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

Xianyou Zhu

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

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

Email: 364715358@qq.com

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

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

Introduction

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

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

Graphical Representation of Protein Sequences

Proteins are composed of 20 different amino acids and these amino acids have many different physicochemical and biological properties such as the molecular weight (\(mW\)), isoelectric point (\(pI\)), the \(pK_a\) value for terminal amino acid groups COOH (\(pK_1\)), the \(pK_a\) value for terminal amino acid groups \(NH_3^+\) (\(pK_2\)), van der waals radius (Vdwa), kDHydrophobicity (kh) (Kyte and Doolittle, 1982), wHydrophobicity (wh) (Wimley and White, 1996), hHydrophobicity (hh) (Hessa et al., 2005), the occurrence in human properties (%) (Oihp (%)), Abundance (Abu), ATP cost in synthesis under aerobic condition (Csa) and ATP cost in synthesis under anaerobic condition (Csan) etc. The names and symbols of the 20 amino acids and the value of their 12 major properties are illustrated in the following Table 1.
Let \( \{F_1, F_2, \ldots, F_{12}\} \) represent these 12 different properties of amino acids illustrated in above Table 1 and \( \Omega = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\} \) be the set of 20 kinds of amino acids, \( \forall r \in \Omega \), let \( X_r, X_r^2, \ldots, X_r^{12} \) be the values of 12 features of \( r \),

\[
Y_r = \frac{\sum_{i \in \{1,2,\ldots,12\}} X^i_r}{20},
\]

then we can standardize the values of 12 features of \( r \) according to the following Formula 1:

\[
Y'_r = Y_r - \bar{Y}
\]

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

Let \( \Psi = p_1 p_2 \ldots p_n \) \((p_i \in \Omega, \forall i \in \{1,2,\ldots,N\}) \) represent a protein sequence with \( n \) amino acids and for any given letter \( u \in \Omega \), suppose that \( u \) represents \( k \) times in the protein sequence \( \Psi \) totally and the location of \( u \) at the \( j^{th} \) time in \( \Psi \) is \( u_j \), then we call the vector \( \langle u_{i1}, u_{i2}, \ldots, u_{in} \rangle \) as the “Feature Appearance Model” of \( u \) in \( \Psi \).

Based on the concept of “Feature Appearance Model” proposed above, then for each protein sequence \( \Psi \), we can obtain its graphical representation according to the above steps:

**Step1:** According to the concept of Feature Appearance Model, obtain 20 different Feature Appearance Models of amino acids in the protein sequence \( \Psi \).

**Step2:** For \( j = 1 \) to \( 12 \), select the \( j^{th} \) feature from these 12 features of amino acids \( \{F_1, F_2, \ldots, F_{12}\} \), \( \forall p_i \in \Psi \), let the standardized values of the \( j^{th} \) feature of \( p_{i1}, p_{i2} \) and \( p_{i1}, p_{i2} \) be \( Y'_{i1}, Y'_{i1} \) and \( Y'_{i2}, Y'_{i2} \) respectively and the Feature Appearance Model of \( p_i \) be \( \Delta_i = \langle u_{i1}, u_{i2}, \ldots, u_{i12} \rangle \), then we can obtain \( k \) different coordinates of \( p_i \) such as \( (x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_k, y'_k) \) according to Formula 2 and 3:

\[
x'_i = u'_i
\]

\[
y'_i = Y'_i + (Y'_i - Y'_{i1}) \left( \frac{Y_{i1} - Y_{i2} + 1}{Y_{i1} - Y'_{i1} + 1} \right)
\]

where, \( i \in \{1,2,\ldots,k\} \) and for \( p_j \in \Psi \), since there isn’t \( p_0 \) in \( \Psi \), then we define \( Y'_0 = 0 \), for \( j \in \{1,2,\ldots,12\} \).

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

**Similarities/Dissimilarities analysis Model of Protein Sequences**

Let \( G_\Psi \) represent a graph of \( \Psi \) obtained by the method given above in section 2 and \( \forall p_i \in \Psi \), \( i \in \{1,20\} \), let the Feature Appearance Model of \( p_i \) be \( \Delta_i = \langle u_{i1}, u_{i2}, \ldots, u_{i12} \rangle \) and \( E_i \in G_{\Psi} \) represent the curve of \( p_i \) in \( G_{\Psi} \), then we can obtain the \( ED, PD, D/D, L/L \) matrixes of \( E_i \) according to the following Formula 5-8 respectively (Randic et al., 2000; 2003a; 2003b; 2003c; Randic, 2003; Randic et al., 2004; Randic and Vracko, 2003; Bajzer et al., 2003). Thereafter, let \( M'_t, M'_e, M'_r, M'_s \) represent the \( ED, PD, D/D, L/L \) matrixes of \( E_i \) respectively, we will finally obtain 80 different matrixes for \( \Psi \), since there are 20 curves in \( G_{\Psi} \):

\[
[ED]_{i,j} = \sqrt{(u_{i1} - u_{j1})^2 + (u_{i2} - u_{j2})^2 + \ldots + (u_{i12} - u_{j12})^2}
\]

\[
[PD]_{i,j} = \left[ [ED]_{i+1,j+1} + [ED]_{i+1,j} + \ldots + [ED]_{i,j+1} \right]_{i < j}
\]
Based on the matrixes $M'_{i}, M'_{i}', M_{i}, M_{i}'$ obtained above, then we can compute out the maximum eigenvalues \( \{\lambda_{i}\} \), the minimum eigenvalues \( \{\lambda_{i}\} \), the average eigenvalues \( \{\bar{\lambda}_{i}\} \), the sum of the maximum and minimum eigenvalues \( \{\lambda_{i} + \lambda_{i}\} \), the index values of ALE \( \{\pi_{i}\} \) of these matrixes respectively (Li and Wang, 1966; Shrock and Tsai, 1997; Biggs, 1974). Hence, for \( E_{t} \in G_{\Psi} \), we can describe it with a 20 dimensional vector \( V_{t} \) as the following Formula 9:

\[
V_{t} = x_{i}'x_{j}', \ldots, x_{i}'d_{i}, x_{j}'d_{j}, \ldots, x_{i}'d_{i}, x_{j}'d_{j}
\]

Thereafter, the graph of \( \Psi \) can be represented as a 20\times20 matrix \( M_{\Psi} = [V_{1}, V_{2}, \ldots, V_{20}]^{T} \), called the Descriptor Matrix.

Based on the Descriptor Matrix obtained above, we randomly select \( k \) \((k \in [1,20])\) columns from \( M_{\Psi} \) each time, then we will obtain a new 20\times\( k \) matrix \( M_{\Psi}^{k} = [V_{1}^{k}, V_{2}^{k}, \ldots, V_{20}^{k}]^{T} \), where \( V_{j}^{k} \) is a \( k \) dimensional vector for any \( j \in [1,20] \). Therefore, for any two protein sequences \( \Psi_{1} \) and \( \Psi_{2} \), supposing that we have obtained two 20\times\( k \) matrix \( M_{\Psi_{1}}^{k} \) and \( M_{\Psi_{2}}^{k} \), then we can obtain the distance \( d(\Psi_{1}, \Psi_{2}) \) between \( \Psi_{1} \) and \( \Psi_{2} \) as follows:

\[
d(\Psi_{1}, \Psi_{2}) = \sum_{i=1}^{20} \sum_{j=1}^{k} (V_{ij}^{1} - V_{ij}^{2})^{2}
\]

where:

\[
\|V_{i}^{k} - V_{i}^{k}\| = \sqrt{\sum_{j=1}^{k} (d_{ij}^{1} - d_{ij}^{2})^{2}}
\]

Table 2. The basic information of 16 ND5 protein sequences

| No. | Name                | abbreviation | Access No   | Length |
|-----|---------------------|--------------|-------------|--------|
| 1   | Human               | human        | ADT80430.1  | 603    |
| 2   | Gorilla             | gorilla      | NP 008222   | 603    |
| 3   | Pigmy Chimpanzee    | pi-chim      | NP 008209   | 603    |
| 4   | Common Chimpanzee   | c-chim       | NP 008196   | 603    |
| 5   | Fin Whale           | fin-whale    | NP 006899   | 606    |
| 6   | Blue Whale          | blue-whale   | NP 007066   | 606    |
| 7   | Rat                 | rat          | AP 006899   | 610    |
| 8   | Mouse               | mouse        | NP 904338   | 607    |
| 9   | Opossum             | opossum      | NP 007105   | 602    |
| 10  | Sheep               | sheep        | ABW22903.1  | 606    |
| 11  | Goat                | goat         | BAN59258.1  | 606    |
| 12  | Lemur               | lemur        | CAD13431.1  | 603    |
| 13  | Cattle              | cattle       | ADN11902.1  | 606    |
| 14  | Hare                | hare         | CAD13291.1  | 603    |
| 15  | Gallus              | Gallus       | BAE16036.1  | 605    |
| 16  | Rabbit              | rabbit       | NP 007559.1 | 603    |

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

| No. | Access No. | Abbreviation | Length |
|-----|------------|--------------|--------|
| 1   | CAA25111   | Bovine       | 145    |
| 2   | CAA26204   | Chimpanzee   | 125    |
| 3   | CAA68429   | Hare         | 147    |
| 4   | CAA23700   | Gallus       | 147    |
| 5   | AAA30913   | Goat         | 145    |
| 6   | CAA43421   | Gorilla      | 121    |
| 7   | AAA16334   | Human        | 147    |
| 8   | AAA36822   | Lemur        | 147    |
| 9   | CAA24101   | Mouse        | 147    |
| 10  | AAA30976   | Opossum      | 147    |
| 11  | CAA24251   | Rabbit       | 147    |
| 12  | CAA29887   | Rat          | 147    |
| 13  | NP_001091117 | Sheep    | 145    |
Based on the Formula 10, we can obtain three other distance matrices \( M_{sN}, M_{sG} \) and \( M_{sS} \) according to three groups of protein sequences such as the 16 ND5 protein sequences, 13 globin protein sequences and 29 sequences of spike protein respectively. The basic information of these three groups of protein sequences are illustrated in the following Table 2 to 4.

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

Considering the above two groups of distance matrices \( \{M_{sN}, M_{sG}, M_{sS}\} \) and \( \{M_{oN}, M_{oG}, M_{oS}\}, \forall \Pi \in \{N, G, S\} \), supposing that the number of columns in matrix \( M_{df} \) is \( N_{df} \), \( M_{df} = [V_{d1}^f, V_{d2}^f, \ldots, V_{dN_{df}}^f] \) and \( M_{df} = [V_{d1}^o, V_{d2}^o, \ldots, V_{dN_{df}}^o] \), then we define the Average Correlation Coefficient \( \mbox{Acorr} \) between \( M_{df} \) and \( M_{df} \) as follows:

\[
\mbox{Acorr} (M_{df}, M_{df}) = \frac{\sum_{i=1}^{N_{df}} \text{corrcoef}(V_{d1}^f, V_{d1}^o)}{N_{df}} \tag{12}
\]

where, \( \text{corrcoef}(V_{d1}^f, V_{d1}^o) \) is the correlation coefficient between the \( f \)th column vectors \( V_{d1}^f \) and \( V_{d1}^o \) in the matrices \( M_{df} \) and \( M_{df} \) respectively. Thus, we will obtain three different Average Correlation Coefficients such as \( \mbox{Acorr}(M_{oN}, M_{sN}), \mbox{Acorr}(M_{oG}, M_{sG}), \mbox{Acorr}(M_{oS}, M_{sS}) \).

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

For the protein sequence \( \Psi \), supposing that we finally obtain \( \Gamma \) different values of \( \mbox{AvgCC} \) such as \( \{\mbox{AvgCC}_1, \mbox{AvgCC}_2, \ldots, \mbox{AvgCC}_\Gamma\} \) and there is \( \mbox{AvgCC}_1 \geq \mbox{AvgCC}_2 \geq \ldots \geq \mbox{AvgCC}_\Gamma \), and in addition, supposing that to obtain the value of \( \mbox{AvgCC}_1 \), we shall select the \( J \)th \((J \in [1, 2])\) feature from these 12 features of amino acids and \( K \) \((K \in [1, 20])\) columns \( \{d_{kj}, d_{k2j}, \ldots, d_{k20j}\} \) from \( M_f \), then, we will call the \( J \)th feature of amino acids as the Optimal Feature for the protein sequence \( \Psi \) and \( \{d_{kj}, d_{k2j}, \ldots, d_{k20j}\} \) as the Optimal Descriptors of the protein sequence \( \Psi \).
Obviously, the Optimal Feature obtained above can be utilized for graphical representation of protein sequences and the Optimal Descriptors obtained above can be utilized for analyzing the similarities/dissimilarities of protein sequences.

Based on three groups of protein sequences such as the 16 ND5 protein sequences, 13 globin protein sequences and 29 sequences of spike protein, through experiments, it is easy to prove that the Optimal Feature will be \( pK2 \) and the Optimal Descriptors will be \( \left\{ v^p, v^d, s^p, s^d, a^i \right\} \) for \( t \in [1, 20] \) and for convenience, we rewrite the 20 different Optimal Descriptors as a matrix

\[
\begin{bmatrix}
v^p_1 & v^d_1 & s^p_1 & s^d_1 & a^1_i \\
v^p_2 & v^d_2 & s^p_2 & s^d_2 & a^2_i \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
v^p_{20} & v^d_{20} & s^p_{20} & s^d_{20} & a^{20}_i
\end{bmatrix}
\]

Hence, we can adopt \( pK2 \) and the parameters to construct a new Similarities/Dissimilarities Analysis Model according to the following steps:

**Step1:** For each protein sequence \( \Psi \) in these two groups of protein sequences such as 16 ND6 proteins and 15 myoglobin proteins, obtain its graphical representation \( G_\Psi \) based on the feature \( pK2 \).

**Step2:** According to the Formula 10 and the matrix of Optimal Descriptors \[\begin{bmatrix} v^p_1 & v^d_1 & s^p_1 & s^d_1 & a^1_i \\
v^p_2 & v^d_2 & s^p_2 & s^d_2 & a^2_i \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
v^p_{20} & v^d_{20} & s^p_{20} & s^d_{20} & a^{20}_i \end{bmatrix}, \]

obtain two distance matrixes \( M_{oND6} \) and \( M_{oGlobin} \) for these two groups of protein sequences such as 16 ND6 proteins and 15 myoglobin proteins respectively.

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

### Results and Analysis

#### Graphical Representation of Protein Sequences

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

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

**Similarity/Dissimilarity Analysis of Protein Sequences**

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

| Homo_sapiens | Gorilla_gorilla | Pan_panicus | Pan_troglodytes | Balaenoptera_phy | Balaenoptera_musculus | Rattus_norvegicus |
|--------------|----------------|-------------|-----------------|------------------|----------------------|------------------|
| Homo_sapiens | 0              | 0           | 0               | 0                | 0                    | 0                |
| Gorilla_gorilla | 428.2748757   | 0           | 0               | 0                | 0                    | 0                |
| Pan_panicus | 245.5669038   | 0           | 0               | 0                | 0                    | 0                |
| Pan_troglodytes | 344.1705447   | 0           | 0               | 0                | 0                    | 0                |
| Balaenoptera_phy | 341.0092666   | 0           | 0               | 0                | 0                    | 0                |
| Balaenoptera_musculus | 343.6789815   | 0           | 0               | 0                | 0                    | 0                |
| Rattus_norvegicus | 338.235028    | 0           | 0               | 0                | 0                    | 0                |

Table 5 the distance matrix of 16 ND5 protein sequences

Xianyou Zhu / American Journal of Biochemistry and Biotechnology 2016, 12 (1): 47-55
DOI: 10.3844/ajbbsp.2016.47.55
### Table 6: The distance matrix of 15 myoglobin protein sequences

| Protein 1          | Notothenia_coriiceps | Bos_taurus | Gallus_gallus | Iguana_iguana | Equus_caballus | Mus_musculus | Homo_sapiens | Neopagetopsis_ionah | Rattus_norvegicus | Chionodraco_rastrospinosus | Sus_scrofa | Ursus_maritimus | Ovis_aries | Physeter_catodon | Leptonychotes_weddellii | Neopagetopsis_ionah | Rattus_norvegicus | Chionodraco_rastrospinosus | Sus_scrofa | Ursus_maritimus | Ovis_aries | Physeter_catodon | Leptonychotes_weddellii |
|--------------------|----------------------|------------|---------------|---------------|----------------|---------------|---------------|----------------------|---------------------|-----------------------------|-------------|----------------|-----------|----------------|--------------------------|----------------------|---------------------------|-----------------------------|-------------|----------------|-----------|----------------|--------------------------|
| Notothenia_coriiceps | 0                    |            |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Bos_taurus         | 4097.44              | 0          |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Gallus_gallus      | 4792.56              | 3560.79    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Iguana_iguana      | 4546.51              | 2783.15    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Equus_caballus     | 4818.31              | 1981.66    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Mus_musculus       | 4281.93              | 1961.60    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Homo_sapiens       | 4172.33              | 1770.81    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Neopagetopsis_ionah| 388.75               | 4103.15    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Rattus_norvegicus  | 4547.07              | 2138.12    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Chionodraco_rastrospinosus | 788.15 | 3832.18 | 5079.82 | 4723.46 | 4499.04 | 4078.14 | 3875.70 | 3890.45 | 4673.29 | 3851.76 | 3889.83 |
| Sus_scrofa         | 4078.14              | 1415.23    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Ursus_maritimus    | 4108.56              | 1623.44    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Ovis_aries         | 3911.86              | 712.94     |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Physeter_catodon   | 4033.47              | 2166.93    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| Leptonychotes_weddellii | 6401.13 | 3911.44 | 3275.70 | 3890.45 | 4673.29 | 3851.76 | 3889.83 | 3875.70 | 4673.29 | 3851.76 | 3889.83 |
| Neopagetopsis_ionah| 4470.33              | 1172.15    |               |               |                |               |               |                      |                     |                             |             |               |           |               |                          |                      |
| (a) (b)            (c) (d) |

**Fig. 1.** Some graphs of 16 ND6 protein sequences based our method (a) Homo sapiens (Human) (b) Gorilla gorilla (Gorilla) (c) Pan paniscus (P-Chim) (d) Pan troglodytes (C-Chim)
From Table 5, it is easy to find that there are some similar pairs such as (Human, P-Chim) with the distance 245.57, (Human, C-Chim) with the distance 340.17, (Human, Gorilla) with the distance 428.27, (Gorilla, P-Chim) with the distance 336.88, (Gorilla, C-Chim) with the distance 447.87, (P-Chim, C-Chim) with the distance 268.25, (Fin-Wha, Blu-Wha) with the distance 1254.55 and (Sheep, Goat) with the distance 906.41, etc., and among them, the Opossum and Gallus seems to be two peculiar mammals, since the shortest distance between Opossum and the remaining mammals is more than 3371.80 and the shortest distance between Gallus and the remaining mammals is more than 3071.20. Obviously, the result is consistent with the fact that Opossum is the most remote specie from the remaining mammals and Gallus is not a kind of mammal.

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

The Phylogenetic Tree of the Protein Sequences

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

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

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

Conclusion

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

Acknowledgement

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

Ethics

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

References

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

Abo-Elkhier, M.M., 2012. Similarity/dissimilarity analysis of protein sequences using the spatial median as a descriptor. J. Biophys. Chem., 3: 142-148. DOI: 10.4236/jbpc.2012.3.32016

Bajzer, Z., M. Randic, D. Plasic and S.C. Basak, 2003. Novel map descriptors for characterization of toxic effects in proteomics maps. J. Mol. Graph. Model., 22: 1-9. DOI: 10.1016/S1093-2636(02)00186-9

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

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

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

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

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

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

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

Qi, Z.H., J. Feng, X.Q. Qi and L. Li, 2012. Application of 2D graphic representation of protein sequence based on Huffman tree method. Comput. Biol. Med., 42: 556-563. DOI: 10.1016/j.compbiomed.2012.01.011

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

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

Randic, M., M. Vracko, A. Nandy and S.C. Basak, 2000. On 3-D graphical representation of DNA primary sequences and their numerical characterization. J. Chem. Inf. Comput. Sci., 40: 1235-1244. DOI: 10.1021/ci000334q

Randic, M., M. Vracko, J. Zupan and A.T. Balaban, 2004. Unique graphical representation of protein sequences based on nucleotide triplet codons. Chem. Phys. Lett., 397: 247-252. DOI: 10.1016/j.cplett.2004.08.118
Randic, M., M. Vracko, N. Lers and D. Plavšić, 2003a. Novel 2-D graphical representation of DNA sequences and their numerical characterization. J. Chem. Phys. Lett., 368: 1-6. DOI: 10.1016/S0009-2614(02)01784-0

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

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

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

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar, 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Molecular Biol. Evolut., 30: 2725-2429. DOI: 10.1093/molbev/msr197

Thompson, J.D., D.G. Higgins and T.J. Gibson, 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res., 22: 4673-4680. PMID: 7984417

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

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

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

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

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