Physica A 389 (2010) 4668–4676

Contents lists available at ScienceDirect

Physica A

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

3D graphical representation of protein sequences and their statistical characterization

Moheb I. Abo el Maaty [*], Mervat M. Abo-Elkhier, Marwa A. Abd Elwahaab

Department of Engineering Mathematics and Physics, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt

ARTICLE INFO

Article history:
Received 12 January 2010
Received in revised form 24 May 2010
Available online 25 June 2010

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

ABSTRACT

Based on three physicochemical properties of amino acid side chains, we proposed a novel unique 3D graphical representation of protein sequences. Then, we constructed two vectors of three components as mathematical objects to characterize protein sequences numerically. The similarity/dissimilarity analysis among nine ND5 protein sequences proved the utility of our approach. A correlation and significance analysis have been provided to compare our results and the sequence homology.

1. Introduction

Similarity/dissimilarity analysis of DNA and protein sequences is very important. Its importance arises from the fact that proteins with similar sequences frequently share similar structures [1]. Letter sequence representation (LSR) of DNA sequences represents each base by a letter of four different letters such as A, T, G and C. Similarly for protein sequences, it represents each amino acid by a letter of twenty different letters such as A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y and V. The LSR is necessary for information storage but is difficult to recognize and compare different sequences.

Many mathematical approaches were proposed to translate DNA or protein sequences from letters to 2D or 3D graphical representations accompanied by mathematical objects such as vectors or matrices to use them as sequence descriptors and compare these mathematical objects. For example, if the mathematical objects are vectors, we can calculate the euclidean distance or correlation angle between each two vectors. Based on their values, the similarity/dissimilarity matrix can be obtained. The first 2D graphical representation of DNA was proposed in 1985 [2]. A 2D graphical representation was proposed by assigning each base of the DNA sequence to the four directions $+x$, $-x$, $+y$ and $-y$ respectively [3]. Many modifications were proposed in DNA 2D graphical representations [4–9]. One of these assigned four horizontal lines separated by unit distances instead of assigning vectors to four DNA bases [7,8]. Another representation was made by assigning the four corners of a square to four DNA bases [9]. Many 3D graphical representations of DNA primary sequences were proposed [10–12].

Graphical representation of proteins emerged only recently [13–21]. A protein sequence consists of twenty different amino acids while a DNA sequence consists of only four bases. As a result of this, there is a delay in protein 2D and 3D graphical representations. A uniform distribution of twenty amino acids on the circumference of a magic circle of unit radius was the modification of the idea of assigning the coordinates of the four corners of a square to DNA bases [14]. Another 2D graphical representation resulted from assigning a vector of two components to each amino acid [19]. These two components were pKa of $NH3^+$ and COOH as x- and y-coordinates respectively. A modification of this work was done by representing the x-coordinate as pKa ($NH3^+$) and the y-coordinate as the difference between each pKa (COOH) of each amino acid and

[*] Corresponding author. Tel.: +20 100172313; fax: +20 50 2244690.
E-mail address: elmaaty@mans.edu.eg (M.I. Abo el Maaty).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.06.031

the average of all pKa (COOH) of all twenty amino acids [20]. A 3D graphical representation of proteins was proposed based on five-letter model of amino acids which converts the twenty letters of amino acids to only five letters [21]. Then a vector of three components was assigned to each letter in the sequence as a point P_i (x_i, y_i, z_i). The final graph was obtained by connecting these points.

After obtaining a graph of a DNA or a protein sequence, it can be numerically characterized by a mathematical object as a matrix and then take some of this matrix invariants to describe the sequence. Examples of matrices of DNA sequences are euclidean distance matrix (ED), graph theoretical distance matrix (GD), path distance matrix (PD), L/L matrix and M/M matrix [7,22]. Matrix invariants are graph radius, normalized geometrical centers [12], leading eigenvalues of L/L matrix, M/M matrix [7], 2D coupling numbers [17] and 3D coupling numbers [21]. In this paper, we proposed a novel 3D graphical representation based on physicochemical properties of amino acids side chains. Then, we characterized our 3D graph numerically to analyze similarities among nine ND5 protein sequences. Finally, we compared our similarity/dissimilarity matrices with a percentage sequence identity matrix. This comparison was provided through a correlation and significance analysis.

2. 3D graphical representation of protein sequences

Amino acids are the basic building blocks of proteins. Each one of them consists of three parts; NH2 basic amino group at one of its terminals, COOH acidic carboxyl group at the other end lying between these two ends a single carbon atom Cα to which are attached a hydrogen atom and also a side group R. The great variety of possible protein structures stems from the fact that the side group R can have any of approximately 20 different chemical compositions. The different side groups interact with the surrounding environment (aqueous surroundings) in different ways, so the physical properties of proteins are determined by the sequence of its R groups [23].

We proposed a new 3D graphical representation of protein sequences based on the physicochemical properties of amino acid side chains. Depending on the polarity of the side chain, amino acids vary in their hydrophilic or hydrophobic character. Hydropathy index value of amino acids represents hydrophobic or hydrophilic properties of side chains [24]. The larger number is the more hydrophobic amino acid. We used it in our 3D graphical representation to represent *x*-coordinate. The amino acid side chain charge depends on the solution pH value. We used the charge at ($\text{pH} = 7$) to represent the *z*-coordinate. Most amino acids have neutral overall charge except Aspartic acid (Asp), Glutamic acid (Glu) have negative charges while Histidine (His), Lysine (Lys) and Arginine (Arg) have positive charges. We used mean accessible surface area (ASA) of side chains from protein fragments to represent *y*-coordinate. As mean ASA depends on the chain length (N), we chose it at ($N = 25$) [25]. The average of the twenty ASA equals 88.995. The *y*-coordinate value of each amino acid is calculated as follows:

$$Y_i = ASA_i - ASA_{\text{ave}} \quad \text{and}$$
$$ASA_{\text{ave}} = \frac{1}{20}\sum_{i=1}^{20} ASA_i.$$

A combination of hydropathy indices, charges and accessible surface areas were used in the field of protein function prediction [26]. These properties and *x*-, *y*- and *z*-coordinates of twenty amino acids are listed in Table 1. Each amino acid is represented numerically by a special vector of three values of *x*-, *y*- and *z*-coordinates. By translating the protein sequence' amino acids to their characterized vectors, our 3D graphical representation is obtained by summing these vectors. For example, if we have a protein sequence of n amino acids $S = s_1, s_2, s_3, \ldots, s_n$, so we have n points in the graph. Each point has three values P_i (x_i, y_i, z_i) which is calculated as follows:

$$X_i = \sum_{k=1}^{i} S_k^1, \qquad Y_i = \sum_{k=1}^{i} S_k^2 \quad \text{and} \quad Z_i = \sum_{k=1}^{i} S_k^3$$

where S_k^j ($j = 1, 2, 3$) represents the jth component of the vector corresponding to S_k. By connecting these points, we obtained our 3D graphical representation. We first applied our approach on the two shorter segments of protein of yeast Saccharomyces cerevisiae. The two segments are called protein I and protein II. Protein I and protein II sequences are “WTFESRNDPAKDPVILWLNGGPGCSSLTGL” and “WFFESRNDPANDPIILWLNGGPGCSSFTGL” respectively. Our 3D graphical representation of protein I and protein II is illustrated in Fig. 1(a) and (b) respectively. By looking at Fig. 1(a) and (b), we can find that the two protein curves are similar on the whole. The two protein segments have only four mismatching amino acids at positions 2, 11, 14 and 27. At the position “2” in protein I where “T” is replaced by “F” in protein II, this mismatch is accompanied a small relative distance.

3. Protein sequence data

Protein sequences that are used to prove our approach were downloaded from GenBank. The same data set were used before [20]. These data of nine ND5 (NADH dehydrogenase subunit 5) proteins are: human (*Homo sapiens*, AP_000649), gorilla (*Gorilla gorilla*, NP_008222), common chimpanzee (*Pan troglodytes*, NP_008196), pygmy chimpanzee (*Pan paniscus*,