A new hybrid coding for protein secondary structure prediction based on primary structure similarity

Zhong Li, Jing Wang, Shunpu Zhang, Qifeng Zhang, Wuming Wu

PII: S0378-1119(17)30169-5
DOI: doi: 10.1016/j.gene.2017.03.011
Reference: GENE 41817

To appear in: *Gene*

Received date: 9 August 2016
Revised date: 2 February 2017
Accepted date: 15 March 2017

Please cite this article as: Zhong Li, Jing Wang, Shunpu Zhang, Qifeng Zhang, Wuming Wu , A new hybrid coding for protein secondary structure prediction based on primary structure similarity. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Gene(2017), doi: 10.1016/j.gene.2017.03.011

A new hybrid coding for protein secondary structure prediction based on primary structure similarity

Zhong Li [1,*], Jing Wang [1], Shunpu Zhang [2], Qifeng Zhang [1], Wuming Wu [1]

1. College of Science, Zhejiang Sci-Tech University, Hangzhou, 30018, China

2. Department of Statistics, University of Central Florida, Orlando, 32816, USA

Abstract: The coding pattern of protein can greatly affect the prediction accuracy of protein secondary structure. In this paper, a novel hybrid coding method based on the physicochemical properties of amino acids and tendency factors is proposed for the prediction of protein secondary structure. The principal component analysis (PCA) is first applied to the physicochemical properties of amino acids to construct a 3-bit-code, and then the 3 tendency factors of amino acids are calculated to generate another 3-bit-code. Two 3-bit-codes are fused to form a novel hybrid 6-bit-code. Furthermore, we make a geometry-based similarity comparison of the protein primary structure between the reference set and the test set before the secondary structure prediction. We finally use the support vector machine (SVM) to predict those amino acids which are not detected by the primary structure similarity comparison. Experimental results show that our method achieves a satisfactory improvement in accuracy in the prediction of protein secondary structure.

Keywords: Hybrid code; Protein secondary structure prediction; Protein primary structure; Support vector machine

1. Introduction

The prediction of protein secondary structure is a key step for the prediction of the 3D structure of a protein. It provides significant insights into protein functions. As the protein sequences stored in PDB grows exponentially, existing methods to make predictions are time consuming and could not meet the need of reality. Therefore, there is an urgent need to use amino acid sequence to predict the structure and function of proteins. In this paper, we focus on the prediction of protein secondary structure, which is an important research field in bioinformatics.

When the prediction accuracy of protein secondary structure is more than 80%, the 3D structure of a protein can generally be determined [1]. Many methods have been proposed for predicting the secondary structure of protein, such as methods by exploiting the physical and chemical properties of amino acids, methods based on sequence homology and statistical analysis, etc. [2-6]. An important distinction among these methods is the protein coding method. For example, Chen et al. [2] proposed an orthogonal coding method, which used a 20-bit binary vector to represent the twenty particular amino acids respectively. Bohr et al. [3] introduced a profile encoding method that the amino acid sequences to be predicted are compared by the multiple sequence alignments, and then looked for similar sequences in the homologous protein sequences in the database, and finally calculated the probability of a basic amino acid type appearing at each position of the amino acid sequence. Lamont [4] used codons (the amino acid were grouped and each group contained three bases) to predict the protein secondary structure, in which each base was represented by a 4-bit binary vector. Aydin et al. [5] provided a method based on the dynamic Bayesian network and support vector machine for the protein secondary structure prediction. Ding

[*] Corresponding author. Email: lizhong@zstu.edu.cn