Novel Method of 3-Dimensional Graphical Representation for Proteins and Its Application

Zhao-Hui Qi1, Ke-Cheng Li1, Jin-Long Ma1, Yu-Hua Yao2 and Ling-Yun Liu1

1School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, Republic of China. 2School of Mathematics and Statistics, Hainan Normal University, Haikou, Republic of China.

ABSTRACT: In this article, we propose a 3-dimensional graphical representation of protein sequences based on 10 physicochemical properties of 20 amino acids and the BLOSUM62 matrix. It contains evolutionary information and provides intuitive visualization. To further analyze the similarity of proteins, we extract a specific vector from the graphical representation curve. The vector is used to calculate the similarity distance between 2 protein sequences. To prove the effectiveness of our approach, we apply it to 3 real data sets. The results are consistent with the known evolution fact and show that our method is effective in phylogenetic analysis.

KEYWORDS: protein sequences, graphical representation, physicochemical properties, BLOSUM62 matrix

Introduction

With the number of available biological sequences developing rapidly, how to mine essential information from a huge amount of biological sequences effectively and reliably has become a critical problem. As a result, many methods in information extraction are proposed by researchers. Among them, the graphical representation of DNA sequences is an effective method for the virtualization and similarity analysis. Graphical representation is a kind of alignment-free method. It provides intuitive information of data by visualization of biological sequences. What is more, it is more generally applicable because its mathematical description of data facilitates numerical analysis without difficult calculations. Therefore, numerous works based on graphical representation have been presented by researchers.1–8 For example, Randić et al1 proposed a graphical representation of RNA secondary structure based on twelve symbols. Bielińska-Waż et al2 proposed a 2D-dynamic representation of DNA sequences in 2007. After that they proposed more dynamic representations of DNA sequences for generalization.6,7

However, the graphical representation of protein sequences is much more difficult because there are 20 amino acids instead of 4 nucleotides. Various approaches have been proposed by researchers only until recently.9–16 Among them, many approaches are based primarily on the physicochemical properties of amino acids. Randić2 early proposed a 2-dimensional graphical representation of proteins based on a pair of physicochemical properties in 2007. After that, Yu et al11 proposed a protein mapping method of protein sequences based on 10 physicochemical properties. Wang et al10 presented a graphical representation of protein sequences based on 9 physicochemical properties. In the works by He et al15 and Hu,16 the physicochemical properties are also indispensable in information extraction from proteins because they have effects on the rate and pattern of protein evolution. From these, we can see that physicochemical properties are widely applied with graphical representation of protein sequences by these researchers and their results seem well.

In this article, we propose a 3-dimensional (3D) graphical representation of protein sequences based on 10 physicochemical properties17–21 of amino acids and the BLOSUM62 matrix.22 The representation can provide good visualization without degeneracy or circuit. Then, we extract a specific vector from the graphical curve of a protein sequence. In addition, we proposed 2 applications based on the vector to analyze the similarity and evolutionary relationship of 3 data sets, respectively. The results are consistent with the evolution fact and works by other researchers. This shows our approach can be applied to hundreds of sequences with different lengths and perform well.

Methods

As we know, a protein sequence is usually composed of 20 kinds of amino acids. Every amino acid has its own particular physicochemical properties. Therefore, to mine essential information from a protein sequence, we propose an effective graphical method combining physicochemical properties of amino acids and the BLOSUM62 matrix.

**BLOSUM62 matrix**

BLOSUM62 matrix by Henikoff and Henikoff22 is a substitution matrix applied to the alignment of protein sequences. The values of the BLOSUM62 matrix represent the probability...
that one amino acid is replaced by other amino acids. In their scoring scheme, a positive score represents a higher similarity between 2 amino acids and a negative score represents a lower similarity.

Physicochemical properties of amino acids

Here, we consider 10 primary physicochemical properties of amino acids, such as the pK1 (–COOH),\textsuperscript{17} the pK2 (–NH3),\textsuperscript{21} the polar requirement,\textsuperscript{21} the isoelectric point,\textsuperscript{18} the hydrogenation,\textsuperscript{20} the hydroxythiolation,\textsuperscript{20} the molecular volume,\textsuperscript{19} the aromaticity,\textsuperscript{20} the aliphaticity,\textsuperscript{20} and the polarity values.\textsuperscript{19} The 10 physicochemical properties of 20 amino acids are shown in Table 1.

The 3D graphical representation of protein sequences based on Blosum62 matrix and physicochemical properties of amino acids

For each physicochemical property, we will use $K$-means clustering method\textsuperscript{21} to classify the 20 amino acids into several groups. $K$-means clustering is an efficient unsupervised clustering method which is widely used in a diverse range of fields such as data mining, bioinformatics, and natural language processing.\textsuperscript{24} However, there are some weaknesses in $K$-means. $K$-means needs to be given the number of clusters beforehand. Silhouette\textsuperscript{25} is a cluster validity index that can be used to determine the number of clusters. It considers 2 factors: cohesion and separation. Its value ranges from −1 to 1 and a higher value represents a better effect of clustering. According to this index, we can obtain a valid number of clusters of the given data set. In this way, we can obtain 10 kinds of clustering classification based on the 10 different properties, which are shown in Table 2.

According to the property pK1 (–COOH), we can divide the 20 amino acids into 7 groups: G1 (A, G, I, L, M, W, V), G2 (H, F), G3 (Q, E, K, S, Y), G4 (T), G5 (N, P), G6 (C), and G7 (D). If 2 or more amino acids are divided into the same group, it denotes that they are similar to each other by the property pK1 (–COOH). Taking all the properties into consideration, we can obtain the number of similar properties between each pair of amino acids.

If $X$ denotes an amino acid and $Y$ denotes another amino acid, then we define the similar degree of 2 amino acids $S_{XY}$ as follows:

$$S_{XY} = \frac{\text{pro1}_X \cdot \text{pro1}_Y + \text{pro2}_X \cdot \text{pro2}_Y + \text{pro3}_X \cdot \text{pro3}_Y + \text{pro4}_X \cdot \text{pro4}_Y + \text{pro5}_X \cdot \text{pro5}_Y + \text{pro6}_X \cdot \text{pro6}_Y + \text{pro7}_X \cdot \text{pro7}_Y + \text{pro8}_X \cdot \text{pro8}_Y + \text{pro9}_X \cdot \text{pro9}_Y + \text{pro10}_X \cdot \text{pro10}_Y}{\sqrt{\text{pro1}_X^2 \cdot \text{pro1}_Y^2 + \text{pro2}_X^2 \cdot \text{pro2}_Y^2 + \text{pro3}_X^2 \cdot \text{pro3}_Y^2 + \text{pro4}_X^2 \cdot \text{pro4}_Y^2 + \text{pro5}_X^2 \cdot \text{pro5}_Y^2 + \text{pro6}_X^2 \cdot \text{pro6}_Y^2 + \text{pro7}_X^2 \cdot \text{pro7}_Y^2 + \text{pro8}_X^2 \cdot \text{pro8}_Y^2 + \text{pro9}_X^2 \cdot \text{pro9}_Y^2 + \text{pro10}_X^2 \cdot \text{pro10}_Y^2}}$$

| AMINO ACID | PRO1 | PRO2 | PRO3 | PRO4 | PRO5 | PRO6 | PRO7 | PRO8 | PRO9 | PRO10 |
|------------|------|------|------|------|------|------|------|------|------|-------|
| A (Ala)    | 2.34 | 9.69 | 7.0  | 6.00 | 0.33 | −0.062| 31   | −0.11 | 0.239| 8.1   |
| R (Arg)    | 2.17 | 9.04 | 9.1  | 10.76| −0.176| −0.167| 124  | 0.079 | 0.211| 10.5  |
| N (Asn)    | 2.02 | 8.80 | 10.0 | 5.41 | −0.233| 0.166 | 56   | −0.136| 0.249| 11.6  |
| D (Asp)    | 2.09 | 9.82 | 13.0 | 2.77 | −0.371| −0.079| 54   | −0.285| 0.171| 13.0  |
| C (Cys)    | 1.71 | 10.78| 4.8  | 5.07 | 0.074 | 0.38  | 55   | −0.184| 0.22 | 5.5   |
| Q (Gln)    | 2.17 | 9.13 | 8.6  | 5.65 | −0.409| −0.025| 85   | −0.246| 0.26 | 10.5  |
| E (Glu)    | 2.19 | 9.67 | 12.5 | 3.22 | −0.254| −0.184| 83   | −0.067| 0.187| 12.3  |
| G (Gly)    | 2.34 | 9.60 | 7.9  | 5.97 | 0.37  | −0.017| 3    | −0.073| 0.16 | 9.0   |
| H (His)    | 1.82 | 9.17 | 8.4  | 7.59 | −0.078| 0.056 | 96   | 0.32  | 0.205| 10.4  |
| I (Ile)    | 2.36 | 9.68 | 4.9  | 6.02 | 0.149 | −0.309| 111  | 0.001 | 0.273| 5.2   |
| L (Leu)    | 2.36 | 9.60 | 4.9  | 5.98 | 0.129 | −0.264| 111  | −0.008| 0.281| 4.9   |
| K (Lys)    | 2.18 | 8.95 | 10.1 | 9.74 | −0.075| −0.371| 119  | 0.049 | 0.228| 11.3  |
| M (Met)    | 2.28 | 9.21 | 5.3  | 5.74 | −0.092| 0.077 | 105  | −0.041| 0.253| 5.7   |
| F (Phe)    | 1.83 | 9.13 | 5.0  | 5.48 | 0.011 | 0.074 | 132  | 0.438 | 0.234| 5.2   |
| P (Pro)    | 1.99 | 10.60| 6.6  | 6.30 | 0.37  | −0.036| 32.5 | −0.016| 0.165| 8.0   |
| S (Ser)    | 2.21 | 9.15 | 7.5  | 5.68 | 0.022 | 0.47  | 32   | −0.153| 0.236| 9.2   |
| T (Thr)    | 2.63 | 10.43| 6.6  | 6.16 | 0.136 | 0.348 | 61   | −0.208| 0.213| 8.6   |
| W (Trp)    | 2.38 | 9.39 | 5.2  | 5.89 | 0.011 | 0.05  | 170  | 0.493 | 0.183| 5.4   |
| Y (Tyr)    | 2.20 | 9.11 | 5.4  | 5.66 | −0.138| 0.22  | 136  | 0.381 | 0.193| 6.2   |
| V (Val)    | 2.32 | 9.62 | 5.6  | 5.96 | 0.245 | 0.212 | 84   | −0.155| 0.255| 5.9   |

Table 1. Numerical values about 10 physicochemical properties of 20 amino acids.

Pro1, the pK1 (–COOH); pro2, the pK2 (–NH3); pro3, the polar requirement; pro4, the isoelectric point; pro5, the hydrogenation; pro6, the hydroxythiolation; pro7, the molecular volume; pro8, the aromaticity; pro9, the aliphaticity; and pro10, the polarity values.
where $N_{XY}$ is the number of similar properties between amino acid $X$ and $Y$. $B_{XY}$ is the value of amino acid $X$ and $Y$ in the Blosum62 matrix. Then, we calculate all the values of $S_{XY}$ and the result is shown in Table 3.

From Table 3, we can find that the similarity degree $S_{XY}$ of different amino acids can be numerically different from each other. To describe the similarity degree graphically, we will use a unitary linear regression to extract characters from every amino acid. Here, we take $I = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$ as independent variables of linear regression. For amino acid $X$, we take the 19 values in its corresponding row in Table 3 as dependent variables. Using unitary linear regression, we can obtain the corresponding slope and intercept of amino acid $X$. The slope and intercept can describe amino acid $X$ effectively. All the slopes and intercepts of 20 amino acids are given in Table 4.

We assume that $P = p_1, p_2, \ldots, p_n$ is an arbitrary protein sequence composed of $n$ amino acids. If $x_i$, $y_i$, and $z_i$ represent the 3D coordinates of $p_i$ in the protein sequence, then the 3D representation of a protein sequence will be constructed as follows:

$$p_i = \begin{cases} x_i = x_{i-1} + a_i \\ y_i = y_{i-1} + b_i \\ z_i = i \end{cases}$$

where $a_i$ and $b_i$ represent the slope and intercept of $p_i$. In addition, the initial condition is $x_0 = y_0 = 0$. Next, we can convert the $n$ points into a graphical curve.

To demonstrate the effectiveness of the 3D graphical method, we take 2 protein sequences as an example. Both the sequences are taken from yeast *Saccharomyces cerevisiae*. The graphical representations of 2 protein sequences are shown in Figure 1.

### Protein I

WTFESRNDPAPIILWNLNGPGCSSLTGL

### Protein II

WFFESRNDPAPILWNLNGPGCSSFTGL

As can be seen from Figure 1, the 2 curves are similar to each other. Furthermore, we can see that there are some differences between 2 figures in $p_{12}$, $p_{11}$, $p_{14}$, and $p_{27}$. This indicates the effectiveness of the proposed 3D graphical representation method.

### Numerical Characterization and Similarity Analysis of Proteins

Based on the constructed graphical curve, we can get a specific vector from a protein sequence. Using this vector we can analyze the similarity between 2 protein sequences effectively.

#### 40-dimensional characteristic vector

Characteristic vector is a common method to calculate the pairwise distance between 2 protein sequences. A good characteristic vector should avoid the problem about different lengths of sequences and complicated calculation.

Here, we define a 2-tuple $(\overline{a}_X, \overline{b}_X)$ of amino acid $X$ for characterization. Given a protein sequence composed of $n$ amino acids $P = p_1, p_2, \ldots, p_n$, we can compute the 2-tuple as follows:

$$\overline{a}_X = a_X \times \frac{N_X}{n}$$
$$\overline{b}_X = b_X \times \frac{N_X}{n}$$

where $a_X$ and $b_X$ are, respectively, slope and intercept of amino acid $X$ in Table 4. $N_X$ is the number of amino acid $X$ in the sequence. From equation (3), we can see that $N_X / n$
Table 3. The similar degree of each pair of amino acids.

| AMINO ACID | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A (Ala)    | −1| −4| −4| 0 | −2| −2| 0 | −2| −4| −3| −2| −3| −4| −4| −4| 5 | 0 | −6| −2| 0 |   |
| R (Arg)    | −1| 0 | −2| −6| 3 | 0 | −2| 0 | −3| −2| 10 | −1| 0 | −2| −2| −2| 0 | −4| −3 |   |
| N (Asn)    | −4| 0 | 2 | −9| 0 | 0 | 0 | 1 | −6| −6| 0 | −6| −3| −6| 2 | 0 | −4| −2| −9 |   |
| D (Asp)    | −4| −2| 2 | −6| 0 | 6 | −2| 0 | −3| −4| −1| −3| 0 | −2| 0 | −2| 0 | 0 | 0 | −3 |   |
| C (Cys)    | 0 | −6| −9| −6| −6| −6| −6| −6| −4| −3| −3| −6| −6| −2| −5| −4| −4| −3 |   |
| Q (Gln)    | −2| 3 | 0 | 0 | −6| 6 | −6| 0 | −6| −4| 2 | 0 | −6| −3| 0 | −2| −2| −3| −8 |   |
| E (Glu)    | −2| 0 | 6 | −4| 6 | −2| 0 | −6| −3| 2 | −2| 0 | −1| 0 | −1| −3| −4| −4 |   |
| G (Gly)    | 0 | −2| 0 | −2| −6| −6| −2| −2| −12| −16| −2| −9| −3| −12| 0 | −6| −4| −3| −12 |   |
| H (His)    | −2| 0 | 1 | 0 | −6| 0 | 0 | −2| −3| −3| −1| −8| −5| −2| −2| −4| −6| 6 | −3 |   |
| I (Ile)    | −4| −3| −6| −3| −4| −6| −6| −12| −3| 18 | −3| 5 | 0 | −6| −4| −3| −9| −2| 12 |   |
| L (Leu)    | −3| −2| −6| −4| −4| −3| −16| −3| 18 | −2| 10 | 0 | −6| −4| −3| −6| −2| 5 |   |
| K (Lys)    | −2| 10| 0 | −1| −3| 2 | 2 | 2 | −1| −3| −2| −2| −3| −1| 0 | −1| 0 | −2| −2 |   |
| M (Met)    | −3| −1| −6| −3| −3| 0 | −2| −9| −8| 5 | 10 | −2| 0 | −4| −3| −2| −5| −4| 6 |   |
| F (Phe)    | −4| 0 | −3| 0 | −6| −6| 0 | −3| −5| 0 | 0 | −3| 0 | −4| −8| −2| 5 | 15 | −2 |   |
| P (Pro)    | −4| −2| −6| −6| −3| −1| −12| −2| −6| −6| −1| −4| −4| −4| −4| −4| −3| −4 |   |
| S (Ser)    | 5 | −2| 2 | 0 | −2| 0 | 0 | 0 | −2| −4| −4| 0 | −3| −8| −4| 3 | −6| −6| −4 |   |
| T (Thr)    | 0 | −2| 0 | −2| −5| −2| −1| −6| −4| −3| −3| −1| −2| −2| −4| 3 | −2| −2| 0 |   |
| W (Trp)    | −6| 0 | −4| 0 | −4| −2| −3| −4| −6| −9| −6| 0 | −5| 5 | −4| −6| −2| 10| −12 |   |
| Y (Tyr)    | −2| −4| −2| 0 | −4| −3| −4| −3| 6 | −2| −2| −2| −4| 15 | −3| −6| −2| 10| −4 |   |
| V (Val)    | 0 | −3| −9| −3| −3| −8| −4| −12| −3| 12 | 5 | −2| 6 | −2| −4| −4| 0 | −12| −4 |   |
indicates the proportion of the amino acid \( X \) in the whole sequence. Taking proportion into consideration, we can eliminate the effects of the lengths of proteins. Thus, for each kind of amino acid, we get a 2-tuple for characterization. As there are 20 amino acids, we can construct a 40-dimensional characteristic vector \( \left( a_1, a_2, a_3, \ldots, a_{20}, b_1, b_2, \ldots, b_{19}, b_{20} \right) \). The component order is the same with Table 4.

Taking a short segment of 10 amino acids, AARRARRNNN, as an example, the numbers of amino acid A, R, and N in the segment are 3, 4, and 3. Therefore, we can obtain the 40-dimensional characterizing vector \((0.015, -0.738, -0.02, -0.16, -0.042, -0.369, 0, 0, \ldots, 0, 0)\) according to Table 4 and equation (3).

### Similarity analysis

The similarity/dissimilarity between 2 protein sequences can be represented by similar distance. There are several calculating methods for measurement of similar distance such as Euclidian distance, City Block distance, and Manhattan distance. Here, we use Euclidian distance as a measure to represent the similarity/dissimilarity between 2 sequences. We will compute the similarity distance using the 40-dimensional characteristic vector. If the two 40-dimensional characteristic vectors are denoted as

| AMINO ACID (X) | SLOPE (A) | INTERCEPT (B) | LINEAR EQUATION |
|---------------|-----------|---------------|-----------------|
| A (Ala)       | 0.05      | −2.46         | \( y = 0.05x + (-2.46) \) |
| R (Arg)       | −0.05     | −0.4          | \( y = -0.05x + (-0.4) \) |
| N (Asn)       | −0.14     | −1.23         | \( y = -0.14x + (-1.23) \) |
| D (Asp)       | 0.01      | −1.35         | \( y = 0.01x + (-1.35) \) |
| C (Cys)       | 0.09      | −5.44         | \( y = 0.09x + (-5.44) \) |
| Q (Gln)       | −0.22     | 0.26          | \( y = -0.22x + (0.26) \) |
| E (Glu)       | −0.19     | 1.0           | \( y = -0.19x + (1.0) \) |
| G (Gly)       | −0.3      | −2.25         | \( y = -0.3x + (-2.25) \) |
| H (His)       | −0.08     | −1.28         | \( y = -0.08x + (-1.28) \) |
| I (Ile)       | 0.32      | −5.26         | \( y = 0.32x + (-5.26) \) |
| L (Leu)       | 0.23      | −4.16         | \( y = 0.23x + (-4.16) \) |
| K (Lys)       | −0.19     | 1.33          | \( y = -0.19x + (1.33) \) |
| M (Met)       | 0.16      | −3.4          | \( y = 0.16x + (-3.4) \) |
| F (Phe)       | 0.32      | −4.61         | \( y = 0.32x + (-4.61) \) |
| P (Pro)       | 0.02      | −4.26         | \( y = 0.02x + (-4.26) \) |
| S (Ser)       | −0.36     | 1.74          | \( y = -0.36x + (1.74) \) |
| T (Thr)       | 0.05      | −2.51         | \( y = 0.05x + (-2.51) \) |
| W (Trp)       | 0.06      | −3.65         | \( y = 0.06x + (-3.65) \) |
| Y (Tyr)       | 0.23      | −3.11         | \( y = 0.23x + (-3.11) \) |
| V (Val)       | 0.05      | −3.09         | \( y = 0.05x + (-3.09) \) |

Figure 1. Graphical representation of the protein I and protein II by our method.
their Euclidean distance is calculated as follows:

\[ d(V, Y) = \sqrt{\sum_{i=1}^{40} (V_i - Y_i)^2} \]  

The smaller the distance \( d \) is, the more similar 2 protein sequences are.

### Applications and Discussion

#### Similarity analysis of 9 ND5 proteins and 29 spike proteins

To show the effectiveness of the proposed similarity analysis method, we apply it to 9 ND5 protein sequences (provided as Supplementary File 1): human, common chimpanzee, pygmy chimpanzee, gorilla, fin whale, blue whale, rat, mouse, and opossum (their accession number in NCBI [National Center for Biotechnology Information] are AP_000649, NP_008196, NP_008209,NP_008222,NP_006899,NP_007066,AP_004902, NP_904338, and NP_007105, respectively). According to the method given in section “Similarity analysis,” we can obtain the similarity distance matrix of these protein sequences. The corresponding result is shown in Table 5.

On the basis of Table 5, we can find that the distance between fin whale and blue whale is the smallest. This indicates that they have a high degree of similarity. The distance between human, common chimpanzee, pygmy chimpanzee, and gorilla is relatively small, which means that they are similar to each other. Besides, opossum is quite dissimilar to other species because the similarity distances between opossum and other species are large. All these results are consistent with the evolution theory and the recent studies.\(^{14-16}\) That is to say the proposed method can analyze the similarities of proteins effectively.

#### Similarity analysis of 560 gene sequences of influenza A (H1N1) virus

In this section, we give an application for the similarity analysis of HA gene sequences of influenza A (H1N1) from March 1, 2009 to April 30, 2009 (available online at https://www.ncbi.nlm.nih.gov). We obtain a data set that consists of 560 gene sequences with full length (provided as Supplementary File 3). To further demonstrate the validity of our method, we apply the method to this data set.

According to our method, for each virus isolate, we can get a corresponding 20-dimensional vector. Thus, we can obtain a vector set of 560 vectors. By computing the similarity distance between pairs of these vectors, we can obtain a similarity distance matrix. Next, we construct the phylogenetic tree based on our method in Figure 3. To analyze the results better, we mark 2 typical strains: A/California/07/2009 (H1N1) and A/Indiana/08/2009 (H1N1). From Figure 3, it is easy to identify that all virus isolates are mainly classified into 2 groups. This illustrates that there are 2 different kinds of influenza A (H1N1) virus isolates from March 1, 2009 to April 30, 2009.

### Table 5. The similarity matrix for the 9 ND5 protein sequences.

| SPECIES       | HUMAN    | COMMON CHIMPANZEE | PYGMY CHIMPANZEE | GORILLA  | FIN WHALE | BLUE WHALE | RAT       | MOUSE | OPOSSUM |
|---------------|----------|-------------------|------------------|----------|-----------|------------|-----------|-------|---------|
| Human         | 0        | 0.03204           | 0.04244          | 0.04723  | 0.07263   | 0.07744    | 0.18081   | 0.21400| 0.24337 |
| Common chimpanzee | 0       | 0.02842           | 0.04783          | 0.08253  | 0.08917   | 0.17555    | 0.21422   | 0.24196|         |
| Pygmy chimpanzee | 0.05134  | 0.07250           | 0.07938          | 0.16509  | 0.20525   | 0.21064    |           |       |         |
| Gorilla       | 0.06761  | 0.07679           | 0.17006          | 0.20336  | 0.23017   |            |           |       |         |
| Fin whale     | 0.06761  | 0.02457           | 0.16685          | 0.18642  | 0.20773   |            |           |       |         |
| Blue whale    | 0.16515  | 0.16039           | 0.21064          |           |           |            |           |       |         |
| Rat           | 0.07539  | 0.11658           | 0.18642          | 0.20773  |           |            |           |       |         |
| Mouse         | 0.12586  | 0.15286           | 0.20773          | 0.21064  |           |            |           |       |         |
| Opossum       | 0        |                   |                  |           |           |            |           |       |         |

To further demonstrate the effectiveness of our method, we apply it to another data set which is widely used in many works.\(^{10,27}\) This data set consists of 29 spike protein sequences of coronavirus (provided as Supplementary File 2). The basic information of the protein sequences is shown in Table 6. We construct the phylogenetic tree for the 29 spike protein sequences based on our method using UPMGA method in Figure 2. From Figure 2, we can see that all the sequences are mainly classified into 4 groups by our method. This is consistent with the works\(^{10,27}\) and the known biology fact that coronavirus are always classified into 4 groups: the group I (contains PEDV, TGEV), the group II (contains BCoV, MHV, RtCoV), the group III (contains IBV, TCov), and the SARS-CoVs (severe acute respiratory syndrome coronavirus).
This result is consistent with the works by Qi et al.14,28 Furthermore, the result is also consistent with the biology fact that a new influenza virus, A/California/07/2009 (H1N1)–like virus, appeared and showed a strong ability to infect human beings in April 2009.23 The branch length in Figure 3 is the similarity distance between 2 virus isolates.

ClustalW is one of the most widely used multiple sequence alignment method for nucleic acid and protein sequence in molecular biology. We construct the phylogenetic tree of the 560 gene sequences using ClustalW method29 under MEGA6.0 software for comparison. From the phylogenetic tree in Figure 4, we can see that all virus isolates are also classified into 2 groups. In the figure, we also mark 2 typical strains: A/California/07/2009 (H1N1) and A/Indiana/08/2009 (H1N1). Observing Figures 3 and 4, one can easily find out that the results by our method are

### Table 6. The information of 29 spike protein sequences.

| NUMBER | ABBREVIATION | ACCESS NUMBER |
|--------|--------------|---------------|
| 1      | TGEVG        | CAB91145      |
| 2      | TGEV         | NP 058424     |
| 3      | PEDVC        | AAK38656      |
| 4      | PEDV         | NP 598310     |
| 5      | HCoVOC43     | NP 937950     |
| 6      | BCoVE        | AAK83356      |
| 7      | BCoVL        | AAL57308      |
| 8      | BCoVM        | AAA66399      |
| 9      | BCoVQ        | AAL40400      |
| 10     | MHVA         | AAB86819      |
| 11     | MHVJHM       | YP 209233     |
| 12     | MHVP         | AAF69334      |
| 13     | MHVM         | AAF69344      |
| 14     | IBVBJ        | AAP92675      |
| 15     | IBVC         | AAS00080      |
| 16     | IBV          | NP 040831     |
| 17     | GD03T0013    | AAS10463      |
| 18     | PC4127       | AAU73318      |
| 19     | PC4137       | AAV9720       |
| 20     | PC4205       | AAU73319      |
| 21     | civet007     | AAU04646      |
| 22     | civet010     | AAU04649      |
| 23     | A022         | AAV91631      |
| 24     | GD01         | AAP51227      |
| 25     | GZ02         | AAS00003      |
| 26     | BJ01         | AAP30030      |
| 27     | FRA          | AAP50485      |
| 28     | TOR2         | AAP4037       |
| 29     | TaiwanTC1    | AAQ01597      |
consistent with those by ClustalW method. Furthermore, it takes about 126 minutes to obtain the multiple sequence alignment result on our Intel Core i5-3230M CPU @ 2.60 GHz 2.60GHz Windows PC with 4GB RAM. However, the computation time of our method is 105.769 seconds by a Python program. It can indicate that our method is a computational efficiency method when dealing with sequences with different lengths.

Conclusions
In this article, a new 3D graphical representation of protein sequences is introduced based on 10 physicochemical properties and BLOSUM62 matrix. On the basis of the graphical representation curve, we extract a specific vector and use the vector to calculate the similarity distance between 2 protein sequences. To prove the effectiveness of our method, we apply our method to 3 real data sets. The results show the validity of our method in phylogenetic analysis compared with related works and evolution facts.

Acknowledgements
This paper has not been submitted elsewhere for consideration of publication.

Author Contributions
Z-HQ conceived and designed the work that led to the submission. K-CL contributed significantly to analysis and manuscript preparation. J-LM and Y-HY helped to perform the analysis with constructive discussions. All the authors reviewed and approved the final manuscript.

REFERENCES
1. Randić M, Plavšić D. Novel spectral representation of RNA secondary structure without loss of information. Chem Phys Lett. 2009;476:277-280.
2. Randić M, Nović M, Plavšić D. Milestones in graphical bioinformatics. Int J Quant Chem. 2013;113:2413–2446.
3. Bielińska-Waż D. Four-component spectral representation of DNA sequences. J Math Chem. 2010;47:41–51.
4. Bielińska-Waż D. Graphical and numerical representations of DNA sequences: statistical aspects of similarity. J Math Chem. 2011;49:2345–2407.
5. Bielińska-Waż D, Clark T, Waż P, Nowak W, Nandy A. 2D-dynamic representation of DNA sequences. Chem Phys Lett. 2007;442:140–144.
6. Bielińska-Waż D, Waż P. Spectral-dynamic representation of DNA sequences. J Biomed Inform. 2017;72:1–7.
7. Waż P, Bielińska-Waż D. 3D-dynamic representation of DNA sequences. J Mol Model. 2014;20:2141
8. Cao Z, Liao B, Li R. A group of 3D graphical representation of DNA sequences based on dual nucleotides. Int J Quantum Chem. 2008;108:1485–1490.
9. Randić M. WITHDRAWN: 2-D graphical representation of proteins based on physico-chemical properties of amino acids. Chem Phys Lett. 2007;444:176–180.
10. Wang L, Peng H, Zheng J. ADLID: a novel graphical representation of protein sequences and its application. Comput Math Method Med. 2014;2014:959753.
11. Yu C, Cheng SY, He RL, Yau SST. Protein map: an alignment-free sequence comparison method based on various properties of amino acids. Gene. 2011;486:110–118.
12. El-Lakkani A, Mahran H. An efficient numerical method for protein sequences similarity analysis based on a new two-dimensional graphical representation. SAR QSAR Environ Res. 2015;26:125–137.
13. Randić M, Mehulić K, Vukičević D, Pisanski T, Vikić-Topić D, Plavšić D. Graphical representation of proteins as four-color maps and their numerical characterization. J Mol Model. 2009;27:637–641.
14. Qi ZH, Jin MZ, Li SL, Feng J. A protein mapping method based on physicochemical properties and dimension reduction. Comput Biol Med. 2015;57:1–7.
15. He PA, Zhang YP, Yao YH, Tan YF, Nan XY. The graphical representation of protein sequences and its applications. J Comput Chem. 2010;31:2136–2142.
16. Hu H. F-Curve, a graphical representation of protein sequences for similarity analysis based on physicochemical properties of amino acids. MATCH Commun Math Comput Chem. 2015;73:749–764.
17. Yao YH, Dai Q., Li C, He PA, Nan XY, Zhang YZ. Analysis of similarity/dissimilarity of protein sequences. Proteins. 2008;73:864–871.
18. Alff-Steinberger C. The genetic code and error transmission. PLoS Natl Acad Sci. 1969;64:538–591.
19. Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974;185:862–864.
20. Sneath PHA. Relations between chemical structure and biological activity. J Theor Biol. 1966;12:157–195.
21. Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC. On the fundamental nature and evolution of the genetic code. Cold Spring Harb Symp. 1966;31:723–736.
22. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. P Natl Acad Sci. 1992;89:10915–10919.
23. McLaughlin L. Automated programming: the next wave of developer power tools. IEEE Software. 2006;23:91–93.
24. Amorim RCD, Hennig C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. Inform Sci. 2015;324:126–145.
25. Randić M, Butina D, Zupan J. Novel 2-D graphical representation of protein sequences. Chem Phys Lett. 2006;419:528–532.
26. Fang L. Structure, function, and evolution of coronavirus spike proteins. Annu Rev Virol. 2011;3:195–1964.
27. Qi ZH, Feng J, Liu CC. Evolution trends of the 2009 pandemic influenza A (H1N1) viruses in different continents from March 2009 to April 2012. Biologia. 2014;69:407–418.
28. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30:2725–2729.