, W. Zhu, A new method to analyze the similarity based on dual nucleotides of the DNA sequence, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 541–552.

[7] Y. Li, G. H. Huang, B. Liao, Z. Liu, H-L curve: a novel 2-D graphical representation of protein sequences, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 519–532.

[8] H. Gonzalez-Diza, L. G. Perez-Montoto, A. Duardo-Sanchez, E. Paniagua, S. Vazquez-Prieto, R. Vilas, M. A. Dea-Ayuela, F. Bolas-Fernandez, C. R. Munteanu, J. Dorado, J. Costas, F. M. Ubeira, Generalized lattice graphs for 2D-visualization of biological information, *J. Theor. Biol*. **261** (2009) 136-147.

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

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

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

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

[13] 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.

[14] M. Randić, J. Zupan, M. Novič, On 3-D graphical representation of proteomics maps and their numerical characterization, *J. Chem. Inf. Comput. Sci*. **45** (2001) 1339–1344.

[15] C. Li, X. Q. Yu, L. Yang, X. Q. Zheng, Z. F. Wang, 3-D maps and coupling numbers for protein sequences, *Physica A* **388** (2009) 1967–1972.

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

[17] B. Liao, B. Y. Liao, X. M. Sun, Q. G. Zeng, A novel method for similarity analysis and protein sub- cellular localization prediction, *Bioinformatics* **26** (2010) 2678–2683.

[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] J. Feng, T. M. Wang, Characterization of protein primary sequences based on partial ordering, *J. Theor. Biol*. **254** (2008) 752–755.

[20] H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Res*. **18** (1990) 2163–2170.

[21] F. L. Bai, T. M. Wang, On graphical and numerical representation of protein sequences, *J. Bimol. Struct. Dyn.* **23** (2006) 537–546.

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

[23] P. A. He, Y. P. Zhang, Y. H. Yao, Y. F. Tang ,X. Y. Nan, The graphical representation of protein sequences based on the physicochemical properties and its applications, *J. Comput. Chem*. **31** (2010) 2136–2142.

[24] G. Farin, *Curves and Surfaces for Computer–Aided Geometric Design: A Practical Guide*, Academic Press, San Diego, 1997.

[25] Z. Li, D. S. Meek, D. J. Walton, A smooth, obstacle-avoiding curve, *Computers Graphics* **30** (2006) 581–587.

[26] Y. Fang, Y. S. Liu, K. Ramani, Three dimensional shape comparison of flexible proteins using the local-diameter descriptor, *BMC Struct. Biol.* **9** (2009) 29–29.

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