Computers in Biology and Medicine 57 (2015) 1–7

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A protein mapping method based on physicochemical properties and dimension reduction

Zhao-Hui Qi*, Meng-Zhe Jin, Su-Li Li, Jun Feng

College of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, 050043, People's Republic of China

ARTICLE INFO

Article history:
Received 24 July 2014
Accepted 19 November 2014

Keywords:
Protein sequence
Graphical representation
Physicochemical property
Dimension reduction
ND6 protein
H1N1

ABSTRACT

Background: The graphical mapping of a protein sequence is more difficult than the graphical mapping of a DNA sequence because of the twenty amino acids and their complicated physicochemical properties. However, the graphical mapping for protein sequences attracts many researchers to develop different mapping methods. Currently, researchers have proposed their mapping methods based on several physicochemical properties. In this article, a new mapping method for protein sequences is developed by considering additional physicochemical properties, which is a simple and effective approach.
Methods: Based on the 12 major physicochemical properties of amino acids and the PCA method, we propose a simple and intuitive 2D graphical mapping method for protein sequences. Next, we extract a 20D vector from the graphical mapping which is used to characterize a protein sequence.
Results: The proposed graphical mapping consists of three important properties, one-to-one, no circuit, and good visualization. This mapping contains more physicochemical information. Next, this proposed method is applied to two separate applications. The results illustrate the utility of the proposed method.
Discussion: To validate the proposed method, we first give a comparison of protein sequences, which consists of nine ND6 proteins. The similarity/dissimilarity matrix for the ssnine ND6 proteins correctly reveals their evolutionary relationship. Next, we give another application for the cluster analysis of HA genes of influenza A (H1N1) isolates. The results are consistent with the known evolution fact of the H1N1 virus. The separate applications further illustrate the utility of the proposed method.

1. Introduction

A graphical representation of a biological sequence is a powerful tool for analyzing and visualizing sequences. Many graphical methods have been designed to analyze DNA or a protein sequence. A DNA sequence is composed of four nucleic bases A, C, G and T. Comparing with the 20 amino acids in a protein sequence, four different symbols of the DNA sequence are beneficial to its graphical representation because of less symbol number. Early graphical methods of the DNA sequence include Hamori in 1983 [1], Hamori in 1989 [2] and Jeffrey in 1990 [3]. Afterwards, more graphical representations were well developed by researchers. For example, Nandy in 1994 [4] gave a graphical representation by assigning A, G, T and C to the four directions, (−x), (+x), (−y) and (+y), respectively. Bielinska-Waz [5–7], Liao [8–10], Randić [11–13], Jaklic [14], Li [15] and Qi [16–18] also proposed their graphical representations of the DNA sequence. These proposed graphical methods can provide direct insights into the local and global characteristics of DNA sequences.

Compared to the graphical representation of DNA sequences, the graphical representation of protein sequences is difficult because of the twenty amino acids. However, the advantages of graphical representation of protein sequences attract many researchers to develop different representation methods. The early graphical methods of protein sequences were presented in 2004 by Randić et al. [19,20]. They constructed the graphical representations of protein sequences based on the 64 codons. After that, researchers developed graphical representations of protein sequences in two main directions. One is not related to the physicochemical properties of amino acids. For instance, in 2006 Randić et al. [21] outlined a novel highly compact graphical representation for proteins based on a 'magic circle' of unit radius. In 2005 Bai and Wang [22] also proposed a 2D graphical representation of protein sequences based on nucleotide triplet codons. Recently, He et al. [23] presented a 3D graphical representation of protein sequences built on the Gray code. These methods don't take the physicochemical properties of amino acids into account.

The other type of graphical representation for protein sequences is closely related to the physicochemical properties of amino acids. The physicochemical properties of amino acids are essential when separating, purifying and studying protein sequence profiles, folding, classifying and functions. The physicochemical properties have

* Corresponding author.
E-mail address: zhqi_wy2013@163.com (Z.-H. Qi).

http://dx.doi.org/10.1016/j.compbiomed.2014.11.012
0010-4825/

strong effects on the pattern of protein evolution. This is also an important reason for connecting physicochemical properties with graphical representations for protein sequences. However, it is difficult to take more physicochemical properties into consideration when graphically representing a protein sequence in 2D or 3D space. In 2007, Randić [24] first presented a 2D graphical representation of proteins built on the partial order of a selected pair of physicochemical properties of amino acids. This is a good contribution. Later, more researchers proposed their graphical representations with two or three types of physicochemical properties [25,26,27] in 2D or 3D space. In [28], Randić gave more comments for the graphical representations considering physicochemical properties. These graphical representations share a common feature that a protein sequence has a corresponding graphical space curve. They contribute to comparing the similarities and dissimilarities of protein sequences. Then, Yao et al. [29] proposed a 2D graphical representation of protein sequences based on six physicochemical properties of amino acids. Yu et al. [30] presented a new protein map which incorporated ten properties of amino acids. These methods which consider more than three physicochemical properties are difficult to visualize. A protein sequence is likely mapped into a set of graphical curves.

The graphical representation of a protein sequence is confronted with two difficulties, good visualization and more major physicochemical properties. In this article, we propose a protein mapping method based on 12 major physicochemical properties and the PCA (Principal Component Analysis) method. The mapping without degeneracy and loss of information provides good visualization. With the graphical representation, the method is used to analyze the evolutionary relationship of two separate datasets. The results agree well with the known evolution fact and show the efficiency of our method. In addition, how to classify which proteins bind to nanoparticles in a sequence dependent manner is a major issue in medical applications in nanotechnology [31,32]. In [31], Monopoli et al. review the basic concept of the nanoparticle corona, and highlight how the properties of the corona may be linked to its biological impacts. If a nanoparticle corona has given evidence of its preference for a protein sequence, here the proposed mapping method can be used to find more candidate proteins. This is a possible application in medical applications in nanotechnology.

2. Methods

2.1. Protein sequence graphical representations based on physicochemical properties and the PCA method

Here, we consider 12 major physicochemical properties of amino acids, such as the chemical composition of the side chain [33], the polar requirement [34], the hydropathy index [35], the isoelectric point [36], the molecular volume [33], the polarity values [33], the aromaticity [37], the aliphaticity [37], the hydrogenation [37], the hydroxythiolation [37], the $pK1(-COOH)$ [25], and the $pK2(\text{-}NH_3^+)$ [25]. The characteristic physicochemical parameters of 20 amino acids are listed in Table 1. Based on the 12 physicochemical properties, we construct a 2D graphical representation of the protein sequence on two quadrants of the Cartesian coordinate system.

Good visualization is an important characteristic of graphic representation of the protein sequence. A normal resolution is to give a graphic representation for each property when we consider different physicochemical properties of amino acids. Then a protein sequence corresponds to several graphical curves. It is difficult for us to identify similar sequences from various groups of graphical curves, especially for longer proteins. A feasible method is to get the characteristic information from the 12 physicochemical properties, and show the distinguishing information in low-dimensional space. Based on the low-dimensional characteristic information, a protein sequence is easily represented as a graphical curve. The dimension reduction method can help us reach the goal for considering both different physicochemical properties and having good visualization. As for a simple and important dimension reduction method, principal component analysis (PCA) [38] is used to gather feature information from the 12 physicochemical properties.

Principal component analysis (PCA) was invented in 1901 by Karl Pearson [39]). In [38], readers can find the detailed description of this method. This method is a mathematical procedure that uses orthogonal transformation to analyze a data set and reduce it from a set of high-dimensional variables into a few hidden variables while keeping the principal information on its variability. The few hidden variables are known as the principal components. Recently, PCA and its expanded methods have been used to solve many biological problems [40–43]. Here we give a simple description for the PCA method. Let $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ denote the sample mean of

Table 1
Main 12 physicochemical characteristic values of 20 amino acids (*p1*: chemical composition of the side chain; *p2*: polar requirement; *p3*: hydropathy index; *p4*: isoelectric point; *p5*: molecular volume; *p6*: polarity; *p7*: aromaticity; *p8*: aliphaticity; p9: hydrogenation; *p10*: hydroxythiolation; *p11*: $pK1(-COOH)$; *p12*: $pK2(\text{-}NH_3^+)$.

Amino acids	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*	*P8*	*P9*	*P10*	*P11*	*P12*
A (*Ala*)	0	7.0	1.8	6.00	31	8.1	−0.11	0.239	0.33	−0.062	2.34	9.69
C (*Cys*)	2.75	4.8	2.5	5.07	55	5.5	−0.184	0.22	0.074	0.38	1.71	10.78
D (*Asp*)	1.38	13.0	−3.5	2.77	54	13.0	−0.285	0.171	−0.371	−0.079	2.09	9.82
E (*Glu*)	0.92	12.5	−3.5	3.22	83	12.3	−0.067	0.187	−0.254	−0.184	2.19	9.67
F (*Phe*)	0	5.0	2.8	5.48	132	5.2	0.438	0.234	0.011	0.074	1.83	9.13
G (*Gly*)	0.74	7.9	−0.4	5.97	3	9.0	−0.073	0.16	0.37	−0.017	2.34	9.60
H (*His*)	0.58	8.4	−3.2	7.59	96	10.4	0.32	0.205	−0.078	0.056	1.82	9.17
I (*Ile*)	0	4.9	4.5	6.02	111	5.2	0.001	0.273	0.149	−0.309	2.36	9.68
K (*Lys*)	0.33	10.1	−3.9	9.74	119	11.3	0.049	0.228	−0.075	−0.371	2.18	8.95
L (*Leu*)	0	4.9	3.8	5.98	111	4.9	−0.008	0.281	0.129	−0.264	2.36	9.60
M (*Met*)	0	5.3	1.9	5.74	105	5.7	−0.041	0.253	−0.092	0.077	2.28	9.21
N (*Asn*)	1.33	10.0	−3.5	5.41	56	11.6	−0.136	0.249	−0.233	0.166	2.02	8.80
P (*Pro*)	0.39	6.6	−1.6	6.30	32.5	8.0	−0.016	0.165	0.37	−0.036	1.99	10.60
Q (*Gln*)	0.89	8.6	−3.5	5.65	85	10.5	−0.246	0.26	−0.409	−0.025	2.17	9.13
R (*Arg*)	0.65	9.1	−4.5	10.76	124	10.5	0.079	0.211	−0.176	−0.167	2.17	9.04
S (*Ser*)	1.42	7.5	−0.8	5.68	32	9.2	−0.153	0.236	0.022	0.47	2.21	9.15
T (*Thr*)	0.71	6.6	−0.7	6.16	61	8.6	−0.208	0.213	0.136	0.348	2.63	10.43
V (*Val*)	0	5.6	4.2	5.96	84	5.9	−0.155	0.255	0.245	0.212	2.32	9.62
W (*Trp*)	0.13	5.2	−0.9	5.89	170	5.4	0.493	0.183	0.011	0.05	2.38	9.39
Y (*Tyr*)	0.20	5.4	−1.3	5.66	136	6.2	0.381	0.193	−0.138	0.22	2.20	9.11