American Journal of Biochemistry and Biotechnology

Original Research Paper

A Novel 2D Graphical Representation and its Application in the Similarities/Dissimilarities Analysis of Protein Sequences

Xianyou Zhu

Department of Computer Science, Hengyang Normal University, 421008, P.R. China

Article history
Received: 08-01-2016
Revised: 04-03-2016
Accepted: 14-03-2016

Email: 364715358@qq.com

Abstract: In this study, a novel 2D graphical representation of protein sequences is proposed based on the physicochemical feature pK2 of amino acids first and then, on the basis of the newly given 2D graphical representation, a new concept of feature appearance model is introduced to analyze the similarity/dissimilarity of protein sequences. Finally, Theoretical and simulation results show that the newly proposed method is effective in similarities/dissimilarities analysis of protein sequences.

Keywords: Graphical Representation, Similarity/Dissimilarity Analysis, Protein Sequence, Feature Appearance Model

Introduction

Graphical representation of protein sequences is a very powerful tool for visual comparison of protein sequences (Yao *et al*., 2010; Wang *et al*., 2014; 2015). Currently, many effective graphical presentation methods have been proposed to facilitate the analysis of similarities/dissimilarities among the protein sequences. For example, Feng and Zhang (2002) proposed a 2D graphical representation of protein sequence based on the hydrophobicity and charged properties of amino acid residues along the primary sequence. Wen and Zhang (2009) proposed a 2D graphical representation of protein sequence with no circuit or degeneracy based on the chosen physicochemical properties of amino acids. Huang *et al*. (2013) introduced a 2D graphical representation of protein sequence, called HR-Curve, based on classification and dual vectors. Qi *et al*. (2012) proposed a 2D graphical representation of protein sequence based on Huffman tree. Abo-Elkhier (2012) proposed a 3D graphical representation of protein sequence on the basis of a right cone of a unit base and unit height on protein sequences interfaces. Hea *et al*. (2012) introduced a 3D graphical representation, which is a cyclic order of 20 amino acids, based on the order of 6-bit binary Gray code. Abo el Maaty *et al*. (2010) introduced a 3D graphical representation of protein sequence based on three physicochemical properties of amino acid side chains.

In this study, a novel 2D graphical representation of protein sequences is proposed based on a chosen physicochemical feature pK2 of amino acids first and then, 4 descriptors are extracted from the 2D graphical representation of protein sequences and adopted to analyze the similarities/dissimilarities of protein sequences quantitatively. Theoretical and simulation results show that the newly given method is effective in similarities/dissimilarities analysis of protein sequences and can achieve results that are consistent with the results of the known fact of evolution.

Graphical Representation of Protein Sequences

Proteins are composed of 20 different amino acids and these amino acids have many different physicochemical and biological properties such as the molecular weight (*mW*), iselectric point (*p*I), the *pKa* value for terminal amino acid groups COOH (*pK*1), the *pKa* value for terminal amino acid groups NH_3^+ (*pK*2), van der waals radius (Vdwa), kdHydrophobicity (kh) (Kyte and Doolittle, 1982), wwHydrophobicity (wh) (Wimley and White, 1996), hhHydrophobicity (hh) (Hessa *et al*., 2005), the occurrence in human properties (%) (Oihp (%)), Abundance (Abu), ATP cost in synthesis under aerobic condition (Csae) and ATP cost in synthesis under anaerobic condition (Csan) etc. The names and symbols of the 20 amino acids and the value of their 12 major properties are illustrated in the following Table 1.

Science Publications

Table 1. The full list of 20 amino acids and the values of their 12 different properties

Amino acid	Short	Abbrev.	*m*W	*p*I	*p*K1	*p*K2	Vdwv	Oihp (%)	Abu	Csae	Csan	kh	wh	hh
Alanine	A	Ala	89.09404	6.01	2.35	9.87	67	7.8	2.90	-1	1	1.8	-0.17	0.11
Cysteine	C	Cys	121.15400	5.05	1.92	10.70	86	1.9	0.52	11	15	2.5	0.24	-0.13
Aspartic acid	D	Asp	133.10380	2.85	1.99	9.90	91	5.3	1.40	0	2	-3.5	-1.23	3.49
Glutamic acid	E	Glu	147.13070	3.15	2.10	9.47	109	6.3	1.50	-7	-1	-3.5	-2.02	2.68
Phenylalanine	F	Phe	165.19180	5.49	2.20	9.31	135	3.9	1.10	-6	2	2.8	1.13	-0.32
Glycine	G	Gly	75.06714	6.06	2.35	9.78	48	7.2	3.50	-2	2	-0.4	-0.01	0.74
Histidine	H	His	155.15630	7.60	1.80	9.33	118	2.3	0.54	1	7	-3.2	-0.96	2.06
Isoleucine	I	Ile	131.17460	6.05	2.32	9.76	124	5.3	1.70	7	11	4.5	0.31	-0.60
Lysine	K	Lys	146.18930	9.60	2.16	9.06	135	5.9	2.00	5	9	-3.9	-0.99	2.71
Leucine	L	Leu	131.17460	6.01	2.33	9.74	124	9.1	2.60	-9	1	3.8	0.56	-0.55
Methionine	M	Met	149.20780	5.74	2.13	9.28	124	2.3	0.88	21	23	1.9	0.23	-0.10
Asparagine	N	Asn	132.11900	5.41	2.14	8.72	96	4.3	1.40	3	5	-3.5	-0.42	2.05
Proline	P	Pro	115.13190	6.30	1.95	10.64	90	5.2	1.30	-2	4	-1.6	-0.45	2.23
Glutamine	Q	Gln	146.14590	5.65	2.17	9.13	114	4.2	1.50	-6	0	-3.5	-0.58	2.36
Arginine	R	Arg	174.20270	10.76	1.82	8.99	148	5.1	1.70	5	13	-4.5	-0.81	2.58
Serine	S	Ser	105.09340	5.68	2.19	9.21	73	6.8	1.20	-2	2	-0.8	-0.13	0.84
Threonine	T	Thr	119.12030	5.60	2.09	9.10	93	5.9	1.50	6	8	-0.7	-0.14	0.52
Valine	V	Val	117.14780	6.00	2.39	9.74	105	6.6	2.40	-2	2	4.2	-0.07	-0.31
Tryptophan	W	Trp	204.22840	5.89	2.46	9.41	163	1.4	0.33	-7	7	-0.9	1.85	0.30
Tyrosine	Y	Tyr	181.19120	5.64	2.20	9.21	141	3.2	0.79	-8	2	-1.3	0.94	0.68

Let $\{F_1, F_2, \ldots, F_{12}\}$ represent these 12 different properties of amino acids illustrated in above Table 1 and Ω = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y} be the set of 20 kinds of amino acids, $\forall \tau \in \Omega$, let $X^1_\tau, X^2_\tau, \ldots, X^{12}_\tau$ be the values of 12 features of τ, $\overline{Y}^i = \dfrac{\sum_{\eta\in\Omega} X^i_\eta}{20}$, then we can standardize the values of 12 features of τ according to the following Formula 1:

$$Y^i_\tau = X^i_\tau - \overline{Y}^i \tag{1}$$

where, $i \in \{1,2,\ldots,12\}$.

Let $\Psi = p_1p_2\ldots p_n$ ($p_i \in \Omega$, $\forall i \in \{1,2,\ldots,N\}$) represent a protein sequence with n amino acids and for any given letter $u \in \Omega$, supposing that u appears K times in the protein sequence Ψ totally and the location of u at the j^{th} time in Ψ is u_j, then we call the vector $<u_1, u_2, \ldots, u_k>$ as the "Feature Appearance Model" of u in Ψ.

Based on the concept of "Feature Appearance Model" proposed above, then for each protein sequence Ψ, we can obtain its graphical representation according to the following steps:

Step1: According to the concept of Feature Appearance Model, obtain 20 different Feature Appearance Models of amino acids in the protein sequence Ψ.

Step2: For j = 1 to 12, select the jth feature from these 12 features of amino acids $\{F_1, F_2, \ldots, F_{12}\}$, $\forall p_i \in \Psi$, let the standardized values of the jth feature of p_{i-1}, p_i and p_{i+1} be Y^j_{i-1}, Y^j_i and Y^j_{i+1} respectively and the Feature Appearance Model of p_i be $\Delta_i = <u^i_1, u^i_2, \ldots u^i_k>$, then we can obtain k different coordinates of p_i such as $(x^i_1, y^i_1), (x^i_2, y^i_2), \ldots, (x^i_k, y^i_k)$ according to Formula 2 and 3:

$$x^i_t = u^i_t \tag{2}$$

$$Y^i_t = Y^i_{t-1} + \left(x^i_t - x^i_{t-1}\right) * \left(\frac{Y^j_{i+1} - Y^j_i + 1}{Y^j_i - Y^j_{i-1} + 1}\right) \tag{3}$$

where, $t \in \{1,2,\ldots,k\}$ and for $p_1 \in \Psi$, since there isn't p_0 in Ψ, then we define $Y^j_0 = 0$, for $j \in \{1,2,\ldots,12\}$.

Obviously, for each different amino acid in the protein sequence Ψ, after connecting all of its coordinates, then we can obtain 20 different curves for the protein sequence Ψ, since it has 20 different amino acids. Therefore, through above steps, we can translate a protein sequence into a graph with 20 curves according to each feature of amino acids and in addition, as for the 12 different features of amino acids, we can obtain 12 groups of curves for the protein sequence Ψ and in each group, there are 20 different curves.

Similarities/Dissimilarities analysis Model of Protein Sequences

Let G_Ψ represent a graph of Ψ obtained by the method given above in section 2 and $\forall p_t \in \Psi$, $t \in [1,20]$, let the Feature Appearance Model of p_t in Ψ be $\Delta_t = <u^t_1, u^t_2, \ldots u^t_k>$ and $E_t \in G_\Psi$ represent the curve of p_t in G_Ψ, then we can obtain the *ED*, *PD*, *D/D*, *L/L* matrixes of E_t according to the following Formula 5-8 respectively (Randic *et al*., 2000; 2003a; 2003b; 2003c; Randic, 2003; Randic *et al*., 2004; Randic and Vracko, 2003; Bajzer *et al*., 2003). Thereafter, let $M^t_e, M^t_p, M^t_d, M^t_l$ represent the *ED*, *PD*, *D/D*, *L/L* matrixes of E_t respectively, we will finally obtain 80 different matrixes for Ψ, since there are 20 curves in G_Ψ:

$$[ED]_{ij} = \sqrt{(u^i_1 - u^j_1)^2 + (u^i_2 - u^j_2)^2 + \ldots + (u^i_k - u^j_k)^2} \tag{5}$$

$$[PD]_{ij} = \begin{cases} [ED]_{i,i+1} + [ED]_{i+1,i+2} + \ldots + [ED]_{j-1,j}, if\ i < j \\ 0.\ \ if\ i = j \end{cases} \tag{6}$$

$$[D/D]_{ij} = \begin{cases} [ED]_{i,j} / |j-i|, if\ i \neq j \\ 0. \quad if\ i = j \end{cases} \tag{7}$$

$$[L/L]_{ij} = \begin{cases} [ED]_{i,j} / [PD]_{i,j}, if\ i \neq j \\ 0. \quad if\ i = j \end{cases} \tag{8}$$

Based on the matrixes $M_e^t, M_p^t, M_d^t, M_l^t$ obtained above, then we can compute out the maximum eigenvalues $\{x_e^t, x_p^t, x_d^t, x_l^t\}$, the minimum eigenvalues $\{n_e^t, n_p^t, n_d^t, n_l^t\}$, the average eigenvalues $\{v_e^t, v_p^t, v_d^t, v_l^t\}$, the sum of the maximum and minimum eigenvalues $\{s_e^t, s_p^t, s_d^t, s_l^t\}$, the index values of ALE $\{a_e^t, a_p^t, a_d^t, a_l^t\}$ of these matrixes respectively (Li and Wang, 1966; Shrock and Tsai, 1997; Biggs, 1974). Hence, for $E_t \in G_\Psi$, we can describe it with a 20 dimensional vector V_t as the following Formula 9:

$$\begin{aligned} V_t = < & x_e^t, n_e^t, v_e^t, s_e^t, a_e^t, x_p^t, n_p^t, v_p^t, s_p^t, \\ & a_p^t, x_d^t, n_d^t, v_d^t, s_d^t, a_d^t, x_l^t, n_l^t, v_l^t, s_l^t, a_l^t > \end{aligned} \tag{9}$$

Thereafter, the graph of Ψ can be represented as a 20×20 matrix $M_\Psi = [V_1, V_2, \ldots, V_{20}]^T$, called the Descriptor Matrix.

Based on the Descriptor Matrix obtained above, we randomly select k ($k \in [1,20]$) columns from M_Ψ each time, then we will obtain a new 20×k matrix $M_\Psi^k = \left[V_{k1}^\Psi, V_{k2}^\Psi, \ldots, V_{k20}^\Psi\right]^T$, where V_{kj}^Ψ is a k dimensional vector for any $j \in [1,20]$. Therefore, for any two protein sequences Ψ_1 and Ψ_2, supposing that we have obtained two 20×k matrix $M_1^k = \left[V_{k1}^1, V_{k2}^1, \ldots, V_{k20}^1\right]^T$ and $M_2^k = \left[V_{k1}^2, V_{k2}^2, \ldots, V_{k20}^2\right]^T$ and $V_{kj}^1 =< d_{kj}^{i1}, d_{kj}^{i2}, \ldots, d_{kj}^{ik} >$ for any $i \in \{1,2\}$ and $j \in [1,20]$, then we can obtain the distance $d(\Psi_1, \Psi_2)$ between Ψ_1 and Ψ_2 as follows:

$$d(\Psi_1, \Psi_2) = \sum\nolimits_{i=1}^{20} V_{ki}^1 - V_{ki}^2 \tag{10}$$

Where:

$$\left\| V_{ki}^1 - V_{ki}^2 \right\| = \sqrt{\sum\nolimits_{t=1}^{k} \left(d_{kj}^{1t} - d_{kj}^{2t}\right)^2} \tag{11}$$

Table 2. The basic information of 16 ND5 protein sequences

No.	Name	abbreviation	Access No	Length
1	Human	human	ADT80430.1	603
2	Gorilla	gorilla	NP_008222	603
3	Pigmy Chimpanzee	pi-chim	NP_008209	603
4	Common Chimpanzee	c-chim	NP_008196	603
5	Fin Whale	fin-whale	NP_006899	606
6	Blue Whale	blue-whale	NP_007066	606
7	Rat	rat	AP_006899	610
8	Mouse	mouse	NP_904338	607
9	Opossum	opossum	NP_007105	602
10	Sheep	sheep	ABW22903.1	606
11	Goat	goat	BAN59258.1	606
12	Lemur	lemur	CAD13431.1	603
13	Cattle	cattle	ADN11902.1	606
14	Hare	hare	CAD13291.1	603
15	Gallus	Gallus	BAE16036.1	605
16	Rabbit	rabbit	NP_007559.1	603

Table 3. The basic information of 13 beta-globin proteins

No.	Access No.	Abbreviation	Length
1	CAA25111	Bovine	145
2	CAA26204	Chimpanzee	125
3	CAA68429	Hare	147
4	CAA23700	Gallus	147
5	AAA30913	Goat	145
6	CAA43421	Gorilla	121
7	AAA16334	Human	147
8	AAA36822	Lemur	147
9	CAA24101	Mouse	147
10	AAA30976	Opossum	147
11	CAA24251	Rabbit	147
12	CAA29887	Rat	147
13	NP_001091117	Sheep	145

Table 4. The basic information of 29 spike proteins

No.	Access No.	abbreviation	length
1	CAB91145	TGEVG	1447
2	NP_058424	TGEV	1447
3	AAK38656	PEDVC	1383
4	NP_598310	PEDV	1383
5	NP_937950	HCoVOC43	1361
6	AAK83356	BCoVE	1363
7	AAL57308	BCoVL	1363
8	AAA66399	BCoVM	1363
9	AAL40400	BCoVQ	1363
10	AAB86819	MHVA	1324
11	YP_209233	MHVJHM	1376
12	AAF69334	MHVP	1321
13	AAF69344	MHVM	1324
14	AAP92675	IBVBJ	1169
15	AAS00080	IBVC	1169
16	NP_040831	IBV	1162
17	AAS10463	GD03T0013	1255
18	AAU93318	PC4127	1255
19	AAV49720	PC4137	1255
20	AAU93319	PC4205	1255
21	AAU04646	civet007	1255
22	AAU04649	civet010	1255
23	AAV91631	A022	1255
24	AAP51227	GD01	1255
25	AAS00003	GZ02	1255
26	AAP30030	BJ01	1255
27	AAP50485	FRA	1255
28	AAP41037	TOR2	1255
29	AAQ01597	TaiwanTC1	1255

Based on the Formula 10, we can obtain three other distance matrixes M_{oN}, M_{oG} and M_{oS} according to three groups of protein sequences such as the 16 ND5 protein sequences, 13 globin protein sequences and 29 sequences of spike protein respectively. The basic information of these three groups of protein sequences are illustrated in the following Table 2 to 4.

And in addition, when adopting the ClustalW algorithm (Thompson *et al*., 1994) and the software MEGA (Tamura *et al*., 2013) to obtain the distance matrixes for each group of protein sequences such as the 16 ND5 protein sequences, 13 globin protein sequences and 29 sequences of spike protein, then we can also obtain three distance matrixes M_{sN}, M_{sG} and M_{sS} according to these three groups of protein sequences respectively.

Cnsidering the above two groups of distance matrixes $\{M_{sN}, M_{sG}, M_{sS}\}$ and $\{M_{oN}, M_{oG}, M_{oS}\}$, $\forall \Pi \in \{N, G, S\}$, supposing that the number of columns in matrix $M_{s\Pi}$ is N_Π, $M_{s\Pi} = \left[V_{s\Pi}^1, V_{s\Pi}^2, \ldots, V_{s\Pi}^{N_\Pi}\right]$ and $M_{s\Pi} = \left[V_{o\Pi}^1, V_{o\Pi}^2, \ldots, V_{o\Pi}^{N_\Pi}\right]$, then we define the Average Correlation Coefficient *Acorr* ($M_{s\Pi}$, $M_{s\Pi}$) between $M_{s\Pi}$ and $M_{s\Pi}$ as follows:

$$Acorr\left(M_{s\Pi}, M_{s\Pi}\right) = \frac{\sum_{i=1}^{N_\Pi} corrcoef(V_{s\Pi}^i, V_{o\Pi}^i)}{N} \tag{12}$$

where, $corrcoef\left(V_{s\Pi}^i, V_{o\Pi}^i\right)$ is the correlation coefficient between the i^{th} column vectors $V_{s\Pi}^i$ and $V_{o\Pi}^i$ in the matrixes $M_{s\Pi}$ and $M_{s\Pi}$ respectively. Thus, we will obtain three different Average Correlation Coefficients such as $Acorr(M_{sN}, M_{oN})$, $Acorr(M_{sG}, M_{oG})$, $Acorr(M_{sS}, M_{oS})$.

Let $AvgCC = (Acorr(M_{sN}, M_{oN}) + Acorr(M_{sG}, M_{oG}) + Acorr(M_{sS}, M_{oS}))/3$, since there are totally 12 kinds of features of amino acids, then we can obtain 12 different graphs for each protein sequence and in each graph, there are 20 different curves. Additionally, according to the Formula 9, we can know that each curve in a graph can be described by a 20 dimensional vector, then, it is obvious that we will obtain lots of values of *AvgCC*.

For the protein sequence Ψ, supposing that we finally obtain Γ different values of *AvgCC* such as $\{AvgCC_1, AvgCC_2, \ldots, \text{Avg}CC_r\}$ and there is $AvgCC_1 \geq AvgCC_2 \geq \ldots \geq \text{Avg}CC_r$ and in addition, supposing that to obtain the value of $AvgCC_1$, we shall select the *J*th ($J \in [1,12]$) feature from these 12 features of amino acids and *K* ($K \in [1,20]$) columns $\left\{d_{KJ}^1, d_{KJ}^2, \ldots, d_{KJ}^K\right\}$ from M_Ψ, then, we will call the *J*th feature of amino acids as the Optimal Feature for the protein sequence Ψ and $\left\{d_{KJ}^1, d_{KJ}^2, \ldots, d_{KJ}^K\right\}$ as the Optimal Descriptors of the protein sequence Ψ.

Obviously, the Optimal Feature obtained above can be utilized for graphical representation of protein sequences and the Optimal Descriptors obtained above can be utilized for analyzing the similarities/dissimilarities of protein sequences.

Based on three groups of protein sequences such as the 16 ND5 protein sequences, 13 globin protein sequences and 29 sequences of spike protein, through experiments, it is easy to prove that the Optimal Feature will be *pK*2 and the Optimal Descriptors will be $\{v_p^t, v_d^t, s_d^t, a_l^t\}$ for $t \in$ [1, 20] and for convenience, we rewrite the 20 different Optimal Descriptors as a matrix $\begin{bmatrix} v_p^1 & v_d^1 & s_d^1 & a_l^1 \\ v_p^2 & v_d^2 & s_d^2 & a_l^2 \\ \cdots & \cdots & \cdots & \cdots \\ v_p^{20} & v_d^{20} & s_d^{20} & a_l^{20} \end{bmatrix}$.

Hence, we can adopt *pK*2 and $\begin{bmatrix} v_p^1 & v_d^1 & s_d^1 & a_l^1 \\ v_p^2 & v_d^2 & s_d^2 & a_l^2 \\ \cdots & \cdots & \cdots & \cdots \\ v_p^{20} & v_d^{20} & s_d^{20} & a_l^{20} \end{bmatrix}$ as parameters to construct a new Similarities/Dissimilarities Analysis Model according to the following steps:

Step1: For each protein sequence Ψ in these two groups of protein sequences such as 16 ND6 proteins and 15 myoglobin proteins, obtain its graphical representation G_Ψ based on the feature *pK*2.

Step2: According to the Formula 10 and the matrix of Optimal Descriptors $\begin{bmatrix} v_p^1 & v_d^1 & s_d^1 & a_l^1 \\ v_p^2 & v_d^2 & s_d^2 & a_l^2 \\ \cdots & \cdots & \cdots & \cdots \\ v_p^{20} & v_d^{20} & s_d^{20} & a_l^{20} \end{bmatrix}$, obtain two distance matrixes M_{oND6} and $M_{oGlobin}$ for these two groups of protein sequences such as 16 ND6 proteins and 15 myoglobin proteins respectively.

Step3: Utilize these distance matrixes M_{oND6} and $M_{oGlobin}$ to analyze the similarity/dissimilarity of protein sequences numerically.

Results and Analysis

Graphical Representation of Protein Sequences

According to the new Similarities/Dissimilarities Analysis Model given above, the following Fig. 1 illustrates some graphs of the protein sequences in the group of 16 ND6 proteins.

From the graphs in Fig. 1, it is easy to see that the four graphs (*a*), (*b*), (*c*) and (*d*) are similar to each other and it is obvious that the phenomenon is totally consistent with the results of the known fact of evolution.

Similarity/Dissimilarity Analysis of Protein Sequences

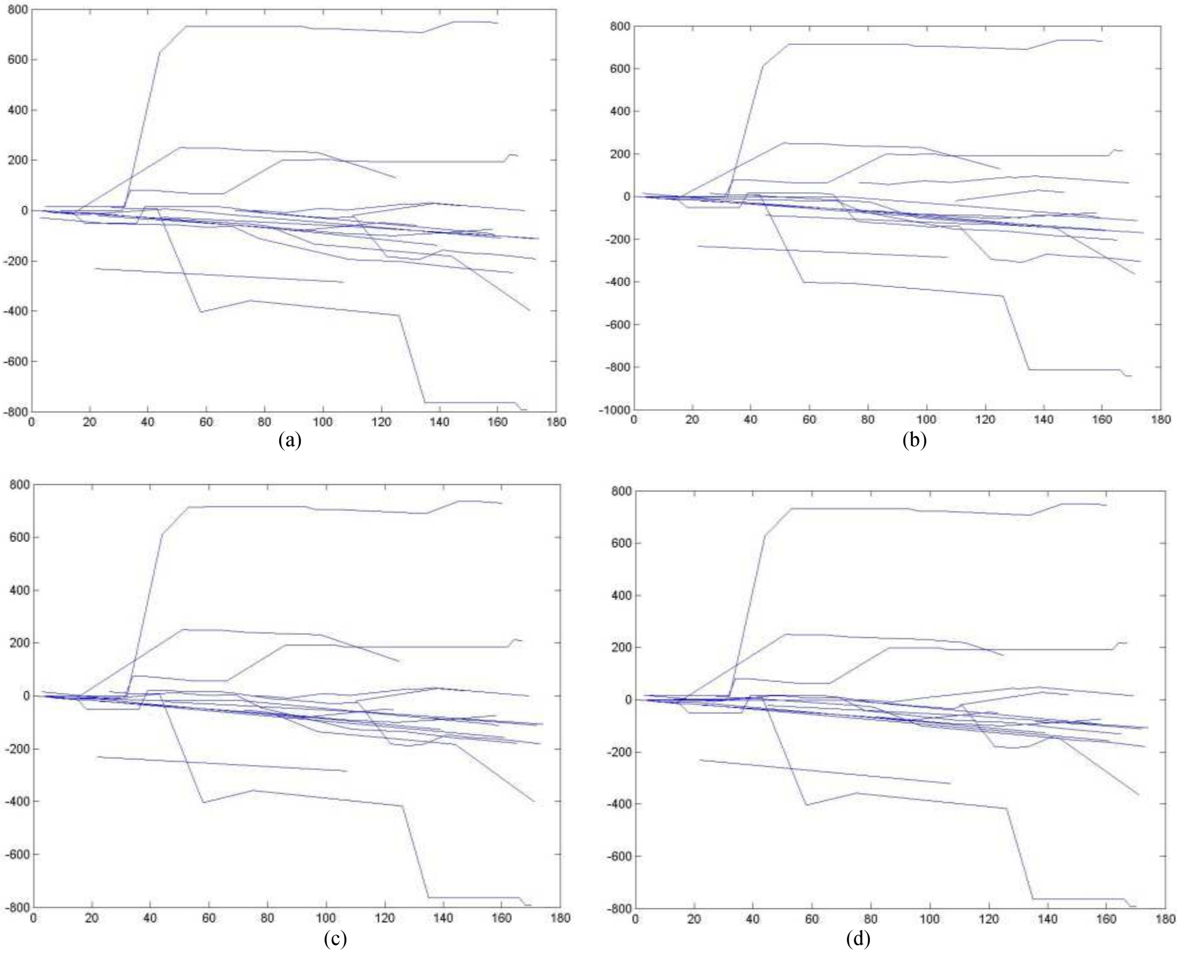
According to the Similarities/Dissimilarities Analysis Model proposed above, the distance matrixes of 16 ND6 protein sequence and 15 myoglobin are illustrated in the following Table 5 and 6 respectively.

Table 5 the distance matrix of 16 ND5 protein sequences

	Homo_sapiens	Gorilla_gorilla	Pan_paniscus	Pan_troglodytes	Balaenoptera_phy	Balaenoptera_musculus	Rattus_norvegicus	
Homo_sapiens	0							
Gorilla_gorilla	428.2748757	0						
Pan_paniscus	245.5669038	336.8824763	0					
Pan_troglodytes	340.1705447	447.8657651	268.2495187	0				
Balaenoptera_physalus	2180.365673	2258.135448	2096.898048	2142.684701	0			
Balaenoptera_musculus	2491.217014	2565.304013	2368.501914	2425.835129	1254.549885	0		
Rattus_norvegicus	2753.373767	2849.159873	2683.337261	2684.03185	2227.793299	2186.302845	0	
Mus_musculus	3662.605201	3799.734652	3633.175375	3648.099271	3335.838573	2700.839095	2678.818806	
Didelphis_virginiana	4316.200696	4104.244838	4263.841271	4143.817458	3981.720493	4360.106801	3371.806463	
Ovis_aries	2409.081221	2394.700083	2332.770699	2353.680291	1877.470505	1517.935117	2135.963633	
Capra_hircus	2469.115214	2601.44016	2393.750355	2427.060012	1752.395879	1277.574038	2195.515599	
Bos_taurus	3206.492969	3365.909897	3142.38396	3185.255944	2471.622027	1864.618091	2688.584535	
Lepus_europaeus	3571.516197	3698.548608	3502.319277	3466.216102	3616.846879	3496.748826	4099.512328	
Oryctolagus_cuniculus	3388.235028	3515.446404	3321.589813	3286.444549	2608.684275	2244.767401	2897.211941	
Lemur_catta	2575.36327	2807.634676	2663.289155	2558.013663	2333.9619	2012.216737	2666.613615	
Gallus_gallus	3241.358429	3360.398648	3238.673151	3301.064643	3071.200182	3124.79419	3241.540081	
Mus_musculus	Didelphis_virginiana	Ovis_aries	Capra_hircus	Bos_taurus	Lepus_europaeus	Oryctolagus_cuniculus	Lemur_catta	Gallus_gallus
0								
5149.92453	0							
2494.320384	4050.077959	0						
2641.029007	4097.168682	906.405623	0					
1608.468722	5060.087872	1586.67092	1458.68122	0				
3047.550859	6290.054664	3818.373622	3537.483632	2546.059499	0			
3323.490281	5048.285883	2575.290879	2407.116075	2682.238934	2645.907326	0		
2963.445449	4459.506653	2455.647047	1925.14231	2204.106617	3434.545587	2857.041294	0	
3462.321596	4887.210993	3707.025511	3533.594378	3569.877383	4054.228312	3461.180234	3432.31125	0

Table 6. The distance matrix of 15 myoglobin protein sequences

	Notothenia_coriiceps	Bos_taurus	Gallus_gallus	Iguana_iguana	Equus_caballus	Mus_musculus	Homo_sapiens
Notothenia_coriiceps	0						
Bos_taurus	4097.44	0					
Gallus_gallus	4792.566084	3560.791165	0				
Iguana_iguana	4546.509684	2783.159901	3096.824624	0			
Equus_caballus	4818.318252	1981.668997	3896.189784	2905.472712	0		
Mus_musculus	4281.901025	1961.600345	3512.368987	2356.68369	1767.894756	0	
Homo_sapiens	4172.328297	1770.811004	3511.86861	1946.100427	1541.611574	1546.595503	0
Neopagetopsis_ionah	388.6511303	4103.159687	4793.162394	4553.035859	4747.32489	4200.922719	4173.495044
Rattus_norvegicus	4547.077255	2138.125443	3916.075065	2579.303332	1520.817094	770.1228191	1208.12297
Chionodraco_rastrospinosus	788.1577625	3832.187413	5079.82546	4723.461676	4499.044348	4078.680073	3846.72076
Sus_scrofa	4078.13495	1415.236077	3182.939886	2155.714222	1328.71691	972.4538495	891.3569498
Ursus_maritimus	4108.557948	1623.441529	3226.768149	2094.694607	1373.023142	940.6168992	866.6113645
Ovis_aries	3911.856584	712.9417429	3603.264337	2532.558954	1678.386853	1868.95617	1524.623599
Physeter_catodon	4033.473164	2166.939307	4032.88236	2389.26319	2582.56704	2436.8821	1836.51924
Leptonychotes_weddellii	6401.127904	3911.447127	3275.703168	3890.454652	4263.29865	3851.768142	3889.825876
Neopagetopsis_ionah	Rattus_norvegicus	Chionodraco_rastrospinosus	Sus_scrofa	Ursus_maritimus	Ovis_aries	Physeter_catodon	Leptonychotes_weddellii
0							
4470.329176	0						
931.5906197	4216.221641	0					
4051.300805	1172.15984	3805.513683	0				
4072.262136	1151.578848	3828.675723	342.7574541	0			
3834.094708	1968.215915	3675.697554	1199.066245	1385.761872	0		
3937.800293	2506.956876	3658.640427	1847.37875	1888.815412	2115.16987	0	
6374.360594	3671.540354	6661.115957	3743.34404	3678.503902	4499.12423	4646.951581	0



(a)

(b)

(c)

(d)

Fig. 1. Some graphs of 16 ND6 protein sequences based our method (a) Homo sapiens (Human) (b) Gorilla gorilla (Gorilla) (c) Pan paniscus (P-Chim) (d) Pan troglodytes (C-Chim)

From Table 5, it is easy to find that there are some similar pairs such as (Human, P-Chim) with the distance 245.57, (Human, C-Chim) with the distance 340.17, (Human, Gorilla) with the distance 428.27, (Gorilla, P-Chim) with the distance 336.88, (Gorilla, C-Chim) with the distance 447.87, (P-Chim, C-Chim) with the distance 268.25, (Fin-Wha, Blu-Wha) with the distance 1254.55 and (Sheep, Goat) with the distance 906.41, etc,. and among them, the Opossum and Gallus seems to be two peculiar mammals, since the shortest distance between Opossum and the remaining mammals is more than 3371.80 and the shortest distance between Gallus and the remaining mammals is more than 3071.20. Obviously, the result is consistent with the fact that Opossum is the most remote specie from the remaining mammals and Gallus is not a kind of mammal.

And from Table 6, we can also obtain some similar pairs such as (Black rockcod, Neopagetopsis ionah) with the distance 788.16, (Cattle, Sheep) with the distance 712.94 and (Norway rat, House mouse) with the distance 770.12. Obviously, although there is a little errors in our experiments, but the basic conclusions are consistent with the results of the known fact of evolution.

The Phylogenetic Tree of the Protein Sequences

To demonstrate the performance of our new Similarities/Dissimilarities Analysis Model, in this section, we illustrate the phylogenetic trees obtained by our model and the phylogenetic trees obtained by utilizing the clutalW algorithm (Thompson *et al*., 1994) in the following Fig. 2.

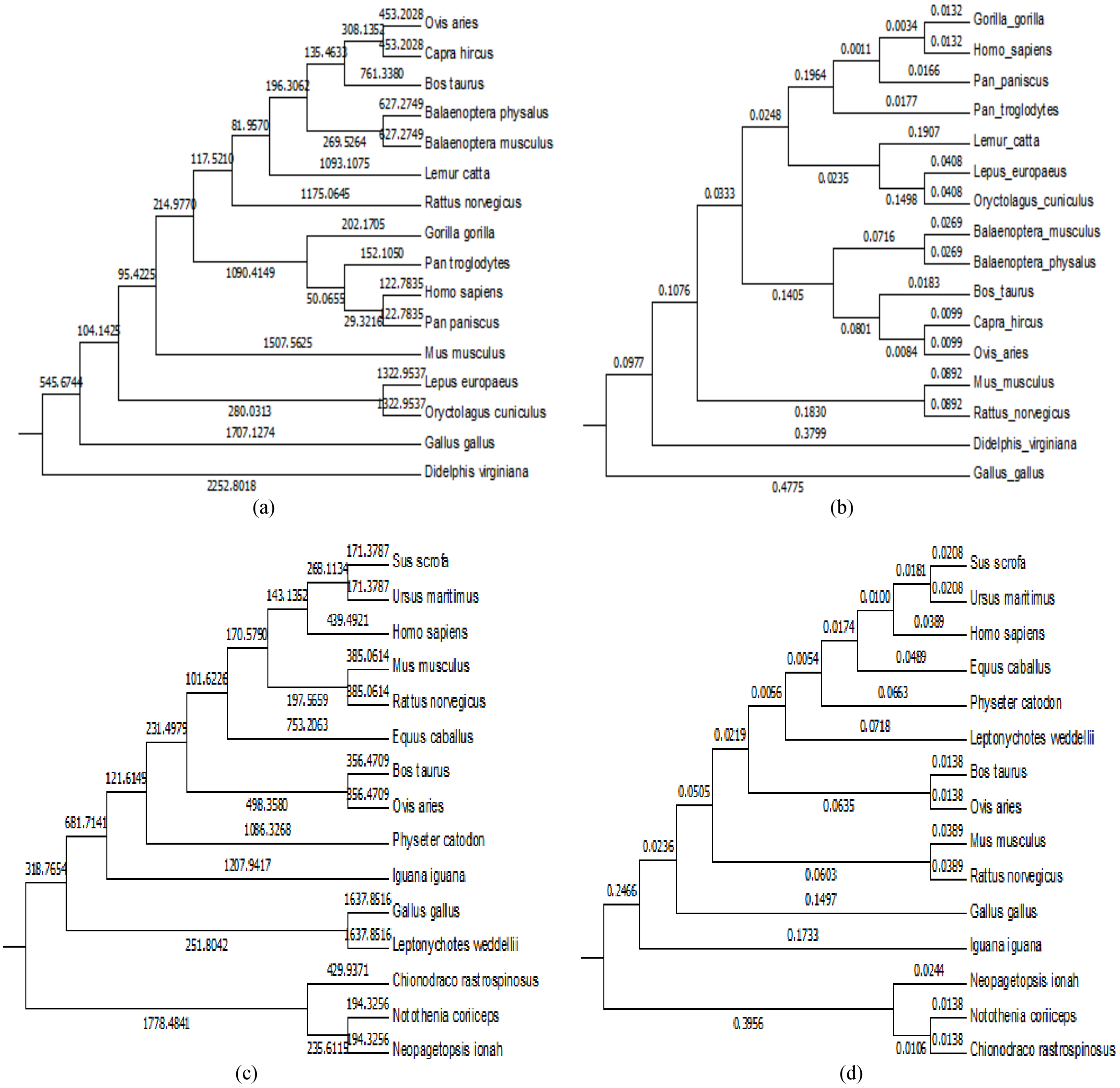


Fig. 2. Phylogenetic trees of ND6 and myglobin obtained by our Model and the clutalW algorithm (a) ND6 (Our Model) (b) ND6 (clustalW) (c) Myoglobin (Our Model) (d) Myoglobin (clustalW)

From Fig. 2, it is easy to know that in the phylogenetic trees of the 16 ND6 protein sequences and 15 myoglobin protein sequences obtained by our Model and the clutalW algorithm are almost the same. For example, in the phylogenetic tree of the 16 ND6 protein sequences obtained by our model, the Human, P_Chim, Gorilla and C_Chim are classified into a same category, the sheep, goat and cattle are classified into a same category, the fin_wha and blu_wha are classified into a same category and the rabbit and hare are classified into a same category also. Obviously, the results obtained by our model meet the reality overall except for the rat and mouse.

Similarly, in the phylogenetic tree of the 15 myoglobin protein sequences obtained by our model, the human, polar bear and pig are classified into a same category, the cattle and sheep are classified into a same category, the black rockcod, Neopagetopsis ionah and ocellated icefish are classified into a same category also, which are the same as that illustrated in the phylogenetic tree of the 15 myoglobin protein sequences obtained by the clutalW algorithm. Thus, we can make a conclusion that our method is correct and effective.

Conclusion

In this study, a new 2D graphical representation of protein sequence by mapping a protein sequence into curves based on the physicochemical and biological features of each amino acid first and then, a new similarities/dissimilarities analysis model for protein sequences is proposed based on the newly given 2D graphical representation of protein sequence, finally, on the basis of three well-known proteins sequence groups, simulation results show that our newly given method is correct and effective.

Acknowledgement

The author thanks the anonymous referees for suggestions that helped improve the paper substantially. And the project is partly sponsored by the Hunan Provincial Natural Science Foundation of China (No.14JJ2070).

Ethics

The experiments performed in accordance with the International Guiding Principles for Biomedical Research Involving Animals as promulgated by the Society for the Study of Reproduction.

References

Abo el Maaty, M.I., M.M. Abo-Elkhier and M.A. Abd Elwahaab, 2010. 3D graphical representation of protein sequences and their statistical characterization. Physica A: Stat. Mechan. Applic., 389: 4668-4676. DOI: 10.1016/j.physa.2010.06.031

Abo-Elkhier, M.M., 2012. Similarity/dissimilarity analysis of protein sequences using the spatial median as a descriptor. J. Biophys. Chem., 3: 142-148. DOI: 10.4236/jbpc.2012.32016

Bajzer, Z., M. Randic, D. Plasic and S.C. Basak, 2003. Novel map descriptors for characterization of toxic effects in proteomics maps. J. Mol. Graph. Model., 22: 1-9. DOI: 10.1016/S1093-3263(02)00186-9

Biggs, N., 1974. Algebraic Graph Theory. 2nd Edn., Cambridge University Press, Cambridge, ISBN-10: 052120335X, pp: 176.

Feng, Z.P. and C.T. Zhang, 2002. A graphic representation of protein sequence and predicting the subcellular locations of prokaryotic proteins. Int. J. Biochem. Cell Biol., 34: 298-307. DOI: 10.1016/S1357-2725(01)00121-2

Hea, P.A., D. Lia, Y. Zhangb, X. Wangc and Y. Yaod, 2012. A 3D graphical representation of protein sequences based on the Gray code. J. Theoretical Biol., 304: 81-87. DOI: 10.1016/j.jtbi.2012.03.023

Hessa, T., H. Kim, K. Bihlmaier, C. Lundin and J. Boekel *et al*., 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature, 433: 377-381. DOI: 10.1038/nature03216

Huang, L., H.L. Tan and B. Liao, 2013. HR-Curve: A novel 2D graphical representation of protein sequence and its multi-application. J. Comput. Theoretical Nanosci., 10: 257-264. DOI: 10.1166/jctn.2013.2688

Kyte, J. and R.F. Doolittle, 1982. A simple method for displaying the hydropathic character of a protein J. Mol. Biol. 157: 105-132. DOI: 10.1016/0022-2836(82)90515-0

Li, C. and J. Wang, 1966. Inequalities in quadratic forms. Duke Math. J., 33: 511-522. DOI: 10.1215/S0012-7094-66-03360-6

Qi, Z.H., J. Feng, X.Q. Qi and L. Li, 2012. Application of 2D graphic representation of protein sequence based on Huffman tree method. Comput. Biol. Med., 42: 556-563. DOI: 10.1016/j.compbiomed.2012. 01.011

Randic, M. and M. Vracko, 2003. On the similarity of DNA primary sequence. J. Chem. Inf. Comput. Sci., 40: 599-606. DOI: 10.1021/ci9901082

Randic, M., 2003. Graphical representations of DNA as 2-D map. Chem. Phys. Lett., 386: 468-471. DOI: 10.1016/j.cplett.2004.01.088

Randic, M., M. Vracko, A. Nandy and S.C. Basak, 2000. On 3-D graphical representation of DNA primary sequences and their numerical characterization. J. Chem. Inf. Comput. Sci., 40: 1235-1244. DOI: 10.1021/ci000034q

Randic, M., M. Vracko, J. Zupan and A.T. Balaban, 2004. Unique graphical representation of protein sequences based on nucleotide triplet codons. Chem. Phys. Lett., 397: 247-252. DOI: 10.1016/j.cplett.2004.08.118

Randic, M., M. Vracko, N. Lers and D. Plavšić, 2003a. Novel 2-D graphical representation of DNA sequences and their numerical characterization. J. Chem. Phys. Lett., 368: 1-6. DOI: 10.1016/S0009-2614(02)01784-0

Randic, M., M. Vracko, N. Lers and D. Plavšić, 2003b. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chem. Phys. Lett., 371: 202-207. DOI: 10.1016/S0009-2614(03)00244-6

Randic, M., M. Vracko, J. Zupan and M. Novič, 2003c. Compact 2-D graphical representation of DNA. Chem. Phys. Lett., 373: 558-562. DOI: 10.1016/S0009-2614(03)00639-0

Shrock, R. and S.H. Tsai, 1997. Upper and lower bounds for the ground state entropy of antiferromagnetic Potts models. Phys. Rev. E, 55: 6791-6791. DOI: 10.1103/PhysRevE.55.6791

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar, 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Molecular Biol. Evolut., 30: 2725-2429. DOI: 10.1093/molbev/mst197

Thompson, J.D., D.G. Higgins and T.J. Gibson, 1994. 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: 4673-4680. PMID: 7984417

Wang, L., H. Peng and J. Zheng, 2014. ADLD: A novel graphical representation of protein sequences and its application. Comput. Math. Meth. Med., 14: 959753-959767. DOI: 10.1155/2014/959753

Wang, L., P. Hui and Z. Jinhua, 2015. Similarities/dissimilarities analysis of protein sequences based on recurrence quantification analysis. Curr. Bioinform., 10: 112-119. DOI: 10.2174/157489361001150309144955

Wen, J. and Y.Y. Zhang, 2009. A 2D graphical representation of protein sequence and its numerical characterization. Chem. Phys. Lett., 476: 281-286. DOI: 10.1016/j.cplett.2009.06.017

Wimley, W.C. and S.H. White, 1996. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat. Struct. Biol., 3: 842-848. DOI: 10.1038/nsb1096-842

Yao, Y.H., Q. Dai, L. Li, X.Y. Nan and P.A. He *et al*., 2010. Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation. J. Comput. Chem., 31: 1045-1052. DOI: 10.1002/jcc.21391