Chemometrics and Intelligent Laboratory Systems 158 (2016) 146–154

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A segmentation based model for subcellular location prediction of apoptosis protein

Qi Dai [a,*], Sheng Ma [a], Yabin Hai [a], Yuhua Yao [a], Xiaoqing Liu [b,*]

[a] College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China
[b] College of Sciences, Hangzhou Dianzi University, Hangzhou 310018, People's Republic of China

ARTICLE INFO

Article history:
Received 28 October 2015
Received in revised form
15 April 2016
Accepted 14 September 2016
Available online 14 September 2016

Keywords:
Subcellular location
Position information
Evolutionary profile
Golden ratio
Support vector machine

ABSTRACT

Subcellular location is very useful to understand the mechanism and functions of apoptosis proteins. Various efficient methods have been proposed to predict subcellular location prediction, but challenges still exist. In this paper, we proposed a segmentation based model to improve subcellular location prediction. In three experiments, the proposed model reported robust results and demonstrated better performance compared with existing methods, which can be contributed to the introduction of the segmentation because it makes the N-segment and C-segment of the proteins gotten differential treatment in subcellular location prediction. This understanding can be useful to design more powerful method for predicting subcellular location. The software, data and supplement material are freely available at http://bioinfo.zstu.edu.cn/GoldenP/.

1. Introduction

Apoptosis, known as programmed cell death (PCD), is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death [1,2]. Cells undergoing apoptosis usually exhibit a characteristic morphology, including fragmentation of the cell into membrane-bound apoptotic bodies, nuclear and cytoplasm condensation and hemolytic cleavage of the DNA into small oligo-nucleosomal fragments [3,4]. Although apoptosis plays a key role in development and tissue homeostasis [1,4], its aberrant activation may contribute to a variety of formidable diseases. For example, blocking apoptosis is related to with cancer [5,6] and autoimmune disease, whereas unwanted apoptosis can possible lead to ischemic damage [7] or neurodegenerative disease [8]. Therefore, it is necessary to study on the mechanism of the apoptosis in order to find many targets for therapeutic intervention [2,9].

Previous studies have shown that the functions of the apoptosis proteins are closely associated with their subcellular locations [10–17]. With the help of the cell fractionation, electron microscopy and fluorescence microscopy, we can determine the subcellular location of apoptosis proteins. But these experimental methods are expensive and time-consuming. With the increasing of the number of the proteins, there is a great need to develop a reliable and effective computational method to predict protein subcellular location.

Because of the importance of the subcellular location prediction, various efficient methods have been proposed to address this problem. Several studies indicated that the subcellular location of apoptosis proteins is strongly related with amino acid composition (AAC), which permits researchers to find them by comparing the divergence of their AAC [17,18]. However, these AAC-based methods treat each peptide or polypeptide separately and, consequently, lose sequence-order information [19]. To overcome this limitation, some new descriptions of the amino acids were proposed, such as pseudo-amino acid composition (PseAAC) [10,11,20,21], functional domain composition [12,22], Gene Ontology (GO) [13,23] and profile of PSI-BLAST [16,24]. Moreover, several classification methods have already been introduced in subcellular location prediction, including support vector machine (SVM) [25–27], neural network [28], fuzzy k-NN [29] and classifier fusion technique [14,15,30].

Recently, evolution information was used to improve the prediction of the subcellular location. Given a query sequence, we searched it against a database of the proteins with the help of the position-specific iterated BLAST (PSI-BLAST), a search tool hanging the double sequence alignment and the multiple sequence alignment together. Position Specific Scoring Matrix (PSSM) can be then obtained from PSI-BLAST profile to represent evolutionary information of protein sequences [31]. This information was first

* Corresponding authors.
E-mail addresses: daiailiu04@yahoo.com (Q. Dai), xiaoqingliu@hdu.edu.cn (X. Liu).

http://dx.doi.org/10.1016/j.chemolab.2016.09.005
0169-7439/

proposed to study protein subcellular localization [32], and widely used in many prediction and classifications areas. For example, Proteome Analyst computed evolution features for classification based on the homologous sequences token from Swiss-Prot database [33]. Guo et al. extracted some features from the position-specific scoring matrix (PSSM) and used them to predict the subcellular location for Gram-negative bacteria proteins [34]. Mundra et al. introduced standard sigmoid function to process the PSSM and combined it with the pseudo amino acid composition to predict protein subnuclear localization [35]. Rashid et al. computed the amino acid, dipeptides and PSSM composition features and combined them to develop more efficient prediction method based on SVM machine [36]. Shen and Chou proposed the pseudo position specific scoring matrix (PsePSSM) and developed a new web-server for predicting protein subnuclear localization, Nuc-PLoc [16]. Zou et al. summed up all rows in the PSSM and used the sequence length to normalized them, and further scaled them to the range of (0, 1) with standard sigmoid function [37]. Xiao et al. computed the auto covariance of position specific scoring matrix and applied them in protein sub-nuclear localization prediction [38].

All above methods have achieved better performance in the subcellular location prediction, but challenges still remain. First, some methods focus mostly on the content of amino acids and, consequently, ignore the position distribution of the amino acids along the sequences. Second, PSSM is processed as a whole matrix without considering which segments represent the most important signals for protein subcellular location prediction. Third, efforts should be made to solve computational problems when searching a long sequence against the Swiss-Prot database. In many cases, only a small part of the Swiss-Prot database are useful for protein subcellular location prediction, and the rest will add the noise to evolutionary profile and reduce the efficiency of prediction model.

To address these problems, we reported herein a segmentation based model for protein subcellular location prediction. We first split a protein sequence into subsequences N-segment and C-segment, in which the ratio of the N-segment to the C-segment is golden ratio. We then computed the AAC and AAP of these subsequences as well as the evolutionary profile. At last, we used principal component analysis to reduce high-dimensional feature space and then fed them into support vector machine to predict the subcellular location of the apoptosis proteins. Through three experiments, we want to address effectiveness of the proposed segmentation based model in comparison with the available competing methods, and whether the amino acid composition, amino acid position and evolutionary profile achieve an improvement when using segmentation technique.

2. Materials and methods

2.1. Datasets

In order to facilitate the comparison and analysis, three datasets from SWISS-PROT (version 49.5) were used in this work. The first dataset is referred to as ZD98. It consists of 43 cytoplasmic proteins, 30 plasma membrane-bound proteins, 13 mitochondrial proteins and 12 other proteins [39]. The secondary dataset, ZW225, contains 41 nuclear proteins, 70 cytoplasmic proteins, 25 mitochondrial proteins and 89 membrane proteins [40]. The third dataset is referred to as CL317. It consists of 112 cytoplasmic proteins, 55 membrane proteins, 34 mitochondrial proteins, 17 secreted proteins, 52 nuclear proteins and 47 endoplasmic reticulum proteins [41].

2.2. Golden ratio

As for two quantities, if the ratio of their sum to the one is equal to the ratio of the larger one to the smaller one, we infer that they are in the golden ratio (*Gr*). For example, *a* and *b* are given quantities, if

$$(a + b)/a = a/b = Gr. \tag{1}$$

They are said to be in the golden ratio. A general way to find *Gr* is to start with the left fraction. Through simplifying the fraction, we get

$$(a + b)/a = 1 + b/a = 1 + 1/Gr, \tag{2}$$

then

$$1 + 1/Gr = Gr, \tag{3}$$

By rearranging them, we obtain

$$Gr^2 - Gr - 1 = 0. \tag{4}$$

Using the quadratic formula, we get two solutions are

$$Gr = (1 \pm \sqrt{5})/2. \tag{5}$$

Because this ratio between length and width of a rectangle should be non-zero, the positive solution must be chosen

$$Gr = (1 + \sqrt{5})/2 = 1.6180339887.$$

It is well known that the golden ratio (0.382/0.618) is widely in the nature, which is also denoted as the golden section or golden mean [42]. Here, we introduced the golden ratio to separate the biological sequences and PSSM matrix to construct segmentation based model.

2.3. Segmentation based model

Various information has been extracted to predict the subcellular location of the apoptosis proteins, such as pseudo-amino acid composition (PseAAC) [10,11,20,21], functional domain composition [12,22], Gene Ontology (GO) [13,23] and profile of PSI-BLAST [16,24]. In this study, we proposed a segmentation based model to extract novel information of the apoptosis proteins for subcellular location prediction.

2.3.1. Amino acid composition based on golden ratio (AAC_Gr)

In word statistics, a biological sequence can be regarded as a succession of symbols and further analyzed the distribution of its small subsequences. A *k*-word *w* is a series of *k* consecutive letters in a sequence *s*. The total number of occurrence of the word *w* in the sequence *s* is the count of the *k*-word *w*, in which these *k*-words can be overlap in the sequence.

With help of *k*-word counts, a sequence can be represented by an *n*-dimensional vector C_k

$$C_k = (c(w_{k,1}), c(w_{k,2}), \cdots, c(w_{k,n}),), \tag{6}$$

where *c(w)* denotes the count of the *k*-word *w* in the sequence *s*, and *n* is the size of the *k*-words set. Then the frequencies of the *k*-words, denoted by F_k, can be calculated as

$$F_k = \left(c(w_{k,1})/(m - k + 1), c(w_{k,2})/(m - k + 1), \cdots, c(w_{k,n})/(m - k + 1)\right). \tag{7}$$

where *m* is the length of the given sequence *s*. In fact, the frequencies of the 1-words are amino acid composition (*AAC*).

The N-terminus refers to the start of a protein terminated by an amino acid with a free amine group (-NH2), and the C-terminus is the end of an amino acid chain, terminated by a free carboxyl group (-COOH). Because the N-terminus of a peptide chain is