Physica A 391 (2012) 93–99

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A novel graphical representation of proteins and its application

Ping-an He [a,*], Jinzhou Wei [a], Yuhua Yao [b], Zhixin Tie [c]

[a] Department of Mathematics, College of Science, Zhejiang Sci-Tech University, Hangzhou 310018, PR China

[b] College of Life, Zhejiang Sci-Tech University, Hangzhou 310018, PR China

[c] Instructional Division of Computer Technology, Zhejiang Sci-Tech University, Hangzhou 310018, PR China

ARTICLE INFO

Article history:
Received 31 March 2011
Received in revised form 18 July 2011
Available online 25 August 2011

Keywords:
Protein
Graphical representation
Similarity
Numerical characterization
Correlation and significance analysis

ABSTRACT

On the basis of three kinds of indices of physicochemical properties of twenty amino acids, a novel 3D graphical representation of protein sequences is presented. Then, the improved cumulative distance of the 3D graphical representations is defined in order to compare the proteins for similarity. Furthermore, the efficiency of our approach is illustrated by performing a comparison of similarities/dissimilarities among sequences of the ND5 proteins of nine different species. A correlation and significance analysis is provided to compare our results on similarities/dissimilarities and some other graphical representation results with the ClustalW results on similarities/dissimilarities. The comparison results show that our approach has better correlations with ClustalW for all nine species than other approaches.

1. Introduction

The number of biological sequences is rapidly rising in various biological databases with sequencing techniques advancing. The comparisons of DNA, RNA and protein sequences are among the most common and important tasks in molecular biology and bioinformatics. One of the most well known algorithms for performing the string-matching operation presented in this task is the Smith–Waterman algorithm, in which a distance function or a score function is used to represent insertion, deletion, and substitution in the compared sequences. Due to it being computationally intensive, however, many researchers have developed alignment-free methods to avoid using it [1]; the graphical representation of biological sequences, for example, is a very useful tool for visual comparison of DNA, RNA and protein sequences [2–13]. Graphical representations of biological sequences have been used in comparisons of biological sequences not only to provide a simple way to view complex relationships but also to numerically characterize the similarities/dissimilarities between biological sequences [2].

The methods of graphical representation of proteins, which are generalized from the graphical representation of DNA, are computationally complicated because of the substitution from four bases to twenty amino acids. Recently, various graphical representations of proteins have been introduced by many researchers [14–38]. For example, some graphical representations of protein sequences based on the genetic code were suggested by Randic [15–17] and Liao et al. [18]. Using the indices of some physicochemical properties of the twenty amino acids, many graphical representations of protein sequences have been proposed by Randic [19], Yao [20], Feng [21], Yau [22], Wen [23], el Maaty [24] and He [25]. Several graphical representations of proteins based on the modification of existing graphical representations of DNA have been constructed for the comparison of proteins [26–33]. In particular, the graphical approach of Jeffrey [13] for DNA representation was generalized to obtain a graphical representation of proteins by Randic and He et al., in which the twenty amino acids were placed on the periphery of the unit circle, and thus a square was replaced by a 20-side polygon [34–38].

* Corresponding author.
E-mail address: pinganhe@zstu.edu.cn (P.-a. He).

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.08.015

Table 1
Three parameters of the 20 amino acids and their coordinates in the new Cartesian system.

Amino acid	$pK1(a\text{-COOH})$	$pK2(NH_3)$	pI (at 25 °C)	x coordinate	y coordinate	z coordinate
A	2.33	9.71	6.00	0.18	0.334	−0.0245
C	1.91	10.28	5.07	−0.24	0.904	−0.9545
D	1.95	9.66	2.77	−0.2	0.284	−3.2545
E	2.16	9.58	3.22	0.01	0.204	−2.8045
F	2.18	9.09	5.48	0.03	−0.286	−0.5445
G	2.34	9.58	5.97	0.19	0.204	−0.0545
H	1.70	9.09	7.59	−0.45	−0.286	1.5655
I	2.26	9.60	6.02	0.11	0.224	−0.0045
K	2.15	9.16	9.74	0	−0.216	3.7155
L	2.32	9.58	5.98	0.17	0.204	−0.0445
M	2.16	9.08	5.74	0.01	−0.296	−0.2845
N	2.16	8.73	5.41	0.01	−0.646	−0.6145
P	1.95	10.47	6.30	−0.2	1.094	0.2755
Q	2.18	9.00	5.65	0.03	−0.376	−0.3745
R	2.03	9.00	10.76	−0.12	−0.376	4.7355
S	2.13	9.05	5.68	−0.02	−0.326	−0.3445
T	2.20	8.96	5.60	0.05	−0.416	−0.4245
V	2.27	9.52	5.96	0.12	0.144	−0.0645
W	2.38	9.34	5.89	0.23	−0.036	−0.1345
Y	2.24	9.04	5.66	0.09	−0.336	−0.3645

Amino acids are the basic building blocks of protein molecules. Various physicochemical AA indexes, such as the hydrophobicity, hydrophilicity, and side chain mass, are often used to enhance the prediction quality of protein attributes [34]. Among these, three parameters, $pK1(\text{COOH})$, $pK2(NH_3^+)$ and pI at 25 °C, were adopted to construct the 3D Cartesian coordinates of amino acids, whose values are listed in Table 1.

It is worthy of note that the interaction of neighbor amino acids plays a major role in molding protein structure. In this paper, considering the interaction between neighbor amino acids, the 3D graphical representation for a protein based on Jeffrey's method [13] is proposed. Then, a novel descriptor is suggested to characterize the 3D graphical representation of a protein, and a distance between two 3D graphical representations is introduced to compare the similarities of two corresponding proteins. The similarities/dissimilarities of the ND5 protein sequences of nine different species are employed to illustrate the utility of our method. Furthermore, a correlation and significance analysis is provided in order to compare our results on similarities/dissimilarities and some other graphical representation results with the ClustalW similarity/dissimilarity results. The comparison results show that our approach has better correlations with ClustalW for all nine species than other approaches.

2. 3D graphical representation of protein sequences

Proteins are linear polymers composed of twenty different amino acids, linked by covalent bonds. Physicochemical properties of amino acids, such as the relative molecular mass, solubility limit, specific rotation, isoelectric point, hydropathy index, melting point, and pK a values for terminal amino acid groups of COOH and NH_3, can be used to study protein sequence profiles, folding and functions.

As is well known, the proton-donating ability of $pK1(\text{COOH})$ is essential for the chemical properties of proteins, and the ionization constant $pK1(\text{COOH})$ determines the catalytic activities of enzymes. The proton-accepting ability of $pK2(NH_3^+)$ is also important in biochemistry to determine the activity of enzymes. The pI is the pH of an aqueous solution of an amino acid (or peptide) at which the molecules on average have no net charge. It could reflect the innate structure of the protein sequence, rather than the apparent legitimate structure.

As shown in Table 1, the values of the three parameters are all positive. If the values are considered as coordinates of the points representing the twenty amino acids in a Cartesian (x, y, z) coordinate system, then those points will all lie in the first quadrant. The averages values of $pK1(\text{COOH})$, $pK2(NH_3^+)$ and pI are 2.15, 9.376 and 6.0245, respectively. We take the point (2.15, 9.376, 6.0245) as the origin, and then form a new Cartesian system. Thus, the 20 points are distributed in six quadrants in the new system. The new components of these points are also listed in Table 1.

Given a protein sequence $S = s_1 s_2, \ldots, s_n$, we inspect it by stepping one amino acid at a time. For step i ($i = 1, 2, \ldots, n-1$), a 3D space point P_i (x_i, y_i, z_i) can be constructed as follows:

$$\begin{cases} x_i = \dfrac{1}{3}(x_{i-1} + s_i^1 + s_{i+1}^1) \\ y_i = \dfrac{1}{3}(y_{i-1} + s_i^2 + s_{i+1}^2) \\ z_i = \dfrac{1}{3}(z_{i-1} + s_i^3 + s_{i+1}^3) \end{cases}$$