Physica A 510 (2018) 477–485

Contents lists available at ScienceDirect

Physica A

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

A novel graphical representation and similarity analysis of protein sequences based on physicochemical properties

Mehri Mahmoodi-Reihani [a], Fatemeh Abbasitabar [b], Vahid Zare-Shahabadi [a,*]

[a] Department of Chemistry, Mahshahr Branch, Islamic Azad University, Mahshahr 6351977439, Iran
[b] Department of Chemistry, Marvdasht Branch, Islamic Azad University, Marvdasht 7371113119, Iran

HIGHLIGHTS

- 2D graphical representation of protein sequences is presented.
- 12 physicochemical properties of amino acids and principal component analysis were considered.
- A unique and meaningful numerical descriptive vector for a given sequence is obtained.
- A new similarity measure on the basis of moving window correlation coefficients is introduced.
- Depicting the moving window correlation coefficients leads to a graph which is easy to interpret.

ARTICLE INFO

Article history:
Received 29 January 2018
Received in revised form 5 July 2018
Available online xxxx

Keywords:
Protein sequence
Graphical representation
Principal component analysis
Physicochemical property
Moving window correlation coefficient

ABSTRACT

One of popular topic in bioinformatics is protein sequence analysis. The graphical representation of protein sequence is a simple and common way to visualize protein sequences. In this study, a numerical descriptive vector for a given protein sequence is calculated based on twelve physicochemical properties of amino acids (AAs) and principal component analysis (PCA). Each entry of the descriptive vector corresponds to one AA in the sequence. By this vector, an intuitive spectrum-like graphical representation of protein sequence is proposed. Squared correlation coefficient as well as moving window correlation coefficient, as a new similarity/dissimilarity measure, were used to compare different sequences. Applicability of the proposed method is assessed by analyzing the nine ND5 proteins. The results revealed the utility of the proposed method.

1. Introduction

Graphical representation of biological sequences facilitates quantitative comparisons and visual inspection of similarities/dissimilarities between biological sequences [1,2]. Compared to the alignment method, which is computationally expensive [3], graphical representation of biological sequence has several advantages and, therefore, many researchers have paid attention to develop new representation methods [4]. There are many graphical methods have been proposed to analyze and visualize DNA or protein sequences [2]. A DNA sequence is comprised of four types of nucleotides A, C, G, and T, while a protein is composed of the 20 amino acids (AAs). In comparison with proteins, appearing only four types AAs in DNA sequence makes it straightforward to graphical representation. Early graphical representations of DNA have been proposed by Hamori [5] in 1983, whilst the first graphical representation of proteins emerged twenty years later by Randić [6]. In the first attempt to graphical representation of proteins, a protein sequence converted into a hypothetical DNA sequence by

[*] Corresponding author.
E-mail address: valizare@gmail.com (V. Zare-Shahabadi).

https://doi.org/10.1016/j.physa.2018.07.011
0378-4371/© 2018 Published by Elsevier B.V.

assuming unique correspondence between each of 20 nucleotide triplets and 20 AAs [6]. By this conversion, it was possible to use available graphical representations of DNA to generate a graphical representation for proteins [6,7]. More recently, novel graphical approaches were developed for proteins that allow a direct representation of proteins as 2D mathematical objects, without a need to know which codons of RNA that encode amino acids were employed in the biosynthesis of proteins. In this regard, different geometric objects were defined by researchers. Randić [8] proposed a 2D graphical representation of proteins based on 2-D map of AAs. He obtained AA map by constructing the partial order on a complementary pair of AA properties. The same procedure was used by Wen et al. [9]. Later, more graphical representations in 2D or 3D space with two or three types of physicochemical properties were reported [10,11]. For example, three physicochemical properties of AA were employed by Maaty et al. [12] as x, y, and z coordinates to obtain a unique 3D graphical representation of protein sequences. Li et al. [13] used a five-letter model to reduce a protein primary sequence and proposed a new 3-D graphical representation of protein sequence based on the five-letter sequence. By this method which is similar to DNA representation some information may be lost. Representation of proteins based on generalized star graphs was done by Randić et al. [14]. Yao et al. examined the usage of six physicochemical properties of amino acids in the 2D graphical presentation of protein sequences, each time one of considered AA properties was employed to create 2D graph. They showed that meaningful 2D graphs for ND6 proteins could be obtained provided that isoelectric point or hydropathy was used [15]. Qu et al. considered 12 physicochemical properties of amino acids and tried to map each AA into the considered 12D space created by physicochemical properties, which did not result into clear and visible curve [16]. They reduced this 12D space to 1D by employing principal component analysis (PCA) in which only the first principal component was retained. Hu et al. [17] employed principal component analysis to reduce the space of nine main physicochemical properties of AAs. They then used a fractal method to construct a graphical representation of protein sequences. In 2012 the exact solution to the protein alignment was reported [18], and its computer program was developed in 2015 by Randić and Pisanski [19]. Strictly speaking, some of these methods utilizing physicochemical properties of AAs and some not. It was shown that 2D mapping of AAs using two or three complementary physicochemical properties of AAs may offer better insights in comparison with codon based representation. Normally, all the above-mentioned methods generate zigzag curves. Such curves offer visual inspection of similarity of different sequences. However, with the increase in length of sequence, one is going to lose visual representations with finer details [20–23].

In this article, we propose a novel way to calculate a numerical descriptive vector for a protein sequence based on 12 physicochemical properties and PCA. The 2D mapping of protein sequence is easily created by plotting the resulted descriptive vector. In addition, a new searching strategy, named moving window correlation coefficient, is introduced to identify similarity/dissimilarity regions between sequences. By this strategy, visual inspection of similarity between protein sequences is possible, even for long length sequences.

2. Method

2.1. Numerical characterization of a sequence

A protein sequence is made from up to 20 different amino acids. Each amino acid is usually represented by one alphabetic letter. In the present work, twelve physicochemical properties of the amino acids are considered in order to measure similarity/dissimilarity of protein sequences. The considered AA properties were hydrophobic parameter, pK$_a$(RCOOH), relative mutability, surrounding hydrophobicity in folded form, average relative fractional occurrence in AR(i), relative preference value at N_2, information measure for pleated-sheet, side chain hydropathy, normalized positional residue frequency at helix termini N, hydropathy scale based on self-information values in the two-state model, relative stability scale extracted from mutation experiments, and weighted diameter based on the atomic number [24,25]. Values of these physicochemical properties for all twenty AAs are given in Table 1.

Numerical descriptive vector of a protein sequence must contain some features about composition of amino acids and the position of each amino acid relative to the other nearby amino acids. Our proposed procedure for the graphical representation for protein sequences was partitioned into the following steps:

1. Let a protein sequence be $\Psi = s_1s_2s_3, \ldots, s_N$ with N amino acids where $s_i \in \{$G, A, V, L, I, F, S, T, Y, C, M, P, W, K, R, H, D, E, N, Q$\}$. For each amino acid in the sequence a numerical matrix X_i with size of $N \times L$ is created, where L is the number of considered physicochemical properties. For a sequence of length N, subsequently, N numerical matrices are created. The entries of ith X matrix is calculated by the following equation:

$$(x_{jl})_i = \frac{(a_{il} + \frac{1}{d_{ij}} a_{jl})}{2} j = 1, 2, 3, \ldots, N \,\&\, l = 1, 2, \ldots, L$$

where a_i is a vector contains the physicochemical properties of the ith AA and d_{ij} is

$$d_{ij} = \begin{cases} = 1 & \text{if } i = j \\ = dist(i, j) & \text{if } i \neq j \end{cases}$$

Here distance between the two AA in the sequence is calculated as Euclidean distance.