Journal of Fiber Bioengineering and Informatics 7:1 (2014) 23–33
doi:10.3993/jfbi03201403

Novel 2D Graphic Representation of Protein Sequence and Its Application

Yongbin Zhao, Xiaohong Li, Zhaohui Qi*

College of Information Science and Technology, Shijiazhuang Tiedao University
Shijiazhuang 050043, China

Abstract

In recent years, more and more researchers presented their graphic representations for protein sequence. Here, we present a new 2D graphic representation of protein sequence based on four physicochemical properties of 20 amino acids. By this graphic representation we define a distance calculation formula to quantificationally calculate the similarity degree of different protein sequences. Then we apply the proposed graphical curve and new distance in the similarity/dissimilarity comparison of 10 ND5 proteins and the protein sub-cellular localization prediction about two common test datasets, ZD98 and CL317. The results show the proposed method is easy and effective.

Keywords: Graphic Representation; Protein Sequence; Sequence Similarity; Sub-cellular Location

1 Introduction

As bio-molecular sequences are rapidly growing, it is of great importance for people to reveal their meaning of life by analyzing these bio-molecular data. Now, people have developed many information methods to research the huge amounts of bio-molecular sequences. The graphic representation of biology sequences such as DNA or protein sequence has been a useful method to get important information from the primary sequences.

In 1983 [1], Hamori first tried to use graphical technique to research and analyze the DNA sequences. Thereafter, many researchers followed this research method and proposed some other graphic representations of DNA sequences [2-11]. For instance, Randić [7-9], Qi [5] and Yuan [10] introduced 2D or 3D graphic representations of DNA sequences. Liao [11] and Chi [3] presented higher dimensional methods to graphically represent DNA sequences. Moreover, researchers like Randić [7-9] presented some graphic representations to analyze protein sequences. Considering the physicochemical properties of amino acids Yao [12] proposed a graphical method based on the PK_α values of $COOH$ and NH_3^+ of the 20 amino acids. Xiao [13] and Wu [13] also developed their 2D graphic representations considering the physicochemical properties of 20 amino acids.

*Corresponding author.
Email address: zhqi_wy2013@163.com (Zhaohui Qi).

An important application of the proposed graphic representations of DNA or protein sequences is to get the similarities or dissimilarities among different biological sequences but not considering the alignment [12]. Based on the graphic representation of sequence, some mathematical descriptors like E, M/M, L/L [7, 8, 14] are developed to quantificationally compute the evolution distance among sequences. Recently, Liao et al. [15] proposes a new 2D graphic representation. They defined a distance computing formula to calculate the similarities or dissimilarities among different protein sequences. Then the method was used for protein sub-cellular localization prediction and got satisfactory results.

In this paper we propose a new graphic representation of protein sequences considering four main physicochemical properties of amino acids. According to the graphic representation we give a 20D characteristic vector to represent the corresponding protein sequence. The vector method can deal with sequences with different length. Then we use the characteristic vector extracted from graphic representation of protein sequence to analyze the similarities or dissimilarities of ten ND5 protein sequences. The test results show that the proposed method is a useful method in finding the similarities or dissimilarities among different protein sequences. Then, we utilized the similarity evaluation ability of the proposed method to take a prediction of sub-cellular localization of proteins. It is well-known that if two proteins are more similar, the two proteins are more likely to exist in the same sub-cellular location. We take two test datasets, the apoptosis proteins ZD98 and CL317. The prediction accuracy in jackknife test shows that the proposed method is effective on protein sub-cellular localization prediction.

2 Novel 2D Graphic Representation of Protein Sequences

Although proteins have many types and are different in nature and functions, they are composed of 20 native amino acids, which everyone knows have a variety of properties. People can study the structures and functions of proteins by these properties of amino acids. Some graphic representations of protein sequence are presented according to the physicochemical properties of amino acids. For example, Randić in [16] proposed a 2D graphic representation of protein. This method considered a pair of physicochemical properties of 20 amino acids, the pKa values of —NH_3 and —COOH. Yao et al. in [12] also proposed a dynamic 2D graphic representation of protein sequence based on the pKa values. In this paper, we present a novel 2D graphic representation based on four main physicochemical characteristic properties of 20 amino acids. They are relative molecular mass, isoelectric point, hydropathy index and melting point. The relative molecular mass can reflect the composition of the side chain. The isoelectric point is the pH value at which a particular surface or molecule carries no net electrical charge. The hydropathy index is a number representing the hydrophobic or hydrophilic properties of side chain of an amino acid. The melting point of amino acid is considered to be that temperature at which its crystalline substance becomes unreliable. The four physicochemical characteristic properties are essential for the protein structure and the catalytic activities of enzymes. The detailed data of the four physicochemical properties are listed in Table 1.

Observing Table 1, we find that similar amino acids have similar physicochemical properties. To clearly know the similarities/dissimilarities among amino acids, we take a similarity analysis of the 20 amino acids based on the four main physicochemical properties. However, the different physicochemical property values in Table 1 have different orders of magnitude. The different magnitude is likely to take bad effect on the similarity analysis of the amino acids. To eliminate