***MATCH***
*Communications in Mathematical and in Computer Chemistry*

MATCH Commun. Math. Comput. Chem. **75** (2016) 463-480

ISSN 0340 - 6253

# An Intuitive Graphical Method for Visualizing Protein Sequences Based on Linear Regression and Physicochemical Properties

**Zhao-Hui Qi*, Meng-Zhe Jin**

[1]*College of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, 050043, P. R. China*

zhqi_wy2013@163.com

(Received August 22, 2015)

**Abstract**

In this paper, a novel protein map is introduced under a linear regression model of eight kinds of physicochemical indices. On the basis of the protein map, graphical representation of proteins is presented to provide clear visualization for protein sequence. Then, to further explore the degree of similarity among proteins, we introduce a 40-component vector for protein numerical characterization. The cosine distances among the vectors are taken as the similarity distances among the corresponding proteins. The analysis results of two real datasets demonstrate that the new scheme is effective in similarity research and phylogenetic analysis.

## 1 Introduction

With the increasing number of available sequences in various biological databases, numerous methods have been proposed to take the analysis of biological sequences in recent years. Among them, graphic methods provide intuitive inspection of data by visualization of sequences and hence become sought after by many researchers. In 1983, Hamori and Ruskin [1] initially applied the H-curve in the visualization and analysis of DNA sequences. They assigned the four corners of a square to the four bases and determined the running index as the third coordinate. The work of Hamori and Ruskin greatly stimulated the following researches on graphic representation of DNA by Gates [2] and Nandy [3]. In those years, the proposed 2D random-walk plot methods had two aspects of limitations: one was the degeneracy generated by circuit, and the other was the loss of

information brought from overlapping. However, with the rapid development of bioinformatics, researchers have proposed various DNA visualizing methods, such as 3D, 4D, 5D [4-9] and Jeffrey's 'magic square' graphical representations [10], to avoid the deficiencies.

The graphical representation of proteins can be more difficult than that of DNA. Comparing with the graphic representation of DNA, the emergence of graphical representations of proteins was delayed for a long time because of the increased complexity in construction of biological descriptors built from the 20 natural amino acids but not the 4 nucleotides. But until recently, various graphical methods have been provided by researchers for visualizing and analyzing protein sequences [11-23]. Several researchers [11-14] first reported their graphical representations of proteins by modifying the existing graphical representations of DNA. Among them, Randić et al [13] proposed the representation of proteins via 'magic circle', which was generalized from Jeffrey's DNA representation method [10] by replacing a square with an icosagon. Moreover, a notion of "Virtual Genetic Code" was provided by Randić [13,15,16] and Liao et al. [17], which offered "transcription" from any protein sequence to a virtual DNA sequence. Besides considering the superficial alphabetical representation of amino acids, more and more researchers [18-23] turned to study the various physicochemical properties of the twenty natural amino acids, such as molecular weight, the *Hydropathy* index, the *Hydrophobicity* values and $PK_a$ values for the terminal amino acid groups $-NH_3^+$ and $-COO^-$. Protein maps were proposed in these researches by the ranking of the physicochemical properties or the transformation from physicochemical values to coordinates in the new Cartesian system.

Other than the advantage of direct visualization of sequences, a variety of available invariant techniques can be directly applied in numerical characterization of biological sequences for the sake of the similarity research and the phylogenetic analysis. Both DNA representations and protein representations can be associated with a characterizing matrix, such as E, D/D, L/L. Matrix invariants such as eigenvalues are used as the characteristic values to describe sequences. However, the construction of such a matrix cost $O(n^2)$ computational complexity, which can be a serious obstacle when sequence length $n$ is larger. Several solutions [19,22,23] have been proposed for bypassing this difficulty. Yao et al. [19,23] suggested a moment of inertial tensor be the descriptor of protein sequences. He et al. [22] proposed a similarity distance computing method corresponding to their 3D graphical representations based on the cumulative distance.

The natural amino acids are the building blocks in modeling the protein structure and the physicochemical properties are reported to be effective on the rate of amino acid substitution [24].

Therefore, it's worthy of consideration in adopting the physicochemical properties of the twenty amino acids in protein sequence analysis.

In this article, we outline a novel protein 2D map method based on the simple regression model of the eight of the physicochemical properties. Adding the running number of amino acid as the third component, we obtain a 3D graphical method for visualizing protein sequences. Then, a novel descriptor is proposed as sequence numerical characterization based on the 2D map. The proposed scheme achieves well performance in the applications on similarity research of the ND5 (NADH dehydrogenase subunit 5) proteins from nine species and phylogenetic analysis of 36 proteins from the Chew-Kedem dataset.

## 2 Intuitive graphical method for visualizing protein sequences

A protein sequence consists of twenty different amino acids, *A* (*Ala*), *C* (*Cys*), *D* (*Asp*), *E* (*Glu*), *F* (*Phe*), *G* (*Gly*), *H* (*His*), *I (lle*), *K* (*Lys*), *L* (*Leu*), *M (Met*), *N* (*A*sn), *P* (*Pro*), *Q* (*Gln*), *R* (*Arg*), *S* (*Ser*), *T* (*Thr*), *V* (*Val*), *W (Trp*) and *Y* (*Tyr*). Each amino acid is bundled with its own physicochemical molecular properties which have important effects on the evolutional pattern of proteins. Given a long protein sequence, it is difficult for us to understand the unintelligible character sequence. However, graphical methods provide intuitive tools for visualizing and analyzing protein sequences. In order to get the good visualization and consider the effects brought by physicochemical properties, the authors suggested some graphical methods of protein sequences based on different physicochemical properties [18-26]. For example, Yau et al. [20] presented a graphical representation method of proteins based on the hydrophobicity of amino acids. Hu [26] selected two different physicochemical properties to construct the 2D graphical representation of proteins. Here, to construct a graphical method of protein representation, we consider eight physicochemical properties essential for protein structure, enzyme catalyzed reaction and gene expression and regulation, such as the Polar Requirement (*PR*) [27], the Isoelectric Point (*IP*) [28], the Hydroxythiolation (*Hth*) [29], the Hydrogenation (*Hg*) [29], the *pK*1(-*COOH*) (*PK*1) [19], the *pK*2(- $NH_3^+$ ) (*PK*2) [19], the Aromaticity (*Am*) [29] and the Hydropathy Index (*HI*) [30]. We list the eight selected physicochemical properties in Table 1 and rearrange the properties according to the mean value of the twenty amino acids. More specially, the mean value of the *HI* of the twenty amino acids is the smallest in the eight properties, so we place the *HI* in the first column in Table 1. After that, we list the rest of properties according to their mean value until the *pK2* is placed.

**Table 1** Eight selected physicochemical properties associated with the 20 amino acids

| Amino acid | *HI* | *Hg* | *Hth* | *Am* | *PK1* | *IP* | *PR* | *PK2* |
|---|---|---|---|---|---|---|---|---|
| *A (Ala)* | 1.8 | 0.33 | -0.062 | -0.11 | 2.34 | 6.00 | 7.0 | 9.69 |
| *C (Cys)* | 2.5 | 0.074 | 0.38 | -0.184 | 1.71 | 5.07 | 4.8 | 10.78 |
| *D (Asp)* | -3.5 | -0.371 | -0.079 | -0.285 | 2.09 | 2.77 | 13.0 | 9.82 |
| *E (Glu)* | -3.5 | -0.254 | -0.184 | -0.067 | 2.19 | 3.22 | 12.5 | 9.67 |
| *F (Phe)* | 2.8 | 0.011 | 0.074 | 0.438 | 1.83 | 5.48 | 5.0 | 9.13 |
| *G (Gly)* | -0.4 | 0.37 | -0.017 | -0.073 | 2.34 | 5.97 | 7.9 | 9.60 |
| *H (His)* | -3.2 | -0.078 | 0.056 | 0.32 | 1.82 | 7.59 | 8.4 | 9.17 |
| *I (Ile)* | 6.02 | 0.149 | -0.309 | 0.001 | 2.36 | 6.02 | 4.9 | 9.68 |
| *K (Lys)* | -3.9 | -0.075 | -0.371 | 0.049 | 2.18 | 9.74 | 10.1 | 8.95 |
| *L (Leu)* | 3.8 | 0.129 | -0.264 | -0.008 | 2.36 | 5.98 | 4.9 | 9.60 |
| *M (Met)* | 1.9 | -0.092 | 0.077 | -0.041 | 2.28 | 5.74 | 5.3 | 9.21 |
| *N (Asn)* | -3.5 | -0.233 | 0.166 | -0.136 | 2.02 | 5.41 | 10.0 | 8.80 |
| *P (Pro)* | -1.6 | 0.37 | -0.036 | -0.016 | 1.99 | 6.30 | 6.6 | 10.60 |
| *Q (Gln)* | -3.5 | -0.409 | -0.025 | -0.246 | 2.17 | 5.65 | 8.6 | 9.13 |
| *R (Arg)* | -4.5 | -0.176 | -0.167 | 0.079 | 2.17 | 10.76 | 9.1 | 9.04 |
| *S (Ser)* | -0.8 | 0.022 | 0.47 | -0.153 | 2.21 | 5.68 | 7.5 | 9.15 |
| *T (Thr)* | -0.7 | 0.136 | 0.348 | -0.208 | 2.63 | 6.16 | 6.6 | 10.43 |
| *V (Val)* | 4.2 | 0.245 | 0.212 | -0.155 | 2.32 | 5.96 | 5.6 | 9.62 |
| *W (Trp)* | -0.9 | 0.011 | 0.05 | 0.493 | 2.38 | 5.89 | 5.2 | 9.39 |
| *Y (Tyr)* | -1.3 | -0.138 | 0.22 | 0.381 | 2.20 | 5.66 | 5.4 | 9.11 |

From Table 1, we can find that properties of distinct amino acids are numerically different from each other. It can be speculated, and rightly so, that the approximative physicochemical properties of amino acids mean their similar effects on the evolutional pattern of proteins and further provide the similarity of different proteins. To describe the approximative relationship graphically, we construct a unitary linear regression model to characterize each natural amino acid, which we call it as physicochemical equation. The equations are determined by linear regression of the eight physicochemical property values. The slope and the intercept of a regression equation are two characterizations and they together compose a 2D point called regression point. Then, we can get a one-to-one mapping between the twenty amino acids and their corresponding regression points.

Specifically, for a certain amino acid denoted as $x$, the values of the eight physicochemical properties, Hydropathy, Hydrogenation, Hydroxythiolation, Aromaticity, *pK*1(-*COOH*), Isoelectric

Point, Polar Requirement and *pK*2(- $NH_3^+$), are written orderly as $V_{HI}, V_{Hg}, V_{Hth}, V_{Am}, V_{pK1}, V_{IP}, V_{PR}$ and $V_{pK2}$. The unitary linear regression analysis is applied here to extract a two-dimensional characteristic of amino acid $X$. To achieve this goal, we append the ordinal numbers for the eight properties and get eight two-tuples, which are represented as eight points as follows:

$$property_X = \{(1, V_{HI}), (2, V_{Hg}), (3, V_{Hth}), (4, V_{Am}), (5, V_{pK1}), (6, V_{IP}), (7, V_{PR}), (8, V_{pK2})\}.$$

A linear regression equation can be fitted out by using these properties based on the Least Squares Approximation. The approach of Least Squares is a standard method in regression analysis to the approximate summarization of a sets of points. For a simple data set consisting of $n$ points $(x_i, y_i)$, $i = 1, 2, \cdots, n$, we want to summarize the information of the points by a straight line. The goal of the Least Squares Approximation is to find the overall solution which minimizes the sum of the squares of the errors made in the results.

The slope (denoted by $\mathbb{S}_X$) and the intercept (denoted by $\mathbb{I}_X$) of a regression equation by the Least Squares Approximation are two important indicators to characterize the properties of amino acid $X$. We fit the properties of each natural amino acid to a straight line and thus obtain twenty linear regression equations and we call them the physicochemical equations.

Here, we take physicochemical equation's establishment of the amino acid *Y (Tyr)* as an example. In Table 1, the eight physicochemical properties are orderly -0.7, 0.136, 0.348, -0.208, 2.63, 6.16, 6.6 and 10.43. The eight constructed two-tuples are shown as follows:

$$property_Y = \{(1, -0.7), (2, 0.136), (3, 0.348), (4, -0.208), (5, 2.63), (6, 6.16), (7, 6.6), (8, 10.43)\}.$$

By using the Least Squares Approximation, a straight line can be fitted out is obtained and the corresponding slope and intercept are respectively 1.41 and -3.67. In Fig. 1, the eight physicochemical values of the amino acid *Y (Tyr)* and its fitted straight line are represented graphically. The fitted physicochemical equation, their corresponding slopes and intercepts are shown in Table 2.

On the basis of the regression model, the 2D map between the twenty amino acids and their corresponding regression points is presented in Fig. 2 on the Cartesian coordinates. The $x$-coordinate represents for the slope of the unitary linear regression model, and the $y$-coordinate for the intercept. For better visualization, we take (0.80,-7.00) as the origin and adjust the coordinates. The modified slope and intercept are also presented in Table 2. It can be seen from Fig. 2 that the amino acid pairs (*C*, *M*) and (*W*, *Y*) own shorter distances than other pairs. That is to say, (*C*, *M*) and (*W*, *Y*) perform a fair amount of similarity in the physicochemical properties.

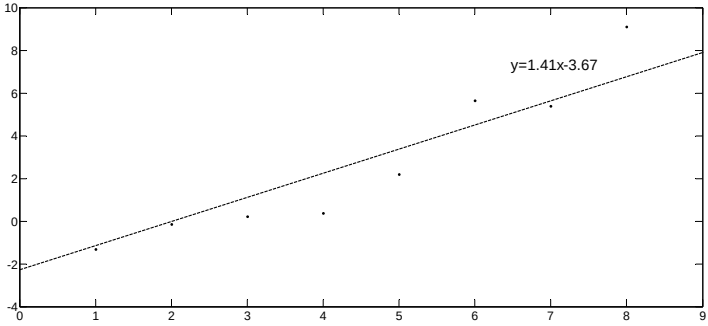


**Figure 1** Eight physicochemical values of the amino acid Y (Tyr) and its linear regression model

**Table 2** Linear equation, slope and intercept of each amino acid from linear regression of the eight physicochemical properties and their corresponding coordinates

| Amino acid( $X$ ) | Physicochemical equation | Slope( $\mathbb{S}$ ) | Intercept( $\mathbb{I}$ ) | Modified slope( $S$ ) | Modified intercept( $I$ ) |
|---|---|---|---|---|---|
| *A (Ala)* | $y = 1.30x - 2.48$ | 1.30 | -2.48 | 0.50 | 4.52 |
| *C (Cys)* | $y = 1.16x - 2.08$ | 1.16 | -2.08 | 0.36 | 4.92 |
| *D (Asp)* | $y = 2.04x - 6.23$ | 2.04 | -6.23 | 1.24 | 0.77 |
| *E (Glu)* | $y = 2.01x - 6.08$ | 2.01 | -6.08 | 1.21 | 0.92 |
| *F (Phe)* | $y = 1.03x - 1.56$ | 1.03 | -1.56 | 0.23 | 5.44 |
| *G (Gly)* | $y = 1.52x - 3.65$ | 1.52 | -3.65 | 0.72 | 3.35 |
| *H (His)* | $y = 1.82x - 5.19$ | 1.82 | -5.19 | 1.02 | 1.81 |
| *I (Ile)* | $y = 0.84x - 0.19$ | 0.84 | -0.19 | 0.04 | 6.81 |
| *K (Lys)* | $y = 2.06x - 5.95$ | 2.06 | -5.95 | 1.26 | 1.05 |
| *L (Leu)* | $y = 1.02x - 1.27$ | 1.02 | -1.27 | 0.22 | 5.73 |
| *M (Met)* | $y = 1.16x - 2.17$ | 1.16 | -2.17 | 0.36 | 4.83 |
| *N (Asn)* | $y = 1.85x - 5.50$ | 1.85 | -5.50 | 1.05 | 1.50 |
| *P (Pro)* | $y = 1.64x - 4.34$ | 1.64 | -4.34 | 0.84 | 2.66 |
| *Q (Gln)* | $y = 1.82x - 5.52$ | 1.82 | -5.52 | 1.02 | 1.48 |
| *R (Arg)* | $y = 2.10x - 6.14$ | 2.10 | -6.14 | 1.30 | 0.86 |
| *S (Ser)* | $y = 1.49x - 3.69$ | 1.49 | -3.69 | 0.69 | 3.31 |
| *T (Thr)* | $y = 1.55x - 3.82$ | 1.55 | -3.82 | 0.75 | 3.18 |
| *V (Val)* | $y = 1.01x - 1.02$ | 1.01 | -1.02 | 0.21 | 5.98 |
| *W (Trp)* | $y = 1.40x - 3.47$ | 1.40 | -3.47 | 0.60 | 3.53 |
| *Y (Tyr)* | $y = 1.41x - 3.67$ | 1.41 | -3.67 | 0.61 | 3.33 |

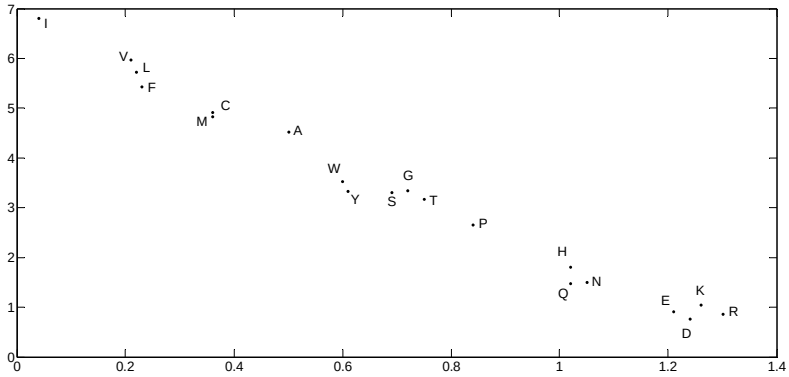


**Figure 2** The 2D map of the twenty amino acids based on their physicochemical properties

After determining the map between the twenty amino acids and their respective modified slope and intercept by linear regression, we are going to figure out the 3D representation of proteins. We suppose $P = a_1 a_2 a_3 \cdots a_n$ to be an arbitrary protein sequence with $n$ amino acids, where $a_i \ (1 \le i \le n)$ denotes the $i$ th amino acid in this protein sequence. If the protein is observed by stepping just one amino acid at a time, the 3D representation of protein sequence can be constructed for step $i$ ($i = 1, 2, 3, \ldots, n$) as follows:

$$\begin{cases} x_i = x_{i-1} + S_i \\ y_i = y_{i-1} + I_i \\ z_i = i \end{cases}, \tag{1}$$

where $x_i$, $y_i$ and $z_i$ denote the 3D coordinates of the $i$ th amino acid in the protein. $S_i$ denotes the modified slope of amino acid $a_i$ and $I_i$ denotes its modified intercept. The presentation is determined by the recursive relation and the initial condition is $\begin{cases} x_0 = 0 \\ y_0 = 0 \end{cases}$. As $i$ ranges from 1 to $n$, we have the graphical representation by connecting the $n$ points into a 3D zigzag curve.

To further illustrate the implementation of the proposed 3D protein graphical representation method, we consider an example. Two short protein segments (*Protein I and Protein II*) are taken from the yeast Saccharomyces cerevisiae, which is introduced by Randić et al [13] as an illustration of the current approach. Both segments are composed of thirty amino acids and respectively shown as follows:

*Protein I*:

WTFESRNDPAKDPVILWLNGGPGCSSLTGL

*Protein II*:

WFFESRNDPANDPIILWLNGGPGCSSFTGL

Fig. 3 (a) and (b) show the 3D visualizing representations for the two segments (*Protein I* and *Protein II*) based on the proposed method.

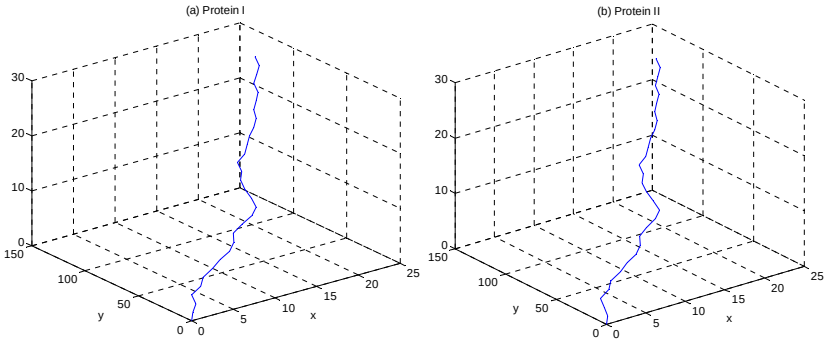


**Figure 3** Graphical representation of Protein I and II based on the proposed method

As it shows in Fig. 3, the two curves exhibit a high degree of similarity, which is consistent with most of the studies [19,23,31]. It can be seen there are several disagreements between the two figures. The differences actually emerge in the $2^{nd}$, $10^{th}$, $14^{th}$ and $27^{th}$ amino acid. The result tells the proposed 3D representation method has its usefulness and validation. In the same time, since the $z$-coordinate value grows from 1 to $n$, there will never be a circuit in the proposed 3D spatial curves.

## 3 Numerical Characterization of Proteins

Sequence comparison for similarity is a basic problem in the area of biological sequence analysis. Once the sequence map method is constructed, quantitative comparison among sequences can be implemented by sequence numerical characterization based on numerous matrix invariants such as the Euclidean matrix (**E** matrix) [32], the **D**/**D** matrix [33], the **M**/**M** matrix, the **L**/**L** matrix and their corresponding higher order matrix [34]. Among them, the **E** matrix is directly composed of the Euclidean distances between any two amino acids in the sequence. And the **D**/**D** matrix and the **L**/**L** matrix are constructed by computing the quotients of any two amino acids' Euclidean distance and the graph theoretical distance, and the quotients of the Euclidean distance and the topological distance. Further more, the cumulative distance and other invariants are used by some

researchers as a sensitive distance invariant in recent studies [17,23,35,36,37]. In this section, we give a novel sequence characterization method. On the basis of the characterization, similarity distance is obtained based on the cosine distance.

### 3.1 40-component characterizing vector

After the 2D protein map is constructed, numerical characterization of a protein becomes possible. Given a sequence $P = a_1 a_2 a_3 \cdots a_n$ (length of $n$), first, we convert each amino acid ($a_i$) into their corresponding modified slope ($S_i$) and intercept ($I_i$) based on the 2D map of the twenty amino acids and their corresponding regression points as is shown in Table 2. Thus, a symbolic sequence is transformed into a $n \times 2$ digital matrix. We suppose $aa$ to be one of the twenty amino acids, such as *A* (*Ala*), *C* (*Cys*), and *D* (*Asp*), and the modified slope and intercept of $aa$ to be $S^{aa}$ and $I^{aa}$, respectively. To construct a characterizing vector, we add up the cumulative slopes of $aa$ in the protein and denote it as the sum of slopes $\tilde{S}^{aa}$; and $\tilde{I}^{aa}$ as the sum of intercept. Thus, for an amino acid $aa$, we have a 2-tuple for characterization. Since there are twenty amino acids, we get a 40-component characterizing vector to describe a protein sequence. As is indicated, the 40-component vector is constructed by the sums of the slope and intercept of the twenty amino acids in a protein, which is shown as follows:

$$V = (\tilde{S}^A, \tilde{I}^A, \tilde{S}^C, \tilde{I}^C, \tilde{S}^D, \tilde{I}^D, ..., \tilde{S}^Y, \tilde{I}^Y)$$

where $\tilde{S}^{aa}$ and $\tilde{I}^{aa}$ are defined as $\begin{cases} \tilde{S}^{aa} = \sum S^{aa} \\ \tilde{I}^{aa} = \sum I^{aa} \end{cases}$. The components are arranged in the alphabetic order of the twenty amino acids: A→C→D→E→F→G→H→I→K→L→M→N→P→Q→R→S→T→V →W→Y.

For example, we consider the short protein segment, AACCDDCDDD, consisting of ten amino acids. First, the symbolic sequence is converted into a $10 \times 2$ digital matrix by the mapping from amino acids to their corresponding modified slopes and intercepts as is shown in Table 3.

**Table 3** the $10 \times 2$ digital matrix of AACCDDCDDD

| Amino acid | *A* | *A* | *C* | *C* | *D* | *D* | *C* | *D* | *D* | *D* |
|---|---|---|---|---|---|---|---|---|---|---|
| Modified Slope | 0.5 | 0.5 | 0.36 | 0.36 | 1.24 | 1.24 | 0.36 | 1.24 | 1.24 | 1.24 |
| Modified intercept | 4.52 | 4.52 | 4.92 | 4.92 | 0.77 | 0.77 | 4.92 | 0.77 | 0.77 | 0.77 |

The sum of *A*'s modified slopes is 0.5+0.5=1, and the sum of *A*'s modified intercepts is 4.52+4.52=9.04. In the same way, the slope sum and intercept sum of *C* are respectively

0.36+0.36+0.36=1.08 and 4.92+4.92+4.92=14.76. And *D*'s are 6.2 and 3.85. It is seen that the other amino acids, such as *E*, *F*, *G*, *H*, ..., *W*, *Y*, do not appear in the segment. So, we then write their slope sums and intercept sums as 0. Finally, the 40-component characterizing vector is established as $(1, 9.04, 1.08, 14.76, 6.2, 3.85, 0, 0, 0, 0, \ldots, 0, 0)$.

## 3.2 Similarity distance

Similarity distance is the primary indicator of the similarity/dissimilarity between two sequences. It is the basis of evolutionary and phylogenetic analysis. Generally, the more similar two sequences are, the smaller the distance between them is. Distance calculation is implemented on the premise of aligning the two sequences to be compared, which may be sort of difficult. However, given two proteins of different lengths, $P_1$ and $P_2$, we are able to conduct the sequence comparison directly based on the proposed 40-component characterizing vector.

Given two characterizing vectors of equal length, researchers normally take two indicators to represent the pair-wise distance, the Euclidean distance and the Cosine distance. The Euclidean indicator is seriously sensitive to the lengths of the sequences and does not have a definite scope, and that can be an obstacle when dealing with the long proteins. However, the cosine distance between characterizing vectors can eliminate the effects of the lengths of proteins. On the other hand, once we normalize the cosine value from (1,-1) to (0,1), the cosine distance owns clear value range and carries more evolutionary meaning.

Consequently, we choose the Cosine function here as a measure to indicate the similarity/dissimilarity between two protein sequences. If the two 40-component characterizing vector are denoted as

$$V_1 = (\tilde{S}_1^A, \tilde{I}_1^A, \tilde{S}_1^C, \tilde{I}_1^C, ..., \tilde{S}_1^Y, \tilde{I}_1^Y)$$

$$V_2 = (\tilde{S}_2^A, \tilde{I}_2^A, \tilde{S}_2^C, \tilde{I}_2^C, ..., \tilde{S}_2^Y, \tilde{I}_2^Y),$$

the Cosine value between the two vectors is calculated as

$$\cos(V_1, V_2) = \frac{V_1 \cdot V_2}{|V_1||V_2|} = \frac{\sum_{aa=A}^{Y} (\tilde{S}_1^{aa} \cdot \tilde{S}_2^{aa} + \tilde{I}_1^{aa} \cdot \tilde{I}_2^{aa})}{\sqrt{\sum_{aa=A}^{Y} (\tilde{S}_1^{aa\,2} + \tilde{I}_1^{aa\,2})} \sqrt{\sum_{aa=A}^{Y} (\tilde{S}_2^{aa\,2} + \tilde{I}_2^{aa\,2})}} \tag{2}$$

where *aa* denotes one of the twenty amino acids arranged in alphabetical order. To give a more reasonable explanation on evolution, the domain scale of the variant is modified from (1,-1) to (0,1) by the following formula,

$$D(V_1,V_2)=\frac{1-\cos(V_1,V_2)}{2} \tag{3}$$

In the traditional matrix invariant techniques, the construction of the $n\times n$ matrix characterizing a sequence needs an $O(n^2)$ computational complexity so that the computing time and space own a higher demand, which can be a serious obstacle when the length $n$ is bigger. However, through the new numerical characterization method, the computational complexity reduces to $O(n)$. This will greatly reduce the computing needs of time and space, especially when we have a large value of sequence length.

# 4 Applications

## 4.1 Application on the similarity research of the nine ND5 proteins

In order to test the effectiveness of the proposed sequence comparison scheme, we apply it to the similarity research of nine ND5 proteins: human (Homo sapiens, AP_000649), gorilla (Gorilla gorilla, NP_008222), common chimpanzee (Pan troglodytes, NP_008196), pigmy chimpanzee (Pan paniscus, NP_008209), fin whale (Balenoptera physalus, NP_006899), blue whale (Balenoptera musculus, NP_007066), rat (Rattus norvegicus, AP_004902), mouse (Mus musculus, NP_904338), and opossum (Didelphis virginiana, NP_007105), which can be downloaded from NCBI Genbank.

First, nine ND5 protein sequences are transformed into digital sequences by the 2D map of twenty amino acids and their modified slopes and intercepts in the linear regression models based on the physicochemical properties. Each amino acid in the sequence can be expressed by a 2-tuple. For example, the map of the first ten amino acids in the human ND5 protein is shown in Table 4.

**Table 4** First ten amino acids in the human ND5 protein and their corresponding 2D map

| Amino acid( $a_i$ ) | *M* | *T* | *M* | *H* | *T* | *T* | *M* | *T* | *T* | *L* |
|---|---|---|---|---|---|---|---|---|---|---|
| Modified Slope( $S_i$ ) | 0.36 | 0.75 | 0.36 | 1.02 | 0.75 | 0.75 | 0.36 | 0.75 | 0.75 | 0.22 |
| Modified intercept( $I_i$ ) | 4.83 | 3.18 | 4.83 | 1.81 | 3.18 | 3.18 | 4.83 | 3.18 | 3.18 | 5.73 |

The 3D visualization of the nine sequences is implemented by using Eq. (1). Among them, the 3D graphical representations of the first one hundred amino acids in Human, Gorilla and Rat are shown in Fig. 4. It is seen under a coarse level that the pair (Human and Gorilla) have a relatively high similarity degree among three species, which generally agrees with the biological facts.

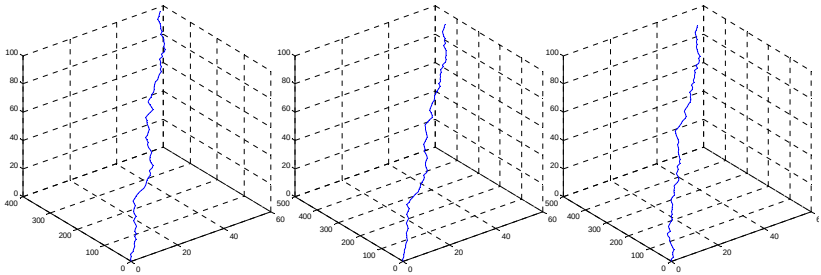

**Figure 4** The 3D zigzag representations of the first 100 amino acids of ND5 proteins from Human, Gorilla and Rat

**Table 5** Construction of the 40-component characterizing vector for Human ND5 protein sequence with a length of 603

| Amino acid | Slope ($S^{aa}$) | Intercept ($I^{aa}$) | $\tilde{S}^{aa} = \sum S^{aa}$ | $\tilde{I}^{aa} = \sum I^{aa}$ |
|---|---|---|---|---|
| *A* | 0.50 | 4.52 | 22.00 | 198.88 |
| *C* | 0.36 | 4.92 | 2.16 | 29.52 |
| *D* | 1.24 | 0.77 | 13.64 | 8.47 |
| *E* | 1.21 | 0.92 | 10.89 | 8.28 |
| *F* | 0.23 | 5.44 | 8.74 | 206.72 |
| *G* | 0.72 | 3.35 | 18.72 | 87.10 |
| *H* | 1.02 | 1.81 | 14.28 | 25.34 |
| *I* | 0.04 | 6.81 | 2.16 | 367.74 |
| *K* | 1.26 | 1.05 | 26.46 | 22.05 |
| *L* | 0.22 | 5.73 | 22.88 | 595.92 |
| *M* | 0.36 | 4.83 | 9.36 | 125.58 |
| *N* | 1.05 | 1.50 | 33.60 | 48.00 |
| *P* | 0.84 | 2.66 | 27.72 | 87.78 |
| *Q* | 1.02 | 1.48 | 20.40 | 29.60 |
| *R* | 1.30 | 0.86 | 10.40 | 6.88 |
| *S* | 0.69 | 3.31 | 33.81 | 162.19 |
| *T* | 0.75 | 3.18 | 48.75 | 206.70 |
| *V* | 0.21 | 5.98 | 3.15 | 89.70 |
| *W* | 0.60 | 3.53 | 7.20 | 42.36 |
| *Y* | 0.61 | 3.33 | 9.76 | 53.28 |

The 3D visualizing representation can roughly show the similarity among several sequences. However, when there are lots of sequences to be compared and a need for accurate description of the similarities, we need to go a step further to conduct a similarity distance calculation among

sequences. The proposed 40-component characterizing vector, of which the component is the sum of one amino acid's modified slope or intercept in the sequence, is set up for each sequence in the above describing. In Table 5, we list the 40-component characterizing vector of the Human ND5 protein sequence with a length of 603.

Next, the similarity matrix of the ND5 sequences from nine species is listed in Table 6 the equations, Eq. (2) and Eq. (3). The experimental results demonstrate the performance of the proposed scheme. One can easily identify the eight entries that are smaller than the rest after a superficial glance at the similarity matrix. They are summarized as: six entries corresponding to the four primate species (*Human*, *Gorilla*, *C.Chim.* and *P.Chim.*), (*F.Whale* - *B.Whale*) and (*Rat* - *Mouse*).

**Table 6** Similarity matrix of the ND5 proteins from nine species

| species | *Gorilla* | *C.Chim.* | *P.Chim.* | *F.Whale* | *B.Whale* | *Rat* | *Mouse* | *Opossum* |
|---|---|---|---|---|---|---|---|---|
| *Human* | **0.00039** | **0.00022** | **0.00045** | 0.00122 | 0.00142 | 0.00767 | 0.01056 | 0.01408 |
| *Gorilla* | | **0.00052** | **0.00066** | 0.00107 | 0.00132 | 0.00714 | 0.00996 | 0.01283 |
| *C.Chim.* | | | **0.00020** | 0.00147 | 0.00167 | 0.00709 | 0.01040 | 0.01371 |
| *P.Chim.* | | | | 0.00107 | 0.00127 | 0.00600 | 0.00931 | 0.01167 |
| *F.Whale* | | | | | **0.00010** | 0.00660 | 0.00832 | 0.01073 |
| *B.Whale* | | | | | | 0.00618 | 0.00759 | 0.01071 |
| *Rat* | | | | | | | **0.00092** | 0.00310 |
| *Mouse* | | | | | | | | 0.00362 |

Observing Table 6, one can note that the entry *Human* - *C.Chim.* is 0.00022 and *Human* - *Gorilla* is 0.00039, which indicates that human and common chimpanzee have a relatively close relationship from evolutionary point of view, and the same conclusion is seen in recent study [23]. Besides, it is not accidental that the entries related to *Opossum* are visibly different from other entries. This phenomenon confirms the opossum is genetically unique among the nine species. As is shown above, the similarity results are basically consistent with the known evolutionary fact and the recent studies on the experimental dataset [19-23].

## 4.2 Application on the phylogenetic analysis of 36 proteins

In order to further test the validity of the proposed method, we use the method to analyze another data set, the Chew-Kedem dataset consisting of 36 proteins, which has been introduced in Refs. [38-40]. All of the proteins identified by their corresponding PDB id are drawn from the PDB

entries of five families, Globin, Alpha, Beta, Alpha-beta and Tim-barrels. The detailed classification is shown as follows,

Goblin (denoted by ‘-’), 1eca, 5mbn,1hlb, 1hlm, 1babA, 1babB, 1ithA, 1mba, 2hbg, 2lhb, 3sdhA, 1ash, 1flp, 1myt, 1lh2, 2vhbA, and 2vhb. Alpha (denoted by ‘/’), 1cnp and 1jhg. Beta (denoted by ‘*’), 1cd8, 1ci5, 1qa9, 1cdb, 1neu, 1qfo and 1hnf. Alpha-beta (denoted by ‘+’), 1aa9, 1gnp, 6q21, 1ct9, 1qra and 5p21. Tim-barrels (denoted by ‘O’), 6xia, 2mnr, 1chr and 4enl.

Among them, protein 2vhb and 2vhbA are the same sequence to check whether the new approach can cluster the two identical proteins into one class.

Fig. 5(a) is the phylogenetic tree for the 36 proteins based on the proposed scheme by using UPMGA method. As a comparison, the phylogenetic tree constructed by using ClustalW method in Mega 6 [41] is shown in Fig. 5(b).

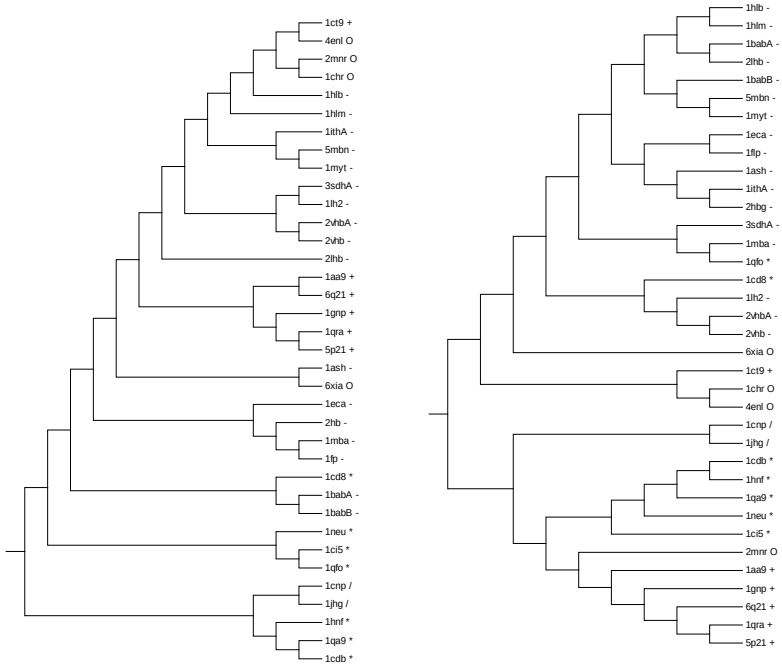


**Figure 5** (a) Phylogenetic tree of 36 proteins by the proposed approach; (b) Phylogenetic tree of 36 proteins by ClustalW

One can find the high similarity between the two figures after careful observation. For instance, in both figures, most of the 36 proteins are classified into the right branches except 1ct9, 6xia and 1cd8. The Alpha family is closer to the Beta family. The Tim-barrels is closer to the Goblin. However, in Fig. 5 there seems to be misplaced proteins such as 2mnr and 1qfo. What's more, the Beta family in Fig. 5 (a) is clustered into only two branches but scatters separately in Fig. 5(b). The result shows that the proposed approach is a valid method in phylogenetic analysis.

## 5 Conclusions

In this paper, we provided a novel protein map method by using the simple linear regression model of eight physicochemical properties of amino acids. Based on the protein map, we construct a corresponding 3D visualizing method, which can roughly indicate the similarity between sequences. Further more, on the basis of the protein map, we set up a 40-component vector to characterize a protein. Similarity research and phylogenetic analysis are implemented by converting each protein sequence to its corresponding 40-component vector. Experiments on two real datasets shows the validity and effectiveness of the new scheme comparing to the related studies and the biological facts.

*Acknowledgment*: This work supported by the National Natural Science Foundation of China (Grant No. 61272254), and by the Natural Science Foundation of Hebei Province, China (Project No. F2012210017), and by the Humanities and Social Sciences Research of Ministry of Education of China (Project name, The Origin, Propagation and Migration of Human Influenza Epidemic (1918-2010) from Space-time Perspective; Project No. 11YJCZH132).

## References

[1] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318-1327.

[2] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol*. **119** (1986) 319-328

[3] A. Nandy, A new graphical representation and analysis of DNA sequence structure. I: Methodology and application to globin genes, *Curr. Sci.* **66** (1994) 309-314.

[4] X. Guo, M. Randić, S. C. Basak, A novel 2-D graphical representation of DNA sequences of low degeneracy, *Chem. Phys. Lett.* **350** (2001) 106-112.

[5] C. T. Zhang, R. Zhang, H. Y. Ou, The Z curve database: a graphic representation of genome sequences, *Bioinformatics* **19** (2003) 593-599.

[6] Z. H. Qi, T. R. Fan, PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **442** (2007) 434-440.

[7] B. Liao, X. Xiang, W. Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence, *J. Comput. Chem.* **27** (2006) 1196-1202.

[8] Z. H. Qi, X. Q. Qi, C. C. Liu, New method for global alignment of 2 DNA sequences by the tree data structure, *J. Theor. Biol*. **263** (2010) 227-236.

[9] Y. Yang, Y. Zhang, M. Jia, Non-degenerate graphical representation of DNA sequences and its applications to phylogenetic analysis, *Comb. Chem. High T.* Scr. **16** (2013) 585-589.

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

[11] M. Randić, K. Mehulić, D. Vukičević, T. Pisanski, D. Vikić-Topić, D. Plavšić, Graphical representation of proteins as four-color maps and their numerical characterization, *J. Mol. Graph. Model.* **27** (2009) 637-641.

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

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

[14] L. Zhang, B. Liao, D. Li, W. Zhu, A novel representation for apoptosis protein subcellular localization prediction using support vector machine, *J. Theor. Biol*. **259** (2009) 361-365.

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

[16] M. Randić, J. Zupan, A.T. Balaban, Unique graphical representation of protein sequences based on nucleotide triplet codons, *Chem. Phys. Lett.* **397** (2004) 247-252.

[17] B. Liao, B. Y. Liao, X. M. Sun, Q. G. Zeng, A novel method for similarity analysis and protein sub-cellular localization prediction, *Bioinformatics* **16** (2010) 2678-2683.

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

[19] Y. H. Yao, Q. Dai, C. Li, P. A. He, X. Y. Nan, Y. Z. Zhang, Analysis of similarity/dissimilarity of protein sequences, *Proteins* **73** (2008) 864-871.

[20] S. S. T. Yau, C. Yu, R. He, A protein map and its application, *DNA Cell Biol*. **27** (2008) 241-250.

[21] Z. H. Qi, M. Z. Jin, S. L. Li, J. Feng, A protein mapping method based on physicochemical properties and dimension reduction, *Comput. Biol. Med.* **57** (2015) 1-7.

[22] Y. Yao, S. Yan, J. Han, Q. Dai, P. A. He, A novel descriptor of protein sequences and its application, *J. Theor. Biol*. **347** (2014) 109-117.

[23] P. A. He, J. Wei, Y. Yao, Z. Tie, A novel graphical representation of proteins and its application, *Physica A* **391** (2012) 93-99.

[24] X. Xia, W. H. Li, What amino acid properties affect protein evolution? *J. Mol. Evol.* **47** (1998) 557-564.

[25] M. K. Gupta, R. Niyogi, M. Misra, A 2D graphical representation of protein sequence and their similarity analysis with probabilistic method, *MATCH Commun. Math. Comput. Chem.* **72** (2014) 519-532.

[26] H. Hu, *F*-curve, A graphical representation of protein sequences for similarity analysis based on physicochemical properties of amino acids, *MATCH Commun. Math. Comput. Chem.* **73** (2015) 749-761.

[27] C. R. Woese, D. H. Dugre, S. A. Dugre, M. Kondo, W. C. Saxinger, On the fundamental nature and evolution of the genetic code, *Cold. Spring. Harb. Sym.* **31** (1996) 723-736.

[28] C. Alff–Steinberger, The genetic code and error transmission, *Proc. Natl. Acad. Sci. USA* **64** (1996) 584-591.

[29] P. H. A. Sneath, Relations between chemical structure and biological activity in peptides, *J. Theor. Biol.* **12** (1966) 157-195.

[30] J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* **157** (1982) 105-132.

[31] Y. Zhao, X. Li, Z. Qi, Novel 2D graphic representation of protein sequence and its application, *J. Fib. Bioen. Inform.* **7** (2014) 23-33.

[32] A. T. Balaban, D. Plavšić, M. Randić, DNA invariants based on non-overlapping triplets of nucleotide bases, *Chem. Phys. Lett.* **379** (2003) 147-154.

[33] M. Randić, A. F. Kleiner, L. M. De Alba, Distance/distance matrixes, *J. Chem. Inform. Comp. Sci.* **34** (1994) 277-286.

[34] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **368** (2003) 1-6.

[35] Z. H. Qi, L. Li, X. Q. Qi, Using Huffman coding method to visualize and analyze DNA sequences, *J. Comput. Chem.* **32** (2011) 3233-3240.

[36] Z. H. Qi, X. Q. Qi, Novel 2D graphical representation of DNA sequence based on dual nucleotides, *Chem. Phys. Lett.* **440** (2007) 139-144.

[37] Z. H. Qi, M. Z. Jin, Y Hong, A measure of protein sequence characteristics based on the frequency and the position entropy of existing *K*-words, *MATCH Commun. Math. Comput. Chem.* **73** (2015) 731-748.

[38] Z. Mu, J. Wu, Y. Zhang, A novel method for similarity/dissimilarity analysis of protein sequences, *Physica A* **392** (2013) 6361-6366.

[39] S. Zhang, L. Yang, T. Wang, Use of information discrepancy measure to compare protein secondary structures, *J. Mol. Struct. (Theochem)* **909** (2009) 102-106.

[40] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, G. Valiente, Compression-based classification of biological sequences and structures via the universal similarity metric: experimental assessment, *BMC Bioinformatics* **8** (2007) #252 (20 pp.).

[41] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: Molecular evolutionary genetics analysis version 6.0, *Mol. Biol. Evol.* **30** (2013) 2725-2729.