Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Evolutionary Bioinformatics

Similarity/Dissimilarity Analysis of Protein Sequences Based on a New Spectrum-Like Graphical Representation

Yuhua Yao, Shoujiang Yan, Huimin Xu, Jianning Han, Xuying Nan, Ping-an He and Qi Dai

College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou, China.

ABSTRACT: Sequence comparison is one of the foundations in bioinformatics, which can be used to study evolutionary relations among the sequences. In this study, a 2D spectrum-like graphical representation of protein sequences is presented based on the hydrophobicity scale of amino acids. The frequencies of amplitudes of 4-subsequences are adopted to characterize a spectrum-like graph, and a 17D vector is used as the descriptor of protein sequence. The χ^2 value of compatibility test is performed. New similarity analysis approach is illustrated on the all protein sequences, which are encoded by the mitochondrion genome of 20 different species. Finally, comparison with the ClustalW method shows the utility of our method.

KEYWORDS: spectral representation, similarities/dissimilarities, protein sequences, compatibility test

CITATION: Yao et al. Similarity/Dissimilarity Analysis of Protein Sequences Based on a New Spectrum-Like Graphical Representation. *Evolutionary Bioinformatics* 2014:10 87–96 doi: 10.4137/EBO.S14713.

RECEIVED: February 10, 2014. **RESUBMITTED:** May 11, 2014. **ACCEPTED FOR PUBLICATION:** May 14, 2014.

ACADEMIC EDITOR: Jike Cui, Associate Editor

TYPE: Original Research

FUNDING: This study is supported by the National Natural Science Foundation of China under Grant Nos. 61272312, 61170316, and 61170110, and by the 521 Talent Cultivation Plan of Zhejiang Sci-Tech University.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest

COPYRIGHT: © the authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC 3.0 License.

CORRESPONDENCE: yaoyuhua2288@163.com

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants.

Introduction

Comparison of bio-sequences, such as DNA, RNA, and protein, is the origin of bioinformatics. Through the comparison, we can identify the similarity/dissimilarity of different species' sequences. Many methods of technologies have been introduced like graphical representation of DNA/RNA and so on. Based on graphical representations, numerical characterization techniques offer a route toward quantitatively estimating the similarities/dissimilarities of sequences.[1–13] The reason for the delay in the emergence of graphical representations of proteins is the increased complexity of biological strings built on a 20-letter alphabet (representing the 20 natural amino acids) in comparison with strings built from only four letters (representing DNA or RNA). According to the genetic code, Randić et al. and Bai and Wang[14–17] gave some graphical representations and the sequence descriptors of proteins. Similar to existing graphical representation of DNA, in order to better compare the similarities/dissimilarities of proteins, we modified some graphical representations of proteins.[18–21] With some physicochemical properties of 20 amino acids, the graphical representations of protein sequence have been introduced.[22–26]

Measuring the similarity between categorical sequences is a fundamental process in many data mining applications. A key issue is extracting and making use of significant features hidden behind the chronological and structural dependencies found in these sequences. This measure can lead to a better understanding of the nature of these sequences. The most important known challenges presented by these data, which are only partially addressed by existing methods, are the following: (1) it is difficult to extract the information underlying the chronological dependencies of structural features which may have significant meaning; (2) the high computational cost involved is also an important problem; (3) this creates ambiguities and complications for the similarity

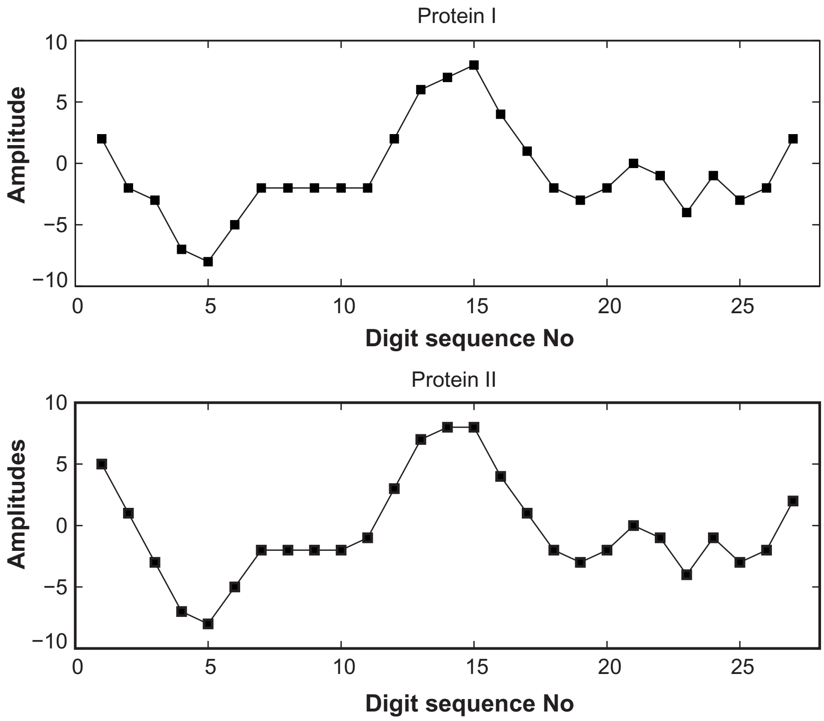

Figure 1. The spectrum-like graphs of two protein fragments I and II of yeast *Saccharomyces cerevisiae*, having 30 amino acids.

measurement task, especially for sequences of significantly different lengths.

In this study, we outline a novel spectrum-like graphical representation, which is based on the hydrophobicity property of amino acids, and introduce a novel strategy for sequence comparison according to the method of calculating the frequencies of all amplitudes of different species' spectral graphs. We will make a comparison for all protein sequences in the mitochondria of 20 species.

Table 1. The digit sequence (d_i) and 4-subsequence (y_i) of the protein I.

i	seq	d_i	y_i	i	seq	d_i	y_i
1	W	2	2	16	L	2	4
2	T	–1	–2	17	W	2	1
3	F	2	–3	18	L	2	–2
4	E	–1	–7	19	N	–2	–3
5	S	–2	–8	20	G	–1	–2
6	R	–2	–5	21	G	–1	0
7	N	–2	–2	22	P	1	–1
8	D	–2	–2	23	G	–1	–4
9	P	1	–2	24	C	1	–1
10	A	1	–2	25	S	–2	–3
11	K	–2	–2	26	S	–2	–2
12	D	–2	2	27	F	2	2
13	P	1	6	28	T	–1	–
14	V	1	7	29	G	–1	–
15	I	2	8	30	L	2	–

Methods

Here we consider a physicochemical property which has important relations with the structure of proteins: hydrophobicity of amino acids. The distribution of hydrophobic amino acids in the primary sequences can be used as an indicator to predict the secondary structure of protein elements.[27] In the following contents, we will construct the spectrum-like graphical representation of protein sequences.

First, each amino acid is characterized by its own physicochemical properties. Twenty amino acids are simplified into two types[28]: hydrophobic amino acids H = {F, L, I, Y, M, W, V, A, P, C}; hydrophilic amino acids P = {S, N, K, D, R, T, H, Q, E, G}. Then twenty amino acids are further simplified into four types[29]: strong hydrophobic amino acids SH = {F, L, I, Y, W}; weak hydrophobic amino acids WH = {M, V, A, P, C}; strong hydrophilic amino acids SP = {S, N, K, D, R}; weak hydrophilic amino acids WP = {T, H, Q, E, G}.

Thus, giving a protein sequence $S = s_1 s_2 \ldots s_N$ with N amino acids, we inspect it by stepping one amino acid at a time. For example, at the step $i (i = 1, 2, \ldots, N)$, S_i is transformed into d_i which may be 2, 1, −1, and −2. Then the digit sequence $D = d_1 d_2 \ldots d_N$ is obtained. In order to more clearly display the differences between hydrophobic amino acids and hydrophilic amino acids, during the construction of the digit sequence, we preset the value of properties:

$$d_i = \begin{cases} 2 & \text{if } s_i \in \mathrm{SH} = \{F, L, I, Y, W\}; \\ 1 & \text{if } s_i \in \mathrm{WH} = \{M, V, A, P, C\}; \\ -1 & \text{if } s_i \in \mathrm{WP} = \{T, H, Q, E, G\}; \\ -2 & \text{if } s_i \in \mathrm{SP} = \{S, N, K, D, R\}. \end{cases} \quad i = 1, 2, \ldots, N.$$

It is sometimes instructive to represent a random walk as a polygonal line, or path, in the plane, where the horizontal axis represents time and the vertical axis represents the value of $\{S_n\}$. Giving a sequence $\{S_n\}$ of partial sums, we first plot the points (n, S_n), and then for each $k < n$, we connect (k, S_k) and $(k+1, S_{k+1})$ with a straight line segment. The length of a path is just the difference in the time values of the beginning and ending points on the path. So, d_j, d_{j+1}, d_{j+2}, d_{j+3}, four consecutive numbers are summed as the partial sums and the summations are the values of vertical axis and are considered as the amplitudes. When i is the value of horizontal axis and runs from 1 to N–3, we have the points $P_1(x_1, y_1)$, $P_2(x_2, y_2)$,…, $P_{N-3}(x_{N-3}, y_{N-3})$. Among them, x_i and y_i are calculated by the following formula:

$$\begin{cases} x_i = i \\ y_i = \sum_{j=i}^{i+3} d_j \end{cases} \quad i = 1, 2, \cdots, N-3.$$

Connecting adjacent points, we obtain a spectrum-like graph of protein sequence.

We will illustrate the current approach on two shorter segments of yeast protein *Saccharomyces cerevisiae*. Figure 1 shows the two spectral graphs, and the corresponding proteins are

Protein I: WTFESRNDPAKDPVILWLNGGPGC-SSLTGL;
Protein II: WFFESRNDPANDPIILWLNGGPGC-SSFTGL.

The digit sequence (d_i) and 4-subsequence (y_i) of the protein I are showed in Table 1.

Observing Figure 1, we know that the two curves are similar on the whole and have several same local sequences' segments. In this method, the reason why we emphasize the same hydrophilic—hydrophobic amino acids is that they are more likely to form a similar or identical structure.

In Figure 2, we apply the new spectral representation to the ND6 (NADH dehydrogenase subunit 6) proteins of nine species, human, gorilla, common chimpanzee, pigmy chimpanzee, blue whale, fin whale, rat, mouse, and opossum. Taking a closer look at Figure 2 and comparing the curves, we find that the curves of the ND6 proteins of human, gorilla, P. chimpanzee, and C. chimpanzee are more similar. Also, the

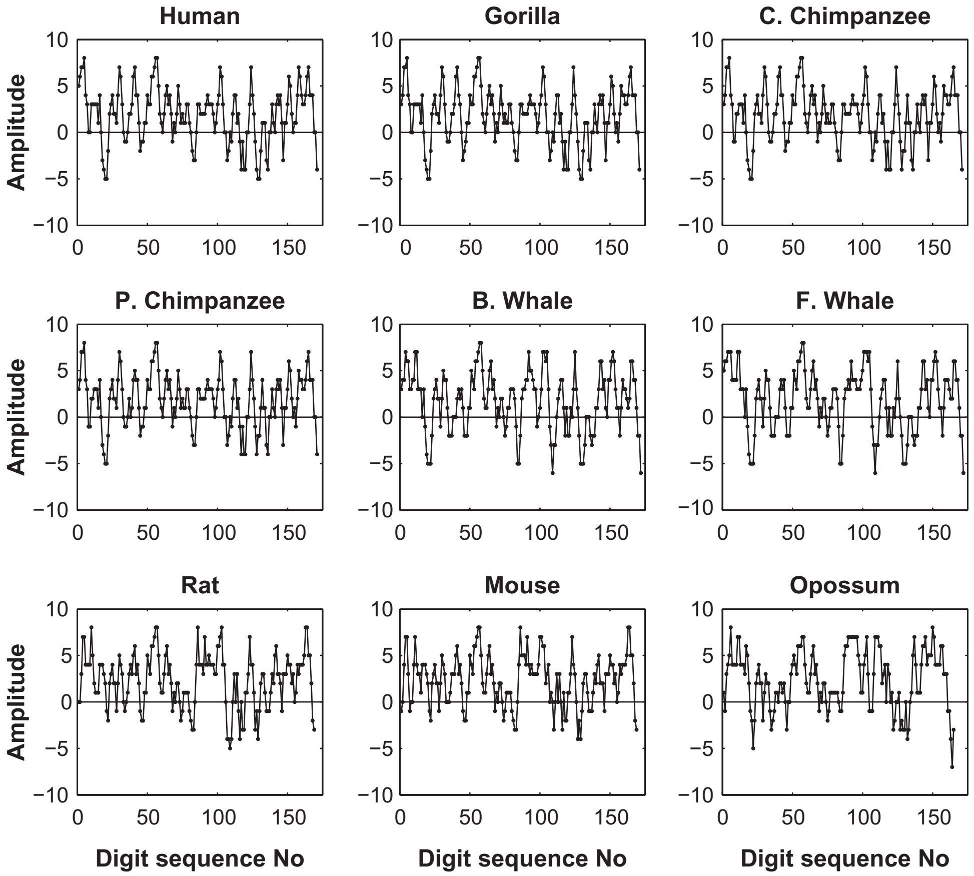

Figure 2. The spectrum-like graphs of the ND6 proteins of nine eutherian species include those for human, gorilla, common chimpanzee, pigmy chimpanzee, blue whale, fin whale, rat, mouse, and opossum.

Table 2. The Information for all protein sequences in the mitochondria of 9 species.

	HUMAN	GORILLA	P. CHIMP	C. CHIMP	F. WHALE	B. WHALE	RAT	MOUSE	OPOSSUM
ND1	CAA24026 (318)	BAA85277 (318)	BAA85294 (318)	BAA85268 (318)	CAA43444 (318)	CAA50995 (318)	CAA32954 (318)	CAA24080 (315)	CAA82677 (318)
ND2	CAA24027 (347)	BAA85278 (347)	BAA85295 (347)	BAA85269 (347)	CAA43445 (347)	CAA50996 (347)	CAA32955 (345)	CAA24081 (345)	CAA82678 (347)
COI	CAA24028 (513)	BAA85279 (513)	BAA85296 (513)	BAA85270 (513)	CAA43451 (516)	CAA50997 (516)	CAA32956 (514)	CAA24082 (514)	CAA82679 (513)
COII	CAA24029 (227)	BAA07303 (227)	BAA07312 (227)	BAA07299 (227)	CAA43452 (227)	CAA50998 (227)	CAA32957 (227)	CAA24083 (227)	CAA82680 (235)
ATP8	CAA24030 (68)	BAA07304 (68)	BAA07313 (68)	BAA07300 (68)	CAA43441 (63)	CAA50999 (63)	CAA32958 (67)	CAA24084 (67)	CAA82681 (69)
ATP6	CAA24031 (226)	BAA85280 (226)	BAA85297 (226)	BAA85271 (226)	CAA43442 (226)	CAA51000 (226)	CAA32959 (226)	CAA24085 (226)	CAA82682 (226)
COIII	CAA24032 (261)	BAA85281 (261)	BAA85298 (261)	BAA85272 (261)	CAA43453 (261)	CAA51001 (261)	CAA32960 (261)	CAA24090 (278)	CAA82683 (281)
ND3	CAA24033 (115)	BAA85282 (115)	BAA85299 (115)	BAA85273 (115)	CAA43446 (115)	CAA51002 (115)	CAA32961 (115)	CAA24086 (114)	CAA82684 (116)
ND4L	CAA24034 (98)	BAA07305 (98)	BAA07314 (98)	BAA07301 (98)	CAA43447 (98)	CAA51003 (98)	CAA32962 (98)	CAA24087 (97)	CAA82685 (98)
ND4	CAA24035 (459)	BAA85283 (459)	BAA85300 (459)	BAA85274 (459)	CAA43448 (459)	CAA51004 (459)	CAA32963 (459)	CAA24091 (474)	CAA82686 (474)
ND5	CAA24036 (603)	BAA07306 (603)	BAA07315 (603)	BAA07302 (603)	CAA43449 (606)	CAA51005 (606)	CAA32964 (610)	CAA24088 (607)	CAA82687 (602)
ND6	CAA24037 (174)	BAA07307 (174)	BAA85301 (174)	BAA85275 (174)	CAA43450 (175)	CAA51006 (175)	CAA32965 (172)	CAA24089 (172)	CAA82688 (168)
CYTB	CAA24038 (380)	BAA85284 (380)	BAA85302 (380)	BAA85276 (380)	CAA43443 (379)	CAA51007 (379)	CAA32966 (380)	CAA24092 (392)	CAA82689 (382)
Total length	3789	3789	3789	3789	3790	3790	3792	3728	3729

ND6 protein graphs are more similar for F. whale, B. whale and rat, mouse too. In addition, we find ND6 protein of opossum is obviously different from the other species. Also their similarities/dissimilarities are consistent with the known fact of evolution.

Unexpectedly, we find that most amplitudes of amino acid are greater than 0, which may mean that amino acids' preferences are hydrophobic in the protein sequence according to the four classifications of amino acids. It is probably because hydrophobic amino acids have an important influence on protein structures.

Results/Discussion

Once we have a matrix to represent a sequence, numerous matrix invariants[25,26,30–33] are used as descriptor of sequences. However, the computational complexity of these matrix invariants techniques is at least $O(N^2)$, which results in the main difficulty in computation. In this section, we overcome the difficulty and introduce a novel way to numerically characterize protein sequence and it is easy to implement. Their computational complexities are reduced to $O(N)$, so it is easy to implement. In addition, the new sequence descriptor is linearly relative to the length of the sequences, so it is appropriate for sequences of significantly different lengths.

When we construct the spectrum-like graph, we calculate the summation of four consecutive numbers of a digit sequence. The summations are considered as the amplitudes, which can be −8, −7, −6, −5, −4, −3, −2, −1, 0, 1, 2, 3, 4, 5, 6, 7, and 8. In order to obtain the numerical representation of protein sequences, we calculate the frequency of amplitude. Therefore, a protein sequence can be characterized by a 17D vector.

The data set consists of 13 proteins (cytochrome oxidase subunits I, II, and III; cytochrome *b* apoenzyme; NADH dehydrogenase subunits 1–6 and 4L; ATP synthase subunits 6 and 8) encoded by the typical mitochondrial genome from mammalian species.

The information of the 13 proteins is listed in Table 2. The 13 proteins are concatenated into one long amino acid sequence and analyzed as one protein sequence. Their frequencies of amplitudes are obtained and listed in Table 3. According to the results obtained in Table 3, we construct 17-component vectors of the spectral graphs corresponding to 9 species proteins, and then the 17-component vectors are first normalized. For a vector X, normalization means: $Z = (X–\mathrm{Mean}(X))/\mathrm{Std}(X)$, $\mathrm{Mean}(X)$, means the mean of X and $\mathrm{Std}(X)$ is the standard deviation of X. In Table 4, the similarity/dissimilarity matrices for the nine species protein sequences are given, which are based on the

Table 3. The frequencies of amplitudes of spectral graphs for all proteins sequences in the mitochondrion of 9 different species.

$f(y_i)$	HUMAN	GORILLA	P. CHIMP	C. CHIMP	F. WHALE	B. WHALE	RAT	MOUSE	OPOSSUM
$f(-8)$	0.0018	0.0018	0.0021	0.0021	0.0008	0.0011	0.0018	0.0032	0.0013
$f(-7)$	0.0040	0.0058	0.0040	0.0037	0.0063	0.0063	0.0071	0.0048	0.0084
$f(-6)$	0.0114	0.0100	0.0129	0.0129	0.0111	0.0106	0.0108	0.0100	0.0118
$f(-5)$	0.0145	0.0143	0.0148	0.0148	0.0177	0.0158	0.0161	0.0177	0.0154
$f(-4)$	0.0341	0.0365	0.0359	0.0351	0.0372	0.0378	0.0430	0.0412	0.0363
$f(-3)$	0.0520	0.0541	0.0534	0.0541	0.0494	0.0470	0.0546	0.0508	0.0523
$f(-2)$	0.0634	0.0565	0.0576	0.0571	0.0562	0.0576	0.0504	0.0526	0.0476
$f(-1)$	0.0726	0.0695	0.0716	0.0716	0.0673	0.0689	0.0697	0.0717	0.0753
$f(0)$	0.1062	0.1096	0.1091	0.1094	0.1128	0.1122	0.1124	0.1124	0.1163
$f(1)$	0.1220	0.1249	0.1236	0.1223	0.1180	0.1215	0.1196	0.1145	0.1147
$f(2)$	0.0914	0.0890	0.0919	0.0909	0.0877	0.0866	0.0823	0.0814	0.0826
$f(3)$	0.1112	0.1059	0.1072	0.1091	0.1109	0.1106	0.1174	0.1174	0.1181
$f(4)$	0.1233	0.1244	0.1244	0.1263	0.1252	0.1231	0.1306	0.1293	0.1283
$f(5)$	0.0713	0.0737	0.0737	0.0737	0.0700	0.0721	0.0678	0.0701	0.0646
$f(6)$	0.0465	0.0465	0.0444	0.0433	0.0544	0.0560	0.0454	0.0500	0.0481
$f(7)$	0.0523	0.0534	0.0515	0.0515	0.0523	0.0523	0.0483	0.0497	0.0562
$f(8)$	0.0219	0.0240	0.0219	0.0222	0.0227	0.0206	0.0224	0.0233	0.0227

Table 4. The similarity matrix of 9 species based on the frequencies of amplitudes.

SPECIES	GORILLA	P. CHIMPAN	C. CHIMPAN.	F. WHALE.	B. WHALE	RAT	MOUSE	OPOSSUM
Human	**4.2144**	**2.7639**	**3.0017**	5.5206	5.1463	6.9385	7.1704	7.4932
Gorilla		**4.0165**	**4.0790**	5.1489	5.7607	6.1994	6.9165	7.2921
P. Chimpan.			**1.0975**	5.6356	5.5562	6.4450	7.1025	7.5040
C. Chimpan.				5.6890	5.9505	6.0357	6.7764	7.1315
F. Whale					**3.2861**	5.5199	5.5947	6.0795
B. Whale						6.5392	6.6137	7.0378
Rat							**4.0634**	5.6101
Mouse								6.1929

Euclidean distances between the 17-component vectors normalized. We give two arbitrary sequences S^1 and S^2. In our approach, the Euclidian distance $D(S^1, S^2)$ between the two vectors is

$$D(S^1,S^2)=\|\vec{v}(S^1)-\vec{v}(S^2)\|_2.$$

The analysis of similarities/dissimilarities represented by the index of similarity/dissimilarity is based on the following assumption: the smaller the distance between two proteins is, the more the two proteins will be similar. We know that the smaller the index of similarity/dissimilarity is, the more similar the two proteins will be. The indexes of similarity/dissimilarity between the nine species are listed in Table 4.

Table 5. The theory values of frequency of the amplitudes.

y_i	SPLIT	COMBINATORIAL NUMBER	THE THEORETICAL FREQUENCY
−8	{−2, −2, −2, −2}	$C_4^4=1$	1/256 ≈ 0.00391
−7	{−2, −2, −2, −1}	$C_4^3=4$	4/256 ≈ 0.01563
−6	{−2, −2, −1, −1}	$C_4^2=6$	6/256 ≈ 0.02344

−5	{−2, −2, −2, 1}
	{−2, −1, −1, −1}
		$C_4^3=4$
		$C_4^1=4$
			8/256 ≈ 0.03125

−4	{−2, −2, −2, 2}
	{−2, −2, −1, −1}
	{−1, −1, −1, −1}
		$C_4^1=4$
		$C_4^1{*}C_3^1=12$
		$C_4^4=1$
			17/256 ≈ 0.06641

−3	{−2, −2, −1, 2}
	{−2, −1, −1, 1}
		$C_4^1{*}C_3^1=12$
		$C_4^1{*}C_3^1=12$
			24/256 ≈ 0.09375

−2	{−2, −2, 1, 1}
	{−2, −1, −1, 2}
	{−1, −1, −1, 1}
		$C_4^2=6$
		$C_4^1{*}C_3^1=12$
		$C_4^1=4$
			22/256 ≈ 0.08594

−1	{−2, −2, 1, 2}
	{−2, −1, 1, 1}
	{−1, −1, −1, 2}
		$C_4^1{*}C_3^1=12$
		$C_4^1{*}C_3^1=12$
		$C_4^1=4$
			28/256 ≈ 0.10938

0	{−2, −2, 2, 2}
	{−2, −1, 1, 2}
	{−1, −1, 1, 1}
		$C_4^2=6$
		$C_4^1{*}C_3^1{*}C_2^1=24$
		$C_4^2=6$
			36/256 ≈ 0.14063

Observing Table 4, we can find that the smaller entries are associated with the pairs in group human, gorilla, P. chimpanzee, and C. chimpanzee; F. whale, B. whale; and rat, mouse. On the other hand, the larger entries in the similarity/dissimilarity matrix appear in the rows belonging to opossum. These results are consistent with the known conclusion of evolution.[12,25]

We calculate the theory values of frequency for the amplitudes which are listed in Table 5. As the theory values are symmetrical, we only show one half. We intend to know whether the frequencies of amplitudes for the 13 proteins in the 9 species are consistent with the ratios of theory values. In Figure 6, we show the comparison charts of 13 proteins of human and the theory values. Then, we calculate the χ^2 values:

$$\chi^2=\sum_{i=-8}^{8}\frac{(O_i-E_i)^2}{E_i}=\sum_{i=-8}^{8}\frac{(n_i-np_i)^2}{np_i}.$$

The χ^2 values of 13 proteins for 9 species are listed in Table 6. Each protein corresponding to one 17-component vector, so all the degrees of freedom are df = 17 − 1 = 16. Significance level is α = 0.01. $\chi^2_{0.01}(16)=32.00$. Nearly all χ2 values are more than $\chi^2_{0.01}=32.00$ in Table 6, so they are not consistent with the ratios of theory values. The amino acid sequences of proteins determine the protein structure and function. So their patterns are not expected to be random.

Firstly, we will make a comparison for helicase protein sequences of 12 baculoviruses, including 3 group I alphabaculovirus: AcMNPV, BmNPV, RoMNPV; 6 group II alphabaculovirus: HearNPV, HzSNPV, MacoNPVA, MacoNPVB, HaSNPV, AgseNPV; 3 betabaculovirus: AdorGV, CpGV, CrleGV. Length and group information of these protein sequences are shown in Table 7. The phylogenetic tree of 12 helicase protein sequences is given in Figure 3. Their

Table 6. The χ^2 values for 13 proteins of 9 species.

SPECIES	ND1	ND2	COI	COII	ATP8	ATP6,	COIII	ND3	ND4L	ND4	ND5	ND6	CYTB
Human	191.16	124.23	209.09	103.63	20.98	179.72	85.36	247.70	73.39	214.32	158.50	80.81	229.47
Gorilla	208.26	130.95	208.19	117.59	30.21	177.31	94.77	197.36	80.11	229.64	220.32	74.96	250.04
P. Chimpan.	177.87	107.64	205.79	107.14	17.95	187.76	88.52	177.68	65.03	222.80	189.38	70.46	278.48
C. Chimpan.	189.97	124.54	208.68	107.14	17.47	200.75	86.82	173.44	75.55	232.19	176.92	75.50	306.85
F. Whale	170.14	139.65	195.96	45.17	45.68	128.02	80.64	154.67	75.31	333.77	187.34	114.51	331.96
B. Whale	171.69	122.01	204.06	47.08	44.93	117.65	82.78	154.67	87.32	322.66	191.10	93.92	239.08
Rat	190.92	114.59	175.99	31.61	41.38	126.04	101.83	120.74	26.18	268.47	170.42	153.39	312.43
Mouse	183.63	207.92	178.29	30.79	53.50	126.70	80.15	142.66	27.57	234.89	199.80	117.80	301.85
Opossum	232.56	107.19	220.40	64.89	31.87	135.80	116.69	189.43	76.05	201.68	207.93	169.48	212.37

similarities/dissimilarities are consistent with classification of these baculovirus proteins.[34–36]

To further verify the validity of our approach, we have done an experiment on a dataset of the 13 proteins encoded by the same strand of the mitochondrial genome from 20 eutherian species: human (*Homo sapiens*), C. chimpanzee (*Pan troglodytes*), P. chimpanzee (*Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), gibbon (*Hylobates lar*), baboon (*Papio hamadryas*), horse (*Equus caballus*), white rhinoceros (*Ceratotherium simum*), harbor seal (*Phoca vitulina*), gray seal (*Halichoerus grypus*), cat (*Felis catus*), F. whale (*Balaenoptera physalus*), B. whale (*Balaenoptera musculus*), cow (*Bos taurus*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), opossum (*Didelphis virginiana*), wallaroo (*Macropus robustus*), and platypus (*Ornithorhynchus anatinus*). Note that we have kept rodent species to murids only and marsupials and monotremes are being used as out-group. The phylogenetic tree of 20 species is given in Figure 4. We also construct a phylogenetic tree by the ClustalW method.[37] The result is shown in Figure 5.

Comparing Figures 4 and 5, we can find that: (1) they all distinguish the marsupials and monotremes, rodents, ferungulates, and primates; (2) it has been debated which two of the three main groups of placental mammals are closely related: primates, ferungulates, and rodents. Figure 4 supports the suggestion that primates and ferungulates are more closely related, whereas Figure 5 shows that primates and rodents are more closely related; (3) in Figure 5, opossum, wallaroo, and platypus as the out-group, was nearly clustered to rodents. The result of Figure 4 is consistent with the known conclusion of evolution and others' partial results[38,39] except for the opossum, so our method is more advantageous in this regard.

To show the efficiency of the proposed approach, based on different protein families, we further make a comparison with the widely used methods, EMBOSS water – pairwise sequence alignment. Then, we test some families by the two methods, including 13 protein families encoded by the same strand of the mitochondrial genome, UDP glucuronosyltransferase family proteins (including the same genus but different species), and so on. The test results show that the similarity distances or scores by different methods are almost in an agreement with each other. Furthermore, for longer protein sequences the test results by the two methods are more consistent.

Table 7. Length and group information of helicase protein sequences of 12 baculovirus.

GENUS (GROUP)	VIRUS NAME	ABBREVIATION	ACCESSION NO.	LENGTH
Alphabaculovirus (Group I NPVs)	*Autographa californica* MNPV	AcMNPV	AAA66725	1221
	Bombyx mori NPV	BmNPV	AAC63764	1222
	Rachiplusia ou MNPV	RoMNPV	AAN28013	1221
Alphabaculovirus (Group II NPVs)	*Helicoverpa armigera* NPV	HearNPV	AEN04007	1253
	Helicoverpa zea SNPV	HzSNPV	AAL56093	1253
	Mamestra configurata NPVA	MacoNPVA	AAM09201	1212
	Mamestra configurata NPVB	MacoNPVB	AAM95079	1209
	Helicoverpa armigera SNPV	HaSNPV	AAG53827	1253
	Agrotis segetum NPV	AgseNPV	AAZ38246	1213
Betabaculovirus (GVs)	*Adoxophyles orona* GV	AdorGV	AAP85713	1138
	Cydia pomonella GV	CpGV	AAK70750	1131
	Cryptophlebia leucotreta GV	CrleGV	AAQ21676	1128

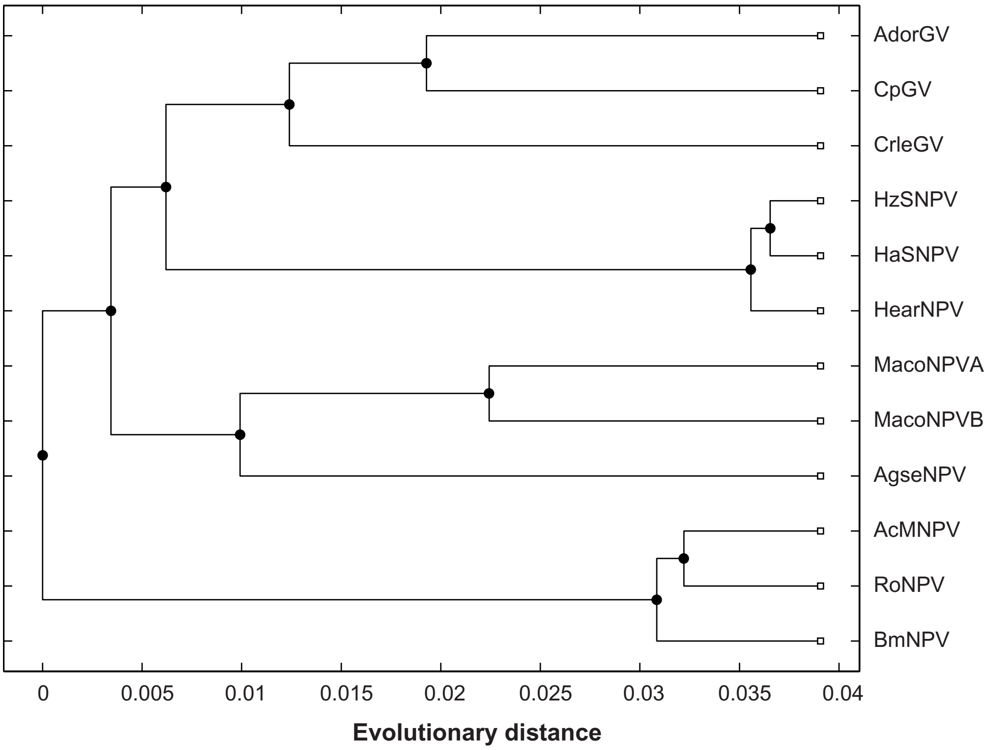

Figure 3. The phylogenetic tree based on protein sequences of 12 baculoviruses. Sequences include those for AcMNPV, BmNPV, RoMNPV, HearNPV, HzSNPV, MacoNPVA, MacoNPVB, HaSNPV, AgseNPV, AdorGV, CpGV, and CrleGV.

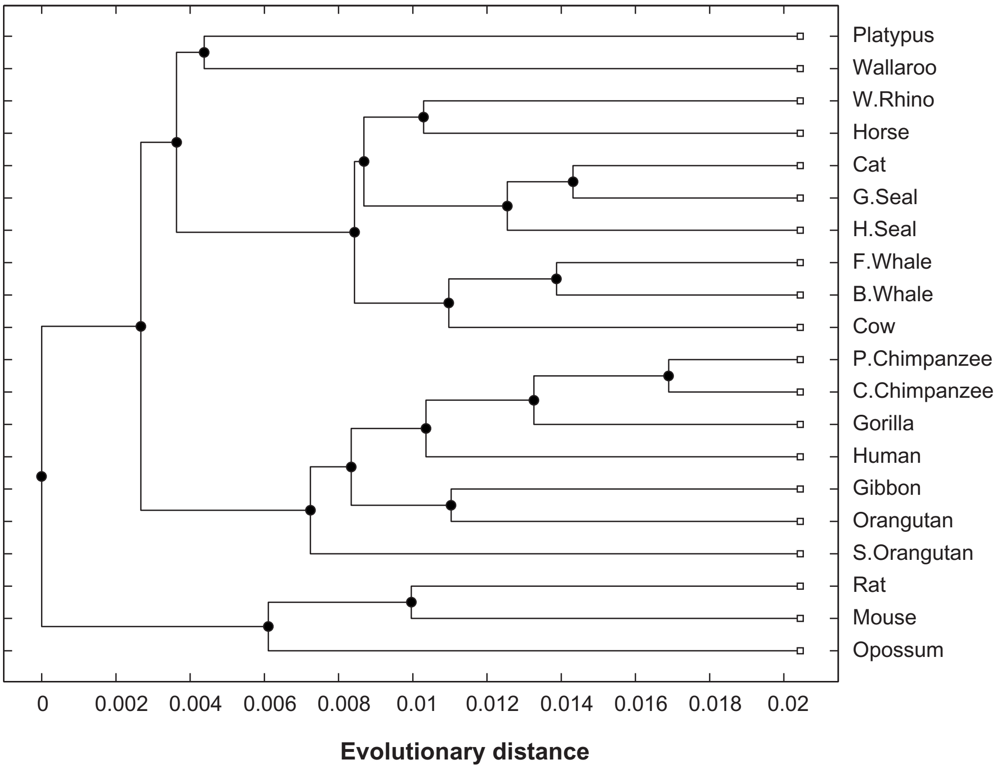

Figure 4. The phylogenetic tree of 20 eutherian species based on our method. Phylogeny was based on analysis of the combined sequences of 13 proteins encoded by the same strand of the mitochondrial genome. Sequences include those for human, common chimpanzee, pigmy chimpanzee, gorilla, orangutan, gibbon, baboon, horse, white rhinoceros, harbor seal, gray seal, cat, fin whale, blue whale, cow, rat, mouse, opossum, wallaroo, and platypus. The sequences of opossum, wallaroo, and platypus were used as out-group.

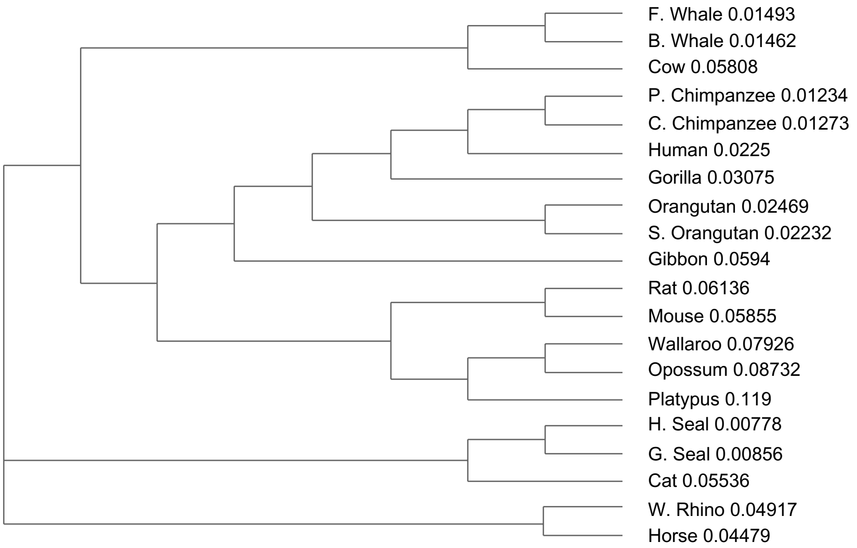

Figure 5. The phylogenetic tree of 20 eutherian species based on ClustalW. Phylogeny was based on analysis of the combined sequences of 13 proteins encoded by the same strand of the mitochondrial genome. Sequences include those for human, common chimpanzee, pigmy chimpanzee, gorilla, orangutan, gibbon, baboon, horse, white rhinoceros, harbor seal, gray seal, cat, fin whale, blue whale, cow, rat, mouse, opossum, wallaroo, and platypus. The sequences of opossum, wallaroo, and platypus were used as out-group.

Conclusions

The graphical techniques of biological sequences have been used as a very powerful tool for the visualization and analysis of protein sequences. Based on the hydrophobicity of amino acids, a new spectral representation of proteins is introduced, in this study.

1. We present a spectrum-like graphical representation of protein sequences, which are based on a significant physicochemical property. The chemical or physical property of amino acids will also be useful to study and solve some bioinformatics problems. The advantage of our approach is that it allows visual inspection of data, which helps

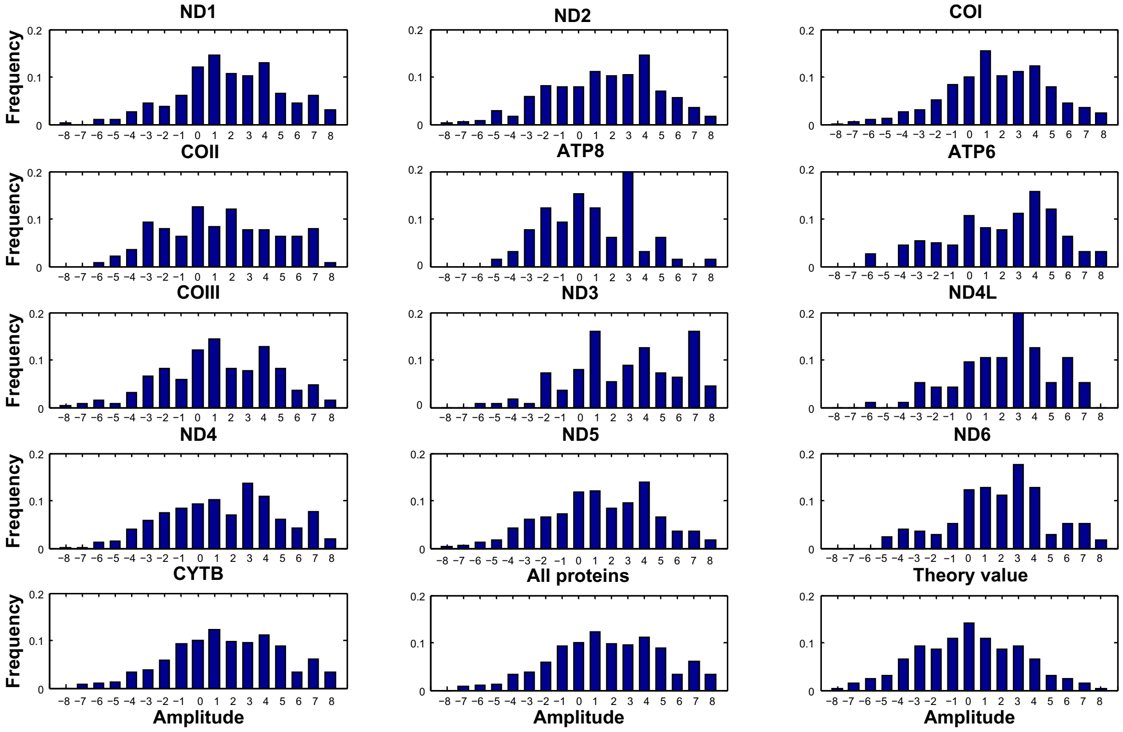

Figure 6. The distributions of amplitudes of 13 proteins of human and the theory value. Proteins include those for cytochrome oxidase subunits I, II, and III (COI, COII, and COIII); cytochrome *b* apoenzyme (CYTB); NADH dehydrogenase subunits 1–6 and 4 L (ND1, ND2, ND3, ND4, ND5, ND6, and ND4L); ATP synthase subunits 6 and 8 (ATP6 and ATP8).

recognize major similarities among different proteins, and even protein structures.

2. For long protein sequences, the frequencies are easily computed and can be used to numerically characterize protein sequences, and the examination of similarity/dissimilarity illustrates the utility of the approach. The computational complexity of alignment method and matrix invariant technique is at least O(N^2). Our method does not require multiple sequence alignments and greatly reduces the computational complexity at the same time.
3. Our approach also gives novel numerical characterization of proteins. One is based on the frequencies of amplitudes of spectral graphs and the other is based on the χ^2, which are used to analyze the similarity of protein sequences. Also, both computational scientists and molecular biologists can use them to analyze protein sequences efficiently.
4. Theory values of frequencies of amplitudes are calculated. The results of the compatibility test show that the distribution of hydrophilic—hydrophobic amino acids may have special biological significance. To a certain degree, our method can extract the information underlying the chronological dependencies of structural features and is successfully applied to sequences comprising similar structural features in chronologically different positions. Also, the other physicochemical properties of amino acids will also be useful to study and solve some bioinformatics problems.

Acknowledgment

We acknowledge the constructive comments of the editor and two anonymous reviewers.

Author Contributions

YY conceived the method and prepared the manuscript. YY, SY, JH, and HX implemented the software and performed the analysis. YY, JH, and HX contributed to writing the paper. XN, PAH, and QD contributed to the discussion and approved the final version. All authors reviewed and approved the final manuscript.

REFERENCES

1. Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J Biol Chem*. 1983;258(2):1318–27.
2. Hamori E. Novel DNA sequence representations. *Nature*. 1985;314(6012):585–6.
3. Gates MA. Simpler DNA sequence representations. *Nature*. 1985;316(6025):219.
4. Nandy A. A new graphical representation and analysis of DNA sequence structure. I methodology and application to globin genes. *Curr Sci*. 1994;66:309–14.
5. Jeffrey HJ. Chaos game representation of gene structure. *Nucleic Acids Res*. 1990;18(8):2163–70.
6. Randić M, Vračko M, Lerš N, Plavšić D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem Phys Lett*. 2003;368(1–2):1–6.
7. Liao B, Wang TM. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. *J Chem Inf Comput Sci*. 2004;44(5):1666–70.
8. Randić M, Zupan J, Balaban AT. Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem Phys Lett*. 2004;397(1–3):247–52.
9. Chi R, Ding K. Novel 4D numerical representation of DNA sequences. *Chem Phys Lett*. 2005;407(1–3):63–7.
10. Yao YH, Nan XY, Wang TM. Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation. *Chem Phys Lett*. 2005;411(1–3):248–55.
11. Yao YH, Nan XY, Wang TM. A new 2D graphical representation—classification curve and the analysis of similarity/dissimilarity of DNA sequences. *J Mol Struct*. 2006;764(1–3):101–8.
12. Yao YH, Dai Q, Nan XY, et al. Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation. *J Comput Chem*. 2008;29(10):1632–9.
13. Bielinska-Waz D. Four-component spectral representation of DNA sequences. *J Math Chem*. 2010;47(1):41–51.
14. Randić M. 2-D graphical representation of proteins based on virtual genetic code. *SAR QSAR Environ Res*. 2004;15(3):147–57.
15. Randić M, Butina D, Novič M, Založnik A, Pisanski T. A novel graphical representation of proteins. *Period Biol*. 2005;107:403–14.
16. Bai F, Wang T. On graphical and numerical representation of protein sequences. *J Biomol Struct Dyn*. 2006;23(5):537–46.
17. Randić M, Mehulic K, Vukicevic D, Pisanski T, Vikic-Topic D, Plavsic D. Graphical representation of proteins as four-color maps and their numerical characterization. *J Mol Graph Model*. 2009;27(5):637–41.
18. Randić M, Vracko M, Novic M, Plavsic D. Spectral representation of reduced protein models. *SAR QSAR Environ Res*. 2009;20(5–6):415–27.
19. He PA. A new graphical representation of similarity/dissimilarity studies of protein sequences. *SAR QSAR Environ Res*. 2010;21(5–6):571–80.
20. Liao B, Sun X, Zeng Q. A novel method for similarity analysis and protein subcellular localization prediction. *Bioinformatics*. 2010;26(21):2678–83.
21. Yao YH, Dai Q, Li L, Nan XY, He PA, Zhang YZ. Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation. *J Comput Chem*. 2010;31(5):1045–52.
22. Randić M. 2-D graphical representation of proteins based on physico-chemical properties of amino acids. *Chem Phys Lett*. 2007;444(1–3):176–80.
23. Yao YH, Dai Q, Li C, He PA, Nan XY, Zhang YZ. Analysis of similarity/dissimilarity of protein sequences. *Proteins Struct Funct Bioinformatics*. 2008;73(4):864–71.
24. Yau SS, Yu C, He R. A protein map and its application. *DNA Cell Biol*. 2008;27(5):241–50.
25. Abo el Maaty MI, Abo-Elkhier MM, Abd Elwahaab MA. 3D graphical representation of protein sequences and their statistical characterization. *Physica A*. 2010;389(21):4668–76.
26. Wu ZC, Xiao X, Chou KC. 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*. 2010;267(1):29–34.
27. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*. 1984;179(1):125–42.
28. Chechetkin VR, Lobzin VV. Characterization and comparison of protein structures. Part I-characterization. *J Theor Biol*. 1999;198(2):197–218.
29. Hu Xiu Zhen LJ. Statistical analysis of application of hydrophilicity-hydrophobicity and molecular size of amino acid. *J Inner Mongolia Polytech Univ*. 2000;19(3):187–91.
30. Qi ZH, Feng J, Qi XQ, Li L. Application of 2D graphic representation of protein sequence based on Huffman tree method. *Comput Biol Med*. 2012;42:556–63.
31. Liao B, Liao B, Lu X, Cao Z. A novel graphical representation of protein sequences and its application. *J Comput Chem*. 2011;32(12):2539–44.
32. Randić M, Vračko M, Nandy A, Basak SC. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J Chem Infor Comput Sci*. 2000;40:1235–44.
33. Randić M, Zupan J, Novic M. On 3-D graphical representation of proteomics maps and their numerical characterization. *J Chem Inf Comput Sci*. 2001;41:1339–44.
34. Nie ZM, Zhang ZF, Wang D, et al. Complete sequence and organization of Antheraea pernyi nucleopolyhedrovirus, a dr-rich baculovirus. *BMC Genomics*. 2007;8:248.
35. Herniou EA, Olszewski JA, O'Reilly DR, Cory JS. Ancient coevolution of baculoviruses and their insect hosts. *J Virol*. 2004;78:3244–51.
36. Jiang Y, Deng F, Wang HL, Hu ZH. An extensive analysis on the global codon usage pattern of baculoviruses. *Arch Virol*. 2008;153:2273–82.
37. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23:2947–48.
38. Cao Y, Janke A, Waddell PJ, et al. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J Mol Evol*. 1998;47(3):307–22.
39. Otu HH, Sayood K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*. 2003;19(16):2122–30.