Artificial Intelligence in Medicine 78 (2017) 41–46

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine

Qilin Xiang[a], Bo Liao[a], Xianhong Li[b], Huimin Xu[b], Jing Chen[b], Zhuoxing Shi[b], Qi Dai[b], Yuhua Yao[b,c,*]

[a] School of Information Science and Engineering, Hunan University, Changsha 410082, China
[b] College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
[c] School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, China

ARTICLE INFO

Article history:
Received 21 November 2016
Received in revised form 8 May 2017
Accepted 11 May 2017

Keywords:
Apoptosis protein
Position-specific scoring matrix
Golden section
Support vector machine

ABSTRACT

Objectives: In this paper, a high-quality sequence encoding scheme is proposed for predicting subcellular location of apoptosis proteins.
Methods: In the proposed methodology, the novel evolutionary-conservative information is introduced to represent protein sequences. Meanwhile, based on the proportion of golden section in mathematics, position-specific scoring matrix (PSSM) is divided into several blocks. Then, these features are predicted by support vector machine (SVM) and the predictive capability of proposed method is implemented by jackknife test
Results: The results show that the golden section method is better than no segmentation method. The overall accuracy for ZD98 and CL317 is 98.98% and 91.11%, respectively, which indicates that our method can play a complimentary role to the existing methods in the relevant areas.
Conclusions: The proposed feature representation is powerful and the prediction accuracy will be improved greatly, which denotes our method provides the state-of-the-art performance for predicting subcellular location of apoptosis proteins.

1. Introduction

The main goal of proteomics is functional annotation of unknown proteins. A critical comment is predicting protein subcellular localization. A large number of efficient and reliable computational approaches [1–6] have been developed to replace or assist the biological experiments, which is both time consuming and costly. The pioneering efforts to predict subcellular location from protein sequence were provided by Nakashima and Nishikawa [7]. Then dipeptide composition and gapped amino acid pairs [8] were proposed. However, ever since the concept of PseAAC was introduced by Chou [9], various modes have been used to predict subcellular localization and other fields. Meanwhile, in order to incorporate more sequence information, many other feature representations were incorporated in to PseAAC, which includes position-specific scoring matrix [10–12], functional domain [13–15], gene ontology [16,17] and so on. The present study is attempted to develop a computational approach for predicting the subcellular localization of apoptosis proteins in hope to stimulate the development of the relevant areas.

Apoptosis proteins, which is also called programmed cell death, is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death [18]. Apoptosis proteins are very important for understanding the mechanism of programmed cell death. These proteins play a central role in development and homeostasis of an organism [19,20]. Obtaining information about the subcellular location of apoptosis proteins is an important and helpful step towards understanding its mechanism and function. This is because the function of an apoptosis protein is closely correlated with its subcellular location [21,22]. Therefore, the study of subcellular location of apoptosis proteins is very important in biology.

Compared to lots of research on protein subcellular location prediction [2], studies on predicting apoptosis protein subcellular location are limited. It is mainly due to the limited number of experimentally validated apoptosis proteins in the database. Actually, many prediction algorithms have been reported in this regard. Subcellular location of apoptosis proteins was first studied by Zhou and Doctor [23]) in 2003. They used amino acid composition (AAC) and covariant discriminate algorithm to predict four kinds of subcellular locations of 98 apoptosis proteins and achieved 72.5% accuracy

* Corresponding author at: School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, China.
E-mail address: yaoyuhua2288@163.com (Y. Yao).

http://dx.doi.org/10.1016/j.artmed.2017.05.007
0933-3657/

by jackknife test. Latter Zhang et al. [24] constructed a new dataset ZW225 with four subcellular locations and adopted EBGW_SVM to predict subcellular localization. The accuracy is 83.1% in jackknife test. Meanwhile, Chen and Li [25,26] have developed two prediction approaches based on increment of diversity (ID) and increment of diversity with support vector machine (ID_SVM), which are validated on a new dataset covering six subcellular compartments and 317 apoptotic proteins and achieved 82.7% and 84.2% accuracy, respectively. Based on these datasets, many prediction algorithms have been put forward one by one, such as [27–32], which involve PseAAC with FKNN, PseAAC with SVM, distance frequency with SVM, auto covariance transformation and some other approaches.

The only difference for SVM-based in predicting subcellular localization is protein sequence encoding schemes. In this study, we introduce two novel feature extraction methods from position specific scoring matrix (PSSM). In order to know which is more closely related with subcellular localization of apoptosis proteins, evolutionary information and conservative information of protein sequences are considered. For PSSM, we also take different segmentations into consideration. Then we use the novel representation of protein sequence and combine with the support vector machine algorithm to predict apoptosis protein. The detailed description of the novel representation will be shown in the following section.

2. Materials and methods

2.1. Datasets

Two benchmark datasets constructed by the previous investigators are used in this work. The ZD98 dataset, constructed by Zhou and Doctor [23], has 43 cytoplasmic proteins, 30 plasma membrane-bound proteins, 13 mitochondrial inner and outer proteins and 12 other proteins. The second dataset, CL317 [25], are divided into six subcellular locations with 112 cytoplasmic proteins, 55 membrane proteins, 34 mitochondrial proteins, 17 secreted proteins, 52 nuclear proteins and 47 endoplasmic reticulum proteins. Proteins in these two datasets are extracted from SWISS-PROT database. Although two datasets have small size, they are widely used in previous studies.

2.2. Feature extraction from position specific scoring matrix

As it has been proved that evolutionary information is more informative than the sequence itself [33]. Evolutionary relationship also had been widely used in many researches such as prediction of protein function [34], membrane proteins types [35], proteins structural class [36] as well as sub-cellular localization of proteins [37]. In this study, two kinds of feature extraction methods are given based on PSSM. We first use each protein sequence as a seed to search and align homogenous sequences from SWISS-PROT database using the PSI-BLAST [38] with parameters h and j set to 0.001 and 3, where h and j denote the E-value threshold for inclusion in PSSM and the maximum of iterations, respectively. Then the protein sequences can be expressed by a PSSM matrix denoted by P. The matrix is composed of $L \times 20$ elements, where L is the length of a given protein sequence. The element $P_{i,j}$ which indicates the relative probability of j-*th* amino acid at the i-*th* location of the protein sequence during biological evolution processes.

2.2.1. Golden section

For the convenience of calculation, the elements in PSSM are scaled to the range from 0 to 1 using the following sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

where x is the original value in the *PSSM*.

After getting the PSSM matrix, we compute the average replaced possibility for all 20 types of amino acids, and finally 20 features are obtained. It can be formulated as

$$\bar{P} = \left[\bar{P}_1, \bar{P}_2, \cdots, \bar{P}_2\right]^T \tag{2}$$

where

$$\bar{P}_j = \frac{1}{L}\sum_{i=1}^{L} P_{i,j}\, j = 1, 2, \cdots, 20 \tag{3}$$

To explore potential information embedded in physicochemical properties of the amino acids, we introduce 544 physicochemical properties values taken from the amino acid index (AAindex) [39]. These physicochemical properties values are used as weight of $\bar{P}_j$, then each given protein can be formulated as:

$$\bar{P}_w = \left[index(1,i)\bar{P}_1, index(2,i)\bar{P}_2, \cdots, index(20,i)\bar{P}_{20}\right]^T \quad i = 1, 2, \cdots, 544 \tag{4}$$

where $index(t,i)(t = 1, 2, \cdots, 20)$ is the selected i-*th* physicochemical property for 20 different amino acids.

In order to extract important information from sequence fragment, many researchers split the sequence into several pieces and then calculate amino acid information of each fragment. As is well known, golden section has great importance in many fields. Accordingly, we introduce golden section to split PSSM and propose segmented evolutionary information. Then segments of a protein sequence is divided by a golden partition value 0.618. First, a given protein sequence, which can be divided into two parts: 0.618*L and L-0.618*L. For each part, we also do the same calculate as Eq. (4). Continually, we split it again for two parts, respectively. Until the segmentation is accomplished by k times. Finally, a protein sequence can be represented by a $2^k \times 20$ dimensional vector.

2.2.2. Evolutionary information and conservative information

PSSM is a $L \times 20$ matrix, in which the value of each element represents the possibility of a particular residue being substituted. The elements in PSSM are positive values, negative values or 0. If the value of $P_{i,j}$ is positive value, we can say that the i-*th* position amino acid of the query sequence being mutated to amino acid type j easily in the process of evolution, and this is called evolutionary mutation. Moreover, the greater the values, the greater the probability of evolution. On the contrary, if $P_{i,j}$ is negative value, then the mutation is not easily and conservative. The smaller the values, the more conservative. Therefore, in order to research the influence of evolutionary information and conservative information of protein sequences on apoptosis proteins subcellular localization, we divide the sequences into two parts. Here, zero value has not been taken into account.

For the PSSM of each given protein sequence, if $P_{i,j}$ is positive, then this value belong to array M otherwise belong to array N.

These two blocks can be represented as:

$$M = \left\{PSSM|P_{i,j} > 0\right\}, N = \left\{PSSM|P_{i,j} < 0\right\}$$

$$i = 1, 2, \cdots, L, j = 1, 2, \cdots, 20.$$

After this, amino acid composition is calculated for these two blocks respectively. At the same time, for the comparison of different forms of information amount, we also consider the amino acid composition of the entire sequence.

2.3. Classification algorithm

After the features are extracted, various classification algorithms can be used to implement apoptosis protein prediction, such as support vector machine (SVM) [40], ensemble classifier [41],