A new 3D graphical representation for similarity/dissimilarity studies of protein sequences

Yan Chen, Kang-Shun Li*, Shan Chang, Lei Yang

College of Information, South China Agricultural University, Guangzhou 510642, China

Received 6 October 2013, www.cmnt.lv

Abstract

With the development of sequencing technology and the rapid growing number of protein sequences, how to find useful information from these large numbers of protein sequences has become an important research focus. The dominant factor of protein's characteristic is each amino acid of it. So this paper uses three-dimensional Cartesian coordinate system to represent three important physical chemistry properties of amino acids: hydrophobicity of amino acids, aromatic amino acids, and side-chain conformations.A new 3D graphical representation of protein sequences is proposed, based on the analysis. Using this graphical approach, 1D sequence of the protein can be expressed as a 3D graphics. At the same time, the similarity comparison of protein-sequences, prediction of functional sites, and other sequence analysis operations can be done further. The paper selects 15 protein sequences of ND6 to conduct the experiment, and the result shows that the analysis of the structures is consistent with the actual results of biological evolution. The experiment illustrates the utility of our approach.

Keywords: Protein sequence; 3D representation; Similarity/Dissimilarity Studies

1 Introduction

The rapid growth of biological sequence such as DNA, RNA, and protein has created many challenges for bio-scientists, in which the considerable efforts have been made to find the reliable and fully automated methods to analyze the vast amount of sequence data. The graphical representation method is one of those methods. It has such an important advantage over other methods: it provides not only visual qualitative inspection of gene data, helping to recognize major differences among similar gene sequences, but also mathematical characterizations of proteome maps[1]. DNA's graphical representations were initiated over 20 years [2-5].

In the 1950s, Anfinsen et al. found all the information of protein structure was hidden in protein sequence, i.e. amino acid sequence. Therefore, studies on protein sequences have become a key issue in the field of Bioinformatics [6-10]. However, it is difficult to directly obtain useful information from the original one-dimensional sequence. For this reason, many researchers have designed a number of methods, for example, converting protein sequences into digital signals, graphics and so on. However, the first graphical representation of proteins was proposed only very recently and just a few representations were outlined [11-17].

Although the existing methods can express protein sequences as 2D or 3D graphics intuitively, they simply regard 20 kinds of amino acids as 20 different symbols, without considering that the amino acids have different physicochemical properties, which has important linkages with the structure and function of a protein. Therefore, a graphical method of protein sequence must consider the properties of amino acids, and extract the information of the properties hidden behind amino acid sequences during the graphical process. This paper proposes a new 3D graphical representation of protein sequences by using three coordinate axes to represent the three different properties of amino acids. The protein sequence can be expressed as a 3D graphics to enable some graphics processing technologies to be applied for similarity comparison and the inherent information of protein sequences can be also better reflected.

2 Materials and methods

2.1 DATABASE

The test data selected by this paper is the protein 1D sequence data of 15 different species from the classic NADH dehydrogenase subunit 6, and the reason is that the test data's commonality and distinguishability is considered. Studies on ND6 proteins have been relatively mature, and the sequence types obtained by sequencing are more plentiful, therefore, this paper selected 15 different species from ND6 as the dataset. The test data is from NCBI database, as shown in Table 1:

* ***Corresponding author*** e- mail: likangshun@scau.edu.cn

TABLE 1 the protein sequence data of 15 different species from the classic NADH dehydrogenase subunit 6

Id	species	length	protein sequence
human	AP_000650	174	mmyalfllsv glvmgfvgfs skpspiyggl vlivsgvvgc viilnfgggy mglmvfliyl ggmmvvfgyt tamaieeype awgsgvevlv svlvglamev glvlwvkeyd gvvvvvnfns vgswmiyege gsgliredpi gagalydygr wlvvvtgwtl fvgvyiviei argn
gorilla	NP_008223	174	mtyvlfllsv glvmgfvgfs skpspiyggl vlivsgvvgc aiilncgggy mglmvfliyl ggmmvvfgyt tamaigeype awgsgvevlv svlvglamev glvlwvkeyd gvvvvvnfnn vgswmiyege gsgliredpi gagalydygr wlvvvtgwtl fvgvyiviei argn
Chimpanzee	NP_008197	174	mtyalfllsv slvmgfvgfs skpspiyggl vlivsgvvgc aiilnygggy mglmvfliyl ggmmvvfgyt tamaieeype awgsgvevlv svlvglamev glvlwvkgyd gmvvvvnfns vgswmiyege gpgliredpi gagalydygr wlvvvtgwtl fvgvyiviei argn
lemur	NP_659299	171	myvmfllsil lvlgfvsiss kpspiyggvg livsgavgcg iimgfggsfm glmvfliylg gmlvvfgytt amateeypet wgsnvviwgv vllgvgmelf mvawmveygg fgvgdvfggv enwmifeske ggviredslg vaslynkasw faaiagwslf isvlivieii r
Goat	NP_877414	175	mmmyivfils vifvmgfvgf sskpspiygg lglivsggvg cgivlnfggs flglmvfliy lggmmvvfgy ttamateqyp eiwvsnkvvl gafitgllme flmvyyvlkd keveivfkfn gmgdwviydt gdsgffseea mgiaalysyg twlvivtgws lligvvvime itrgn
sheep	NP_008417	175	mmtyivfils iifvmgfvgf sskpspiygg lglivsggvg cgivlnfggs flglmvfliy lggmmvvfgy ttamateqyp evwvsnkvvl gtfitgllme flmvyyvlkd keveivfkfn gmgdwviydt gdsgffseea mgiaalysyg twlvivtgws lligvvvime itrgn
bovine	YP_209216	175	mmlyivfils vifvmgfvgf sskpspiygg lglivsggvg cgivlnfggs flglmvfliy lggmmvvfgy ttamateqyp eiwlsnkavl gafvtgllme ffmvyyvlkd kevevvfefn glgdwviydt gdsgffseea mgiaalysyg twlvivtgws lligvvvime itrgn
rabbit	NP_007560	174	mtyvvfllsv mfvmgfvgfs skpspiyggl glivsggvgc givlsfggsf lglmmfliyl ggmlvvfgyt tamateeype twgsnvmilg mfvlgvlmev glvvymvmsd gveivvdfkn mgdwvvfegd evgliredsm gvaalysygs wlmvvagwsl fvsifiviei trga
European hare	NP_659336	174	mtymvfllsv mfvigfvgfa skpspiyggl glivsggvgc giilsfggsf lglmmfliyl ggmlvvfgyt tamateeype twgsnimils mlvlgvllea glvmfmavsd evevvvsfkn mgdwvvfegd evgliredsm gvaalysygs wlmvvagwsl fvsifiviei trgg
mouse	NP_904339	172	mnnyifvlss lflvgclgla lkpspiyggl glivsgfvgc lmvlgfggsf lglmvfliyl ggmlvvfgyt tamateeype twgsnwlilg flvlgvimev flicvlnyyd evgvinldgl gdwlmyevdd vgvmleggig vaamyscatw mmvvagwslf agifiiieit rd
Rat	YP_665640	172	mtnymfilsl lfltgclgla lkpspiyggl glivsgcigc lmvlgfggsf lglmvfliyl ggmlvvfgyt tamateeype twgsnwfifs ffvlglfmel vvfylfslnn kvelvdfdsl gdwlmyeidd vgvmleggig vaaiyscatw mmvvagwslf agifiiieit rd
opossum	NP_007106	168	mkmmtiyiis lllmigfvaf askpspiygg lslvvsgglg cgmvvsledv flglvvflvy lggmlvvfgy ttamateeyp etwvgnvvaf imllfvlllq vgwyfmsklv yiimaiklfd fvetslvgqd yngvsqlyyc ggwalallgw ilfmtiyvvl evvrersy
gallus	NP_006927	173	mtyfviflgi cfmlgvlava snpspyygvv glvvasvmgc gwlvslgvsf vslalflvyl ggmlvvfvys vslaadpype awgdwrvvgy glgfvlvvwm gvvlgglvdf wkvgvvtvdg ggvsfarldf sgvavfyscg vglflvagwg lllalfvvle lvrglsrgai rav
zebra finch	YP_514831	172	mmefvlflgl cfvlgglgva snpspyygvm glvvaavagc gwlvslgvsf vslvlvmvyl ggmlvvfvys vslaadpype swadwgvvgy gfgmglvvvv glvvggvsgv lvdegtvnsg gllsvrsdfs gvavlysega glliggwgl lltlfvvlel vrglsrgair av
Muscovy duck	YP_001974	173	mtyfvfflgv cfvlgilgva snpspyygvv glvlasvagc gwllslgvsf valvlfmvyl ggmlvvfvys valaaepfpe awgdwrvlgy vvalvvvvlg glvlggfvgs wgfgvntvds vgmfvvrldf sgvamlysrg vgmfliagwg llltlfvvle lvrglsrgai rav

2.2 OUR 3D GRAPHICAL REPRESENTATION OF PROTEIN SEQUENCE

The basic units what make up proteins are amino acids. The primary structure of a protein refers to its amino acid sequence. The protein is folded into a 3D tertiary structure through the interaction between the hydrophobicity, chargeability and other properties of its amino acid residues. According to the research of Anfinsen, the primary structure of protein determines its 3D structure, and the properties of amino acids making up the protein's primary structure play a very important role in the folding of the protein[18-20]. Therefore, this paper selected amino acids' three physicochemical properties with important influence and function on folding as the basis for the construction of its 3D structure.

2.2.1 Hydrophilicity of amino acids

"Hydrophilicity" is an important property of amino acids. Hydrophobic amino acids tend to stay away from the surrounding water molecules, and embed themselves into the protein. This trend, 3D space conditions and other factors determine the folding 3D conformation of a protein. 20 kinds of amino acids are divided into two classes, in which the hydrophobic residues include C(Cys), F(Phe), Y(Tyr), W(Trp), M(Met), L(Leu), I(Ile) and V(Val), and the hydrophilic ones are G(Gly), P(Pro), A(Ala), T(Thr), S(Ser), N(Asn), H(His), Q(Gln), E(Glu), D(Asp), R(Arg), K(Lys). X-axis is used to coordinate to express hydrophobic property of amino acids in this paper, where H1 stands for hydrophobic amino acids, and H2 stands for hydrophilic amino acids.

2.2.2 Aromatic amino acids

Aromatic amino acids are amino acids that including an aromatic ring. Aromatic amino acids include phenylalanine (phe), tyrosine (tyr) and tryptophan (trp), which play an extremely important role in life activities. Metabolic disorders of aromatic amino acids can cause a variety of diseases, such as phenylketonuria (PKU) and tyrosinemia which are of great significance in liver diseases, kidney diseases, neuropsychiatric disorders, cancer and other diseases. Y-axis is used to coordinate to express aromatic or non-aromatic amino acids in this paper, where A1 stands for aromatic amino acids, and A2 stands for non-aromatic amino acids.

2.2.3 Side-chain conformations

The corresponding dihedral angles of bonds between atoms on side chains of residues are named as X1, X2 and X3. The side chains have a variety of different conformations, but each type of residues has several relatively stable conformations of side chains. Side-chain conformations of amino acids play a very important role in folding protein structure. Z-axis is used to coordinate to express side-chain conformational property of amino acids in this paper, where C1 stands for the atoms with more than 2 side-chain conformations, and C2 stands for the remaining.

Based on the above three important properties of amino acids, 3D curves of protein sequences by X, Y and Z-axis are shown as Table 2.

TABLE 2 the 3D curves of protein sequences by X, Y and Z-axis

Division Method	1	2	Coordinate
H (hydrophobic)	C(Cys), F(Phe), Y(Tyr), W(Trp), M(Met), L(Leu), I(Ile), V(Val)	G(Gly), P(Pro), A(Ala), T(Thr), S(Ser), N(Asn), H(His), Q(Gln), E(Glu), D(Asp), R(Arg), K(Lys)	x - component
A (aromatic)	W(Trp), F(Phe), Y(Tyr),	A(Ala), V(Val), C(Cys), G(Gly), I(Ile), L(Leu), M(Met), S(Ser), T(Thr), N(Asn), Q(Gln), K(Lys), R(Arg), D(Asp), E(Glu), P(Pro), H(His)	y - component
C (Conformational)	D(Asp), E(Glu), F(Phe), H(His), I(Ile), K(Lys), L(Leu), M(Met), N(Asn), Q(Gln), R(Arg), S(Ser), T(Thr), W(Trp), Y(Tyr)	A(Ala), C(Cys), G(Gly), P(Pro), V(Val)	z - component

Therefore, for a protein sequence with amino acids, the construction process of its 3D curve is as follows:

Step 1: calculate the number of amino acids with certain property from the first amino acid successively. For example, to plot the coordinate of X-axis, assuming that before positions, the number of hydrophobic amino acids is H_i^1 and the number of non-hydrophobic amino acids is H_i^2. Then the value of the curve coordinate x_i is $H_i^1 - H_i^2$. Accordingly, the values of y_i-axis and z_i-axis of the 3D curve can be obtained, as follows formula:

$$P_i = \begin{cases} x_i = \sum_{k=1}^{i} H_i^1 - \sum_{k=1}^{i} H_i^2 \\ y_i = \sum_{k=1}^{i} A_i^1 - \sum_{k=1}^{i} A_i^2 \\ z_i = \sum_{k=1}^{i} C_i^1 - \sum_{k=1}^{i} C_i^2 \end{cases} . \quad (1)$$

$$(i = 1, 2, 3, \ldots, n)$$

Step 2: Since the number of amino acids is different according to the classification of different properties, it is difficult for the follow-up analysis of the relevant characteristics of protein generated based on the nature of the curve. Therefore, according to the number of classification, scale factor are set: $e_x^1 = 12$, $e_x^2 = 8$, $e_y^1 = 15$, $e_y^2 = 5$, $e_z^1 = 5$, $e_z^2 = 15$. And the formula (1) is converted into the following formula:

$$P_i = \begin{cases} x_i = \sum_{k=1}^{i} e_x^1 H_i^1 - \sum_{k=1}^{i} e_x^2 H_i^2 \\ y_i = \sum_{k=1}^{i} e_y^1 A_i^1 - \sum_{k=1}^{i} e_y^2 A_i^2 \\ z_i = \sum_{k=1}^{i} e_z^1 C_i^1 - \sum_{k=1}^{i} e_z^2 C_i^2 \end{cases} . \quad (2)$$

$$(i = 1, 2, 3, \ldots, n)$$

Step 3: Assuming that the length of the protein sequence for 3D graphical representation is n, the three coordinates are standardized in accordance with the length of the protein sequence, as the following formula:

$$P_i = \begin{cases} x_i = \dfrac{\sum_{k=1}^{i} e_x^1 H_i^1 - \sum_{k=1}^{i} e_x^2 H_i^2}{n} \\ y_i = \dfrac{\sum_{k=1}^{i} e_y^1 A_i^1 - \sum_{k=1}^{i} e_y^2 A_i^2}{n} \\ z_i = \dfrac{\sum_{k=1}^{i} e_z^1 C_i^1 - \sum_{k=1}^{i} e_z^2 C_i^2}{n} \end{cases} . \quad (3)$$

$$(i = 1, 2, 3, \ldots, n)$$

Following the above three steps, the 3D curve of a protein sequence can be constructed. Taking human's ND6 protein for example; the 3D curve and the curves of the three coordinate components are shown below:

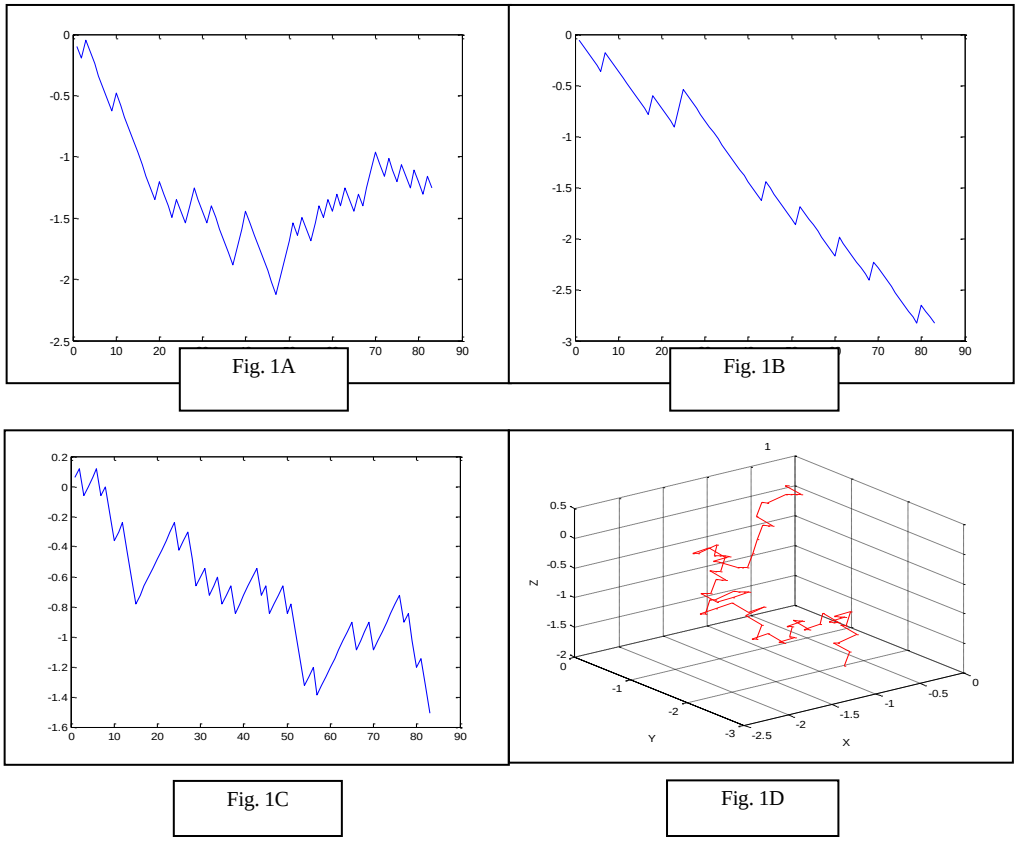

Fig. 1A Fig. 1B

Fig. 1C Fig. 1D

FIGURE 1 The new 3D representation of the serine protein's inhibitor CI-2 from barley seeds

2.3 NUMERICAL CHARACTERIZATION OF PROTEIN SEQUENCES

In order to find some sensitive invariants to represent our graphical, we will transform the graphical into another mathematical object, a matrix. Once we have a matrix representing protein sequence, we can use some of matrix invariants as descriptors of the sequence. For example, Similarity/Dissimilarity was leading eigenvalue. Commonly there are three kinds of matrix: ED, L/L and M/M, which are introduced by M. Randic.

2.3.1 The Euclidean matrix E

The E matrix is the symmetric matrix, E=ET, the (i, j) element is defined as the Euclidean distance between vertices i and j of the curve, which is defined as:

$$e_{ij} = \begin{cases} d_{ij} & (j \neq i) \\ 0 & (j = i) \end{cases}. \tag{4}$$

2.3.2 The M/M matrix

The off-diagonal entries of the M/M matrix are given as a quotient of the Euclidean distance between two vertices of the curve and the number of edges (the so-called graph theoretical distance). The entries on the main diagonal are defined as zero. The M/M matrix is symmetric, which is defined as:

$$m_{ij} = \begin{cases} \dfrac{e_{ij}}{|j-i|} & (j \neq i) \\ 0 & (j = i) \end{cases}. \tag{5}$$

2.3.3 The L/L matrix

The L/L matrix is the symmetric matrix whose off-diagonal elements are defined as a quotient of the Euclidean distance between two vertices of the curve and the sum of geometrical lengths of edges. All diagonal entries are zero, which is defined as:

$$l_{ij} = \begin{cases} \dfrac{e_{ij}}{\sum_{k=i}^{j-1} e_{k,k+1}} & (j \neq i) \\ 0 & (j = i) \end{cases}. \tag{6}$$

2.4 SIMILARITY/DISSIMILARITY STUDIES OF PROTEIN SEQUENCE

Given two arbitrary sequences, $S^1 = s_1^1 s_2^1 s_3^1 \cdots s_{N_1}^1$ and $S^2 = s_1^2 s_2^2 s_3^2 \cdots s_{N_2}^2$, their lengths are N_1 and N_2. In the graphical approaches, more than one graph is usually indicated to completely represent a sequence, so a set of leading eigenvalues from graphs can be obtained $\lambda_1^1, \lambda_2^1, \lambda_3^1 \cdots \lambda_k^1$ and $\lambda_1^2, \lambda_2^2, \lambda_3^2 \cdots \lambda_k^2$ are respective k-

dimensions vectors composed of the leading eigenvalues of characteristic curves based on k different patterns of the sequence S_1 and S_2. So far, almost all such similarity/dissimilarity comparisons of sequence S_1 and S_2 are based as three ways:

The E matrix is the symmetric matrix, E=ET, the (i, j) element is defined as the Euclidean distance between vertices i and j of the curve, which is defined as:

$$D(S_1, S_2) = \sqrt{\sum_{i=1}^{k} \left(\frac{\lambda_i^1}{N_1} - \frac{\lambda_i^2}{N_2} \right)^2} . \tag{7}$$

The off-diagonal entries of the M/M matrix are given as a quotient of the Euclidean distance between two vertices of the curve and the number of edges (the so-called graph theoretical distance). The entries on the main diagonal are defined as zero. The M/M matrix is symmetric, which is defined as:

$$\theta(S_1, S_2) = \arccos \frac{\sum_{i=1}^{k} \lambda_i^1 \lambda_i^2}{\sqrt{\sum_{i=1}^{k} (\lambda_i^1)^2} \sqrt{\sum_{i=1}^{k} (\lambda_i^2)^2}} . \tag{8}$$

$(0 \le \theta \le 4\pi)$

The L/L matrix is the symmetric matrix whose off-diagonal elements are defined as a quotient of the Euclidean distance between two vertices of the curve and the sum of geometrical lengths of edges. All diagonal entries are zero, which is defined as:

$$r^2(S_1, S_2) = \frac{[k\sum_{i=1}^{k} \lambda_i^1 \lambda_i^2 - (\sum_{i=1}^{k} \lambda_i^1)(\sum_{i=1}^{k} \lambda_i^2]^2}{[k\sum_{i=1}^{k} (\lambda_i^1)^2 - \sum_{i=1}^{k} (\lambda_i^1)^2][k\sum_{i=1}^{k} (\lambda_i^2)^2 - \sum_{i=1}^{k} (\lambda_i^2)^2]} . \tag{9}$$

3 Results and discussion

In this section, we illustrate the use of the quantitative characterization of protein sequences with an examination of the similarity among 15 species of ND6 proteins. Figure 2 shows x, y and z-component of curves for the human, gorilla and Muscovy duck of ND6 proteins. Observing these curves, we can see that they are very similar curves between human and gorilla, but they are different from Muscovy duck. Their similarities/dissimilarities are consistent with the known fact of evolution.

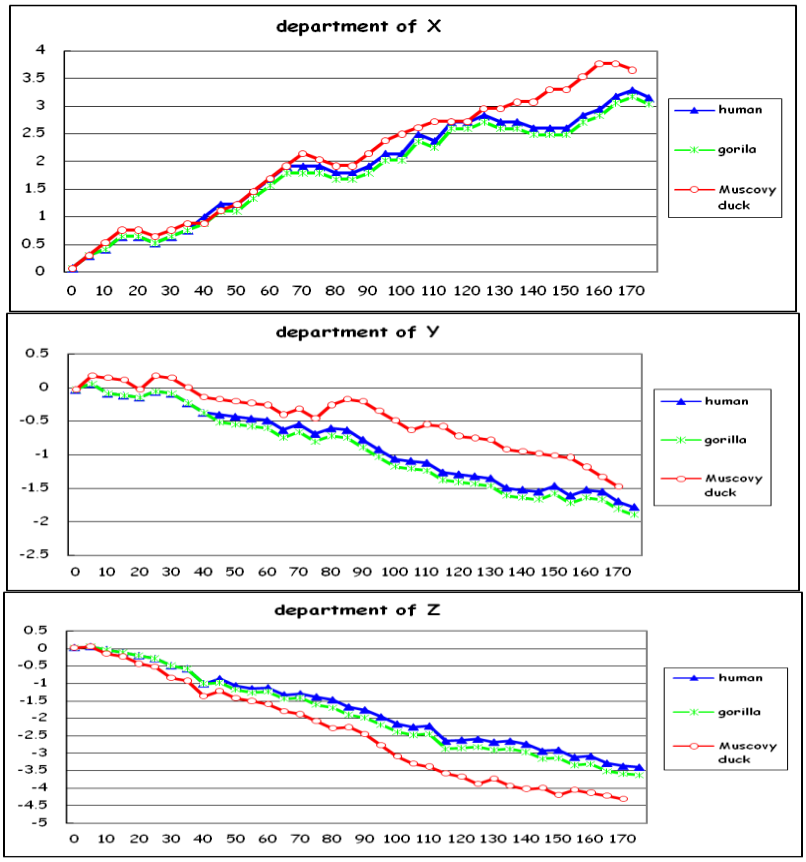

FIGURE 2 x-component, y-component and z-component of curves for the human, gorilla and Muscovy duck of ND6 proteins

The analysis of similarity/dissimilarity is based as three following ways: (1)to calculate the Euclidean distance between the end points of two normalized vectors; (2) to calculate the correlation angle of two vectors; (3) to

calculate the coefficient of determination of two vectors. In Tables 3 and 4, we give the similarity/dissimilarity matrices for fifteen ND6 protein sequences based on the Euclidean distances and correlation angles between 6-component vectors of the normalized leading eigenvalues of the L/L matrices. From table 3 and 4, we can find ND6 proteins of human, gorilla, common chimpanzee and Lemur are more similar with each other, and ND6 proteins are more similar with (Goat, Sheep), (Rabbit, E_hare), (rat, mouse, Opossum) and (Gallus, Z-finch, M-duck). On the other hand, we find ND6 protein of Bovine is very dissimilar to others among eight species because its corresponding row has larger entries.

TABLE 3 Euclidean distances

species	Goat	Sheep	Bovine	Human	Gorilla	chimpanzee	Lemur	Rabbit	E_hare	Mouse	Rat	Opossum	Gallus	Z-finch	M-duck
Goat	0	0.098868	0.26447	3.5431	4.1453	3.7058	2.344	1.3948	0.62834	1.0399	0.67915	0.37594	6.3649	6.6714	6.0221
Sheep	0.098868	0	0.35173	3.6289	4.2316	3.79	2.4261	1.4808	0.71129	1.1252	0.60308	0.47471	6.4516	6.7593	6.1102
Bovine	0.26447	0.35173	0	3.2789	3.8813	3.4414	2.0797	1.1306	0.36434	0.7755	0.94309	0.22143	6.101	6.4081	5.7589
Human	3.5431	3.6289	3.2789	0	0.60399	0.19747	1.2183	2.1484	2.9178	2.5037	4.2184	3.2337	2.8238	3.1378	2.4911
Gorilla	4.1453	4.2316	3.8813	0.60399	0	0.47077	1.8216	2.7509	3.5208	3.1064	4.8196	3.8322	2.2202	2.534	1.8879
chimpanzee	3.7058	3.79	3.4414	0.19747	0.47077	0	1.3679	2.3118	3.0787	2.6659	4.3828	3.4031	2.6718	2.9929	2.35
Lemur	2.344	2.4261	2.0797	1.2183	1.8216	1.3679	0	0.95768	1.7157	1.3063	3.0226	2.0567	4.0385	4.3555	3.7093
Rabbit	1.3948	1.4808	1.1306	2.1484	2.7509	2.3118	0.95768	0	0.7705	0.35579	2.071	1.0992	4.9708	5.2793	4.6303
E_hare	0.62834	0.71129	0.36434	2.9178	3.5208	3.0787	1.7157	0.7705	0	0.41473	1.3074	0.40436	5.7409	6.0498	5.4008
Mouse	1.0399	1.1252	0.7755	2.5037	3.1064	2.6659	1.3063	0.35579	0.41473	0	1.7173	0.75957	5.3264	5.635	4.9861
Rat	0.67915	0.60308	0.94309	4.2184	4.8196	4.3828	3.0226	2.071	1.3074	1.7173	0	0.99813	7.038	7.3419	6.6924
Opossum	0.37594	0.47471	0.22143	3.2337	3.8322	3.4031	2.0567	1.0992	0.40436	0.75957	0.99813	0	6.0477	6.3485	5.6989
Gallus	6.3649	6.4516	6.101	2.8238	2.2202	2.6718	4.0385	4.9708	5.7409	5.3264	7.038	6.0477	0	0.35533	0.39126
Z-finch	6.6714	6.7593	6.4081	3.1378	2.534	2.9929	4.3555	5.2793	6.0498	5.635	7.3419	6.3485	0.35533	0	0.64955
M-duck	6.0221	6.1102	5.7589	2.4911	1.8879	2.35	3.7093	4.6303	5.4008	4.9861	6.6924	5.6989	0.39126	0.64955	0

TABLE 4 Correlation angles

Species	Goat	Sheep	Bovine	Human	Gorilla	chimpanzee	Lemur	Rabbit	E_hare	Mouse	Rat	Opossum	Gallus	Z-finch	M-duck
Goat	0	0.005781	0.019333	0.13141	0.13669	0.14913	0.12927	0.072424	0.048311	0.061733	0.07487	0.023361	0.15951	0.14548	0.13796
Sheep	0.005781	0	0.013574	0.12565	0.13092	0.14336	0.1235	0.066657	0.042544	0.055964	0.080639	0.029126	0.15374	0.13971	0.13219
Bovine	0.019333	0.013574	0	0.11208	0.11735	0.1298	0.10994	0.053091	0.028978	0.042401	0.094202	0.042694	0.14018	0.12615	0.11863
Human	0.13141	0.12565	0.11208	0	0.00527	0.017714	0.002321	0.058995	0.083105	0.069683	0.20628	0.15477	0.0281	0.01407	0.006546
Gorilla	0.13669	0.13092	0.11735	0.00527	0	0.012442	0.007464	0.064266	0.088377	0.074954	0.21155	0.16004	0.022828	0.008799	0.001275
chimpanzee	0.14913	0.14336	0.1298	0.017714	0.012442	0	0.019866	0.076708	0.10082	0.087395	0.22399	0.17248	0.010387	0.003645	0.011168
Lemur	0.12927	0.1235	0.10994	0.002321	0.007464	0.019866	0	0.056864	0.080968	0.067545	0.20414	0.15262	0.030249	0.016223	0.008729
Rabbit	0.072424	0.066657	0.053091	0.058995	0.064266	0.076708	0.056864	0	0.024113	0.010695	0.14729	0.095783	0.087094	0.073064	0.065541
E_hare	0.048311	0.042544	0.028978	0.083105	0.088377	0.10082	0.080968	0.024113	0	0.013423	0.12318	0.071669	0.1112	0.097173	0.089651
Mouse	0.061733	0.055964	0.042401	0.069683	0.074954	0.087395	0.067545	0.010695	0.013423	0	0.1366	0.08509	0.097781	0.08375	0.076228
Rat	0.07487	0.080639	0.094202	0.20628	0.21155	0.22399	0.20414	0.14729	0.12318	0.1366	0	0.051514	0.23438	0.22035	0.21283
Opossum	0.023361	0.029126	0.042694	0.15477	0.16004	0.17248	0.15262	0.095783	0.071669	0.08509	0.051514	1.49E-08	0.18287	0.16884	0.16132
Gallus	0.15951	0.15374	0.14018	0.0281	0.022828	0.010387	0.030249	0.087094	0.1112	0.097781	0.23438	0.18287	0	0.014031	0.021554
Z-finch	0.14548	0.13971	0.12615	0.01407	0.008799	0.003645	0.016223	0.073064	0.097173	0.08375	0.22035	0.16884	0.014031	0	0.007525
M-duck	0.13796	0.13219	0.11863	0.006546	0.001275	0.011168	0.008729	0.065541	0.089651	0.076228	0.21283	0.16132	0.021554	0.007525	0

The similarity/dissimilarity matrix for the fifteen ND6 protein sequences is based on the coefficient of determination of the leading eigenvalues of the L/L matrices. And the results are shown in fig3. The fig3 provide us with additionally physicochemical meanings of graphical representation: it indicates the similarity between two protein sequences by regarding to the physicochemical properties of amino acids (percent). For example, the values corresponding to (human, gorilla) and (human, common chimpanzee) is 0.9830 and 0.9356. This implies similarity of human and gorilla is 98.30%, but similarity of human and common chimpanzee is only 93.56% for ND6 proteins.

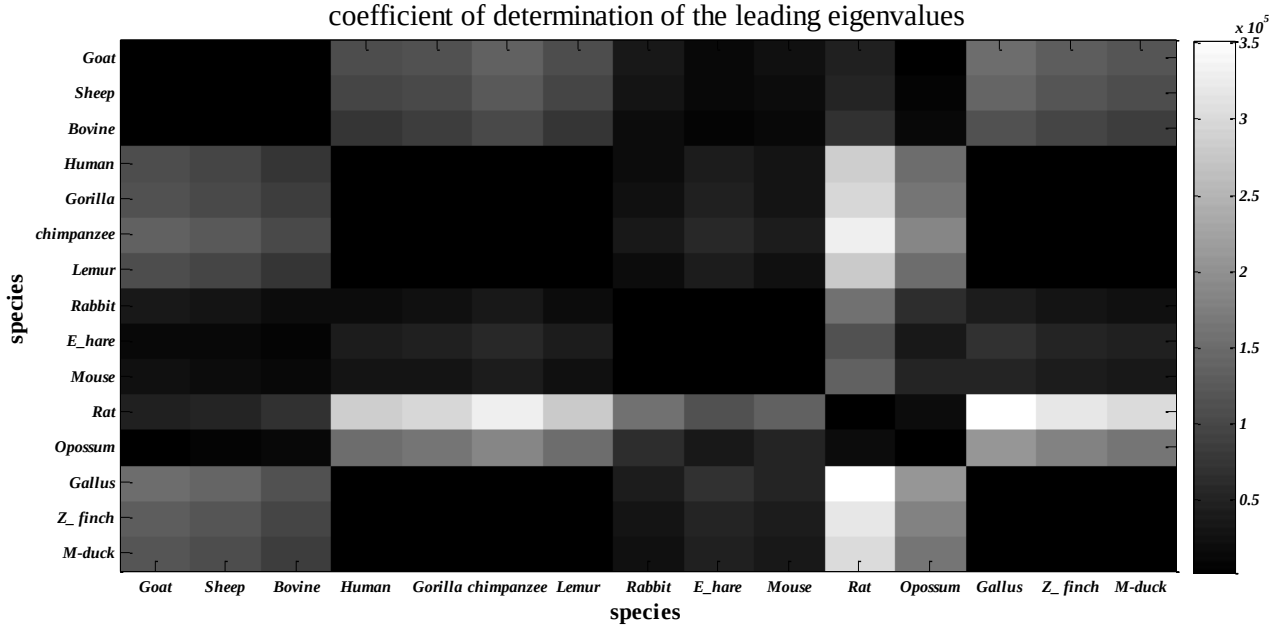

FIGURE 3 The coefficient determination of leading eigenvalues of ND6 proteins

4 Conclusion

Protein is a complex system that has one property only reflecting some characteristics. This paper proposes a new 3D graphica andthis 3D graphical representation considers three important properties of protein: hydrophilicity of amino acids as x-component, aromatic amino acids as y-component and side-Chain Conformations as z-component. This new 3D graphical representation represents the physicochemical properties of protein sequence very well. In order to verify the effectiveness of this algorithm, we use 3D graphical to represent 15 protein sequences , and then analyze their similarities/dissimilarities. The results match well with the evolutionary chronology of the organisms. So this method is useful for sequence comparison and helpful for the evolutionary study, structure and function prediction.

References

[1] J.Wen, Y.Y.Zhang 2009 A 2D graphical representation of protein sequence and its numerical characterization *Chemical Physics Letters* **476**(4-6), 281-6.

[2] E.Hamori, J.Ruskin 1983 H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences *J. Biol. Chem* **258**(2), 1318-27.

[3] E.Hamori 1985 Novel DNA sequence representations *Nature* **314**(6012), 585-6.

[4] M.A.Gates 1985 Simpler DNA sequence representations *Nature* **316**(6025), 219.

[5] M.A.Gates 1986 A simple way to look at DNA *J. Theor. Biol* **119**(3), 319-28

[6] *Deleted by CMNT Editor*

[7] L.Shen, H.F.Ji 2011 A multiscale approach to the simulation of asphaltenes *Computational and Theoretical Chemistry* **975**(1-3), 76-82

[8] E.Henriksson, J.Pesonen, P.Chacon 2012 Curvilinear Dynamics of Protein Complexes *Theoretical and Computational Chemistry* **11**(3), 675-96.

[9] A.Banerjee, A.Jana, B.R.Pati, K.C.Mondal, P.K.Mohapatra 2012 Characterization of Tannase Protein Sequences of Bacteria and Fungi: An In Silico Study *Journal of Protein Chemistry* **31**(4), 306-27.

[10] X.Zou, T.K.Pham, P.C.Wright, J.Noirel 2012 Bioinformatic study of the relationship between protein regulation and sequence properties *Genomics* **100**(4), 240-4.

[11] *Deleted by CMNT Editor*

[12] M.Randić 2004 2-D Graphical representation of proteins based on virtual genetic code *SAR QSAR Environ.Res* **15**(3), 147-57.

[13] C.T.Zhang, R.Zhang 1991 Analysis of distribution of bases in the coding sequences by a digrammatic technique *Nucl. AcidsRes* **19**(22), 6313-7

[14] R.Zhang, C.T.Zhang 1994 Z Curves, An Intutive Tool for Visualizing and Analyzing the DNA Sequences *J. Biomol.Struct.Dyn* **11**(4) 767-82

[15] M.Randić, J.Zupan, D.Vikić-Topić 2007 On representation of proteins by star-like graphs *J. Mol. Graphics Modell* **26**(1), 290-305.

[16] M.Randić, A.T.Balaban, M.Novic, A.Zaloznik, T.Pisanski 2005 A novel graphical representation of proteins *Period. Biol* **107**(4), 403-14

[17] L.P.Zhao, Y.H.Lv, C.Li, M.H.Yao, X.Z,Jin 2010 An S-Curve-Based Approach of Identifying Biological Sequences *Acta Biotheoretica* **58**(1), 1-14.

[18] L.Roland, J.R.Dunbrack, E.Fred, Cohen 1997 Bayesian statistical analysis of protein side-chain rotamer preferences *Protein Science* **6**(8), 1661-81.

[19] *Deleted by CMNT Editor*

[20] S.J.Sun, B.Rachel, H.S.Chan 1995 Designing amino acid sequences to fold with good hydrophobic cores *Protein Engineering* **8**(12), 1205-13.

Authors

< Yan Chen >, <1977.4>,< Guangzhou, Guangdong, P.R. China>

Current position, grades: the Lecturer of College of Information, South China Agricultural University, China.
University studies: received her M.E. in computer science and technology from Xi'an University of Architecture and Technology in China.
Scientific interest: Her research interest fields include evolutinary algorithm, bioinformatics
Publications: more than 5 papers published in various journals.
Experience: She has teaching experience of 10 years, has completed 3 scientific research projects.

< Kang-shun Li >, <1962.3>,< Guangzhou, Guangdong, P.R. China>

Current position, grades: the Professor of College of Information, South China Agricultural University, China.
University studies: received his PhD. in Computer Software and Theory from Wuhan University in China.
Scientific interest: His research interest fields include evolutinary algorithm
Publications: more than 85 papers published in various journals.
Experience: he has teaching experience of 20 years, has completed 20 scientific research projects.

< Shan Chang >, <1982.6>,< Guangzhou, Guangdong, P.R. China>

Current position, grades: the Associate Professor of College of Information, South China Agricultural University, China.
University studies: received his PhD. in Biotechnology from Beijing University of Technology in China.
Scientific interest: His research interest fields include bioinformatics
Publications: more than 50 papers published in various journals.
Experience: he has teaching experience of 5 years, has completed 5 scientific research projects.

< Lei Yang >, <1978.9>,< Guangzhou, Guangdong, P.R. China>

Current position, grades: the Lecturer of College of Information, South China Agricultural University, China.
University studies: received his M.E. in Computer Software and Theory from Shaanxi Normal University in China.
Scientific interest: His research interest fields include evolutinary algorithm
Publications: more than 5 papers published in various journals.
Experience: he has teaching experience of 8 years, has completed 3 scientific research projects.