Mathematical Biosciences 282 (2016) 61–67

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Detrended cross-correlation coefficient: Application to predict apoptosis protein subcellular localization

Yunyun Liang[*], Sanyang Liu, Shengli Zhang

School of Mathematics and Statistics, Xidian University, Xi'an 710071, PR China

ARTICLE INFO

Article history:
Received 21 April 2016
Revised 27 September 2016
Accepted 28 September 2016
Available online 6 October 2016

Keywords:
Subcellular localization
Detrended cross-correlation coefficient
Position-specific scoring matrix
Support vector machine

ABSTRACT

Apoptosis, or programed cell death, plays a central role in the development and homeostasis of an organism. Obtaining information on subcellular location of apoptosis proteins is very helpful for understanding the apoptosis mechanism. The prediction of subcellular localization of an apoptosis protein is still a challenging task, and existing methods mainly based on protein primary sequences. In this paper, we introduce a new position-specific scoring matrix (PSSM)-based method by using detrended cross-correlation (DCCA) coefficient of non-overlapping windows. Then a 190-dimensional (190D) feature vector is constructed on two widely used datasets: CL317 and ZD98, and support vector machine is adopted as classifier. To evaluate the proposed method, objective and rigorous jackknife cross-validation tests are performed on the two datasets. The results show that our approach offers a novel and reliable PSSM-based tool for prediction of apoptosis protein subcellular localization.

1. Introduction

Life science research have entered the post-genome era, and have focused on functional genomics. Protein subcellular localization is closely related to its function, and also maintain highly ordered cell guarantee for normal operation of the system [1]. Studies of protein subcellular localization is very helpful to understand the properties and functions of protein, understand the interaction between proteins and regulation mechanism, understand the pathogenesis of some diseases and develop new drug. However, the traditional biological experiments are both time consuming and costly. Therefore, development of fast and effective machine learning method for predicting protein subcellular localization is very necessary. In this paper, we will investigate apoptosis protein subcellular localization.

Apoptosis is a form of death the cell itself initiates, regulates, and executes using cellular and molecular machinery. so, apoptosis is also called “programed cell death”. When apoptosis malfunctions, a variety of formidable diseases can ensue: excessive apoptosis causes hypotrophy, such as in ischemic damage [2,3], whereas blocking apoptosis results in uncontrolled cell proliferation, such as cancer [4]. The function of an apoptosis protein is closely correlated with its subcellular location [5], therefore, predicting the subcellular location of apoptosis proteins is important and helpful to understand its mechanism and function. Over the past decade, various important methods that have been made to establish a really useful statistical predictor to solve this problem. These methods mainly focus on two aspects: the first aspect is feature extraction, by which the different length sequences are converted into a fixed-length vector. The methods include amino acid composition (AAC) [6,7], pseudo-amino acid composition (PseAAC) [8–10], dipeptide composition [11], encoding method with grouped weight [12], distance frequency [13] and discrete wavelet transform [14], the second aspect is a choice of favorable classification algorithm. Currently, the algorithm contains fuzzy k-nearest neighbor [15], support vector machine (SVM) [9,12–14], adaptive boosting technique [16], Bayesian classifier [17], hierarchical ensemble of Bayesian classifiers [17] and increment of diversity algorithm [11].

Despite some of the existing methods have shown the excellent performance, the information mainly is from apoptosis protein primary sequences. To extract the evolutionary information, which represented in the form of position-specific score matrix (PSSM) is adopted. In this paper, we focus on developing the features extraction technique to predict subcellular location of apoptosis proteins using the detrended cross-correlation (DCCA) coefficient of non-overlapping windows based only on PSSM. The DCCA coefficient (ρ_{DCCA}) [18] is a new method to quantify the level of cross-correlation between two non-stationary time series. This method is defined in terms of detrended fluctuation analysis (DFA) [19] and detrended cross-correlation analysis (DCCA) [20], and can be considered as an evolution of the DCCA method. The ρ_{DCCA} is applied in meteorology [21], in time series of homicide and attempted

[*] Corresponding author. Fax: +86 29 88202860.
E-mail address: yunyunliang88@163.com (Y. Liang).

http://dx.doi.org/10.1016/j.mbs.2016.09.019
0025-5564/© 2016 Elsevier Inc. All rights reserved.

Table 1
The compositions of two datasets adopted in this paper.

Dataset	Cy	Me	Mi	Se	Nu	En	Total
CL317	112	55	34	17	52	47	317
Dataset	Cy	Me	Mi	Other	–	–	Total
ZD98	43	30	13	12	–	–	98

homicide [22], in economy [23], in aero-engine system [24], among others. In this paper, each column of PSSM is a non-stationary time series, so we can obtain a ρ_{DCCA} for two different columns. PSSM of each protein sequence contains 20 different columns, hence, we can constructed a 190D feature vector for SVM classifier. To evaluate our method, jackknife test is employed on CL317 and ZD98 benchmark datasets, the experimental results demonstrate that our approach is a potential candidate for prediction of apoptosis protein subcellular localization.

2. Materials and methods

2.1. Datasets

In order to compare the accuracies of our prediction with the previous works, two benchmark datasets: CL317 and ZD98 are adopted in this paper. CL317 dataset is constructed by Chen and Li [8], and the ZD98 dataset is constructed by Zhou and Doctor [6]. Proteins in those datasets are extracted from SWISS-PROT [25] database. The CL317 consists of 317 apoptosis protein sequences, which include 112 cytoplasmic proteins (CY), 55 membrane proteins (ME), 34 mitochondrial proteins (MI), 17 secreted proteins (SE), 52 nuclear proteins (NU) and 47 endoplasmic reticulum proteins (EN). The ZD98 dataset consists of 98 apoptosis protein sequences, which include 43 cytoplasmic proteins (CY), 30 plasma membrane-bound proteins (ME), 13 mitochondrial proteins (MI) and 12 other proteins (OTHER). The number of proteins belonging to each class for CL317 and ZD98 datasets is listed in Table 1.

2.2. Feature extraction

One of the key steps in developing a powerful predictor for the apoptosis protein subcellular localization based on PSSM is to formulate the protein samples with an effective mathematical expression that truly reflect their intrinsic correlation. Here, we define detrended cross-correlation (DCCA) coefficient on PSSM to extract features.

2.2.1. Position-specific scoring matrix

Recently, multiple sequence alignment information in form of position-specific scoring matrix (PSSM) has been used for developing methods, and PSSM has been used for predicting protein subcellular localization. To extract the evolutionary information, PSSM is generated by using PSI-BLAST program [26] to search the SWISS-PROT database through three iterations with 0.001 as *E*-value for multiple sequence alignment against the protein sequence *P*. The (*i*, *j*)th entry of the position-specific scoring matrix represents the score of the amino acid residue in the *i*th position of the protein sequence being mutated to amino acid type *j* in the biology evolution process. PSSM is denoted as

$$PSSM = (P_1, P_2, \ldots, P_j, \ldots, P_{20}), \tag{2.1}$$

where $P_j = (P_{1,j}, P_{2,j}, \ldots, P_{L,j})^T \ (j = 1, 2, \ldots, 20)$, L represents the length of the protein sequence *P*, and *T* is the transpose operator. We further normalize each element for the original PSSM scores using the following sigmoid function:

$$f(s) = 1/(1 + e^{-s}), \tag{2.2}$$

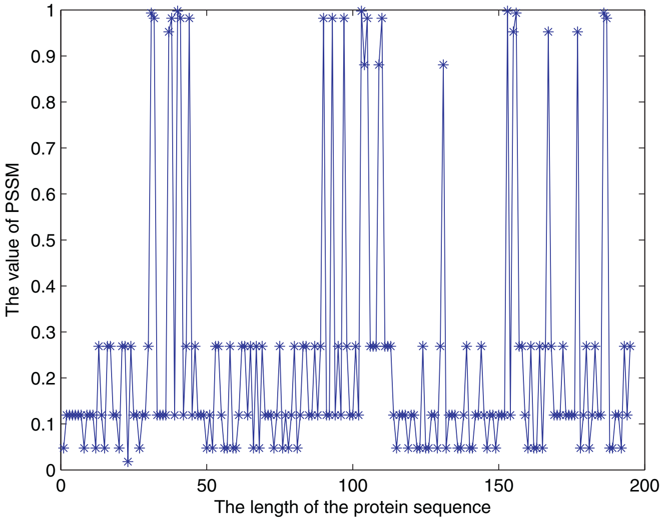

Fig. 1. Any column from any PSSM for CL317 dataset.

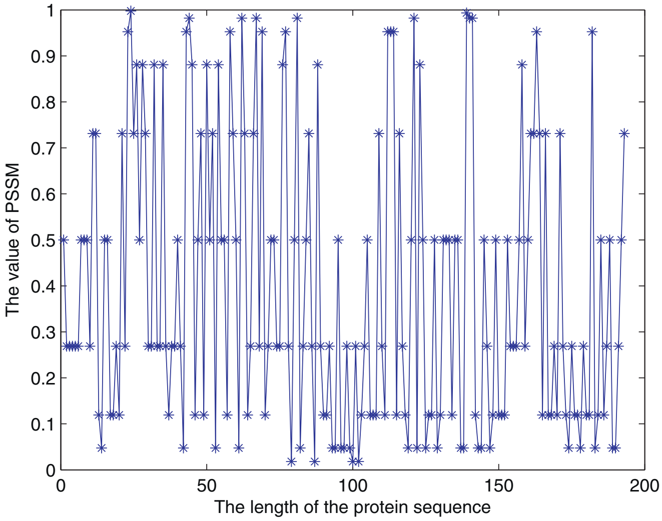

Fig. 2. Any column from any PSSM for ZD98 dataset.

where *s* represents the original value of each element for PSSM, this process can reduce the bias and noise contained in the original scores.

2.2.2. DCCA coefficient method based on PSSM

For non-stationary time series analysis, it is widely carried out by detrended fluctuation analysis (DFA), detrended cross-correlation analysis (DCCA), and detrended cross-correlation (DCCA) coefficient (ρ_{DCCA}). The DCCA coefficient method is an extension of the DCCA and the DFA. A protein sequence can be viewed as a non-stationary time series of the corresponding physicochemical property. Here, only the evolutionary information represented in the form of PSSM is adopted as the considered property. In this work, each amino acid is taken as one property and the PSSM is considered as the time series of all properties. Each PSSM of our adopted datasets contains 20 columns, in other words, contains 20 non-stationary time series. We randomly select any column from any PSSM for CL317 and ZD98 datasets, respectively. Figs. 1 and 2 indicate that the columns of PSSM can be viewed as non-stationary time series.

For two different columns of PSSM corresponding to the protein sequence *P*, $\{x_k\}$ and $\{y_k\}$, with $k = 1, 2, \ldots, L$, L represents the length of the protein sequence *P*. The DCCA coefficient method