Prediction of protein subcellular location using a combined feature of sequence

Qing-Bin Gao*, Zheng-Zhi Wang, Chun Yan, Yao-Hua Du

Institute of Automation, National University of Defense Technology, Changsha 410073, People's Republic of China

Received 9 April 2005; revised 10 May 2005; accepted 10 May 2005

Available online 31 May 2005

Edited by Takashi Gojobori

Abstract **To understand the structure and function of a protein, an important task is to know where it occurs in the cell. Thus, a computational method for properly predicting the subcellular location of proteins would be significant in interpreting the original data produced by the large-scale genome sequencing projects. The present work tries to explore an effective method for extracting features from protein primary sequence and find a novel measurement of similarity among proteins for classifying a protein to its proper subcellular location. We considered four locations in eukaryotic cells and three locations in prokaryotic cells, which have been investigated by several groups in the past. A combined feature of primary sequence defined as a 430D (dimensional) vector was utilized to represent a protein, including 20 amino acid compositions, 400 dipeptide compositions and 10 physicochemical properties. To evaluate the prediction performance of this encoding scheme, a jackknife test based on nearest neighbor algorithm was employed. The prediction accuracies for cytoplasmic, extracellular, mitochondrial, and nuclear proteins in the former dataset were 86.3%, 89.2%, 73.5% and 89.4%, respectively, and the total prediction accuracy reached 86.3%. As for the prediction accuracies of cytoplasmic, extracellular, and periplasmic proteins in the latter dataset, the prediction accuracies were 97.4%, 86.0%, and 79.7, respectively, and the total prediction accuracy of 92.5% was achieved. The results indicate that this method outperforms some existing approaches based on amino acid composition or amino acid composition and dipeptide composition.**

Keywords: Protein subcellular location; Combined feature; Amino acid composition; Dipeptide composition; Physicochemical property; Nearest neighbor; Jackknife test

1. Introduction

In post-genome era, the explosive growth of biological data is increasingly widening the gap between the number of protein sequences deposited in public databases and the experimental annotation of their functions. Protein subcellular location data are a valuable information resource helpful in elucidating protein functions, knowing the compartment in which a protein resides may give important insights as to its structure and function. There are three conventional approaches being applied to experimental validation of protein subcellular locations, namely, the cell fractionation, electron microscopy and fluorescence microscopy [1], whereas, these approaches are time-consuming and costly. Therefore, developing a reliable computational method for predicting protein subcellular locations would be very significant for genome annotation.

A number of protein subcellular localization methods have been developed in the past. Most of them can be divided into two categories: one is based on the identification of protein N-terminal sorting signals and the other is based on amino acid composition [2]. Nakai and Kanehisa [3,4] initially developed an expert system and knowledge base for predicting protein subcellular locations using their N-terminal sorting signals. Subsequently, a computational program based on the same approach, called PSORT [5], was presented. TargetP [6] is another congeneric prediction system. Several machine learning methods [7–9] have been proposed to detect such sorting signals, the most popular one is SignalP [10]. The reliability of this method is strongly dependent on the gene 5′-region or protein N-terminal sequence assignment. However, the assignments of gene 5′-regions are usually unreliable in genome sequencing projects.

Method based on amino acid composition was proposed by Nakashima and Nishikawa [11]. They concerned two subcellular locations, i.e., the intracellular and extracellular compartments. From then on, there are many different approaches have been introduced to predict protein subcellular locations by amino acid composition or dipeptide composition. Cedano et al. [12] suggested a prediction program called ProtLock using the Mahalanobis distance. Reinhardt and Hunnard [13] used neural networks. Chou and Elrod [14] proposed a covariant discriminant algorithm. Yuan [15] constructed a Markov chain model. Fujiwara and Asogawa [16] integrated neural networks with hidden Markov model. Furthermore, Hua and Sun [17], Park and Kanehisa [18] adopted support vector machine (SVM). Huang and Li [19] introduced a fuzzy *k*-nearest neighbors algorithm. However, the methods mentioned above might miss some sequence order and sequence length information, concerning this information may improve the prediction performance to some extent. Some efforts focused on this purpose have been made in recent years.

Chou [20] first introduced a new concept of quasi-sequence-order to reflect the sequence order effect based on the physicochemical distance between amino acids, and a remarkable improvement was observed in the prediction results using the augmented covariant discriminant algorithm. Later, Chou [21] proposed another novel concept called pseudo-amino acid composition. Chou and Cai [22] defined the so-called

*Corresponding author.
E-mail address: gqb_kd@yahoo.com.cn (Q.-B. Gao).

0014-5793/$30.00 © 2005 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2005.05.021

functional domain composition. Recently, a hybrid approach [23–26] by incorporating pseudo-amino acid composition, functional domain composition and gene ontology (GO) was proposed. Other methods, such as Zp curve [27], lexical analysis [28], combining evolutionary and structural information [29], LOCtarget [30], hybrid modules [31], spectral analysis technique [32], supervised locally linear embedding (SLLE) technique [33], complexity measure factor [34], cellular automata [35] and digital signal processing approach [36] were also suggested.

In this paper, we have introduced a nearest neighbor based method for predicting protein subcellular location via a combined feature of protein primary sequence, which consists of amino acid composition, dipeptide composition and physicochemical properties. This method supplies a novel technique for extracting features from protein primary sequence and achieves a high prediction performance in a jackknife test. The dipeptide composition can be considered as a representation form of proteins incorporating neighborhood information and has been used by previous investigators in predicting protein subcellular locations [16,18,19] and protein secondary structure contents [37,38]. On the other hand, the physicochemical properties of amino acids may influence the structure and function of a protein, considering them will provide some helpful information for protein subcellular localization.

2. Materials and methods

2.1. Dataset

The dataset constructed by Reinhardt and Hunnard [13] was adopted in our work. Proteins in this dataset were extracted from Swiss-Prot 33.0 and all transmembrane proteins were excluded, for transmembrane proteins could be properly predicted by several existing methods [39,40]. The dataset consists of 3424 non-redundant proteins with less than 90% sequence identity whose subcellular locations are experimentally determined, including 2427 eukaryotic proteins and 997 prokaryotic proteins. The former contains 684 cytoplasmic, 325 extracellular, 321 mitochondrial and 1097 nuclear proteins, and the latter contains 688 cytoplasmic, 107 extracellular and 202 periplasmic proteins, as shown in Table 1.

2.2. Combined feature vector

Reviewing the existing prediction methods mentioned before, most of them used amino acid composition as the basic feature to represent protein primary sequence. Amino acid composition denotes the occurrence frequencies of the 20 amino acids in the protein sequence and is usually represented by a 20D feature vector $\vec{S}_a$, written as:

$$\vec{S}_a = [p_1, p_2, \ldots, p_{20}]^{\mathrm{T}}$$

where p_i ($i = 1,2,\ldots,20$) is the occurrence frequency of amino acid i, T is the transpose operator. Obviously, it loses the sequence order and sequence length information completely. However, this information may have some close correlation with protein subcellular location. Therefore, to improve the prediction performance, we have to incorporate it via other protein features. In the present work, the dipeptide composition and physicochemical properties of amino acid are utilized to represent a protein together with amino acid composition.

Dipeptide composition attempts to extract the information about amino acid composition along local order of amino acid. By using this component, we can add some sequence order information into the amino acid composition vector. The dipeptide composition denotes the occurrence frequencies of two consecutive residues in the primary sequence and thus deduces a 400D feature vector $\vec{S}_b$, described as:

$$\vec{S}_b = [q_1, q_2, \ldots, q_{400}]^{\mathrm{T}}$$

where q_j ($j = 1,2,\ldots,400$) is the occurrence frequency of each dipeptide.

Furthermore, the physicochemical properties of amino acid residues have a deep influence on protein structure and function, incorporating such effect might lead to a prospective improvement on protein subcellular localization. Thus, we also use the auto correlation function based on physicochemical properties to interpret a protein. It comprises two steps. The first involves the transformation of the protein primary sequence into a numerical series, i.e., $h_1, h_2, \ldots, h_L$, here h_l ($l = 1,2,\ldots,L$) is the amino acid index for the lth residue in the protein sequence, and L is the length of protein sequence. Each amino acid residue in the primary sequence is replaced by the value of amino acid index [41], which represents the physicochemical property of the residue. The 10 physicochemical properties are selected in our work to represent a protein as shown in Table 2. The auto correlation functions r_k are then calculated according to the following expressions [42]:

$$r_k = \frac{1}{L-k}\sum_{l=1}^{L-k} h_l h_{l+k}, \quad k = 1, 2, \ldots, 10.$$

Finally, we obtain the 10D feature vector $\vec{S}_c$, given by:

$$\vec{S}_c = [r_1, r_2, \ldots, r_{10}]^{\mathrm{T}}.$$

Now, we add the components of $\vec{S}_b$ and $\vec{S}_c$ on $\vec{S}_a$ to form a combined feature vector $\vec{S}$, defined by:

$$\vec{S} = \left[\vec{S}_a, \vec{S}_b, \vec{S}_c\right]^{\mathrm{T}}.$$

In the nearest neighbor algorithm, this vector is accepted to represent a protein sequence.

2.3. Nearest neighbor algorithm

The nearest neighbor algorithm [43] is one of the simplest and oldest methods for performing general, non-parametric classification. It is attractive because it is easy to implement and has a low probability of error [44]. Furthermore, the nearest neighbor classifier often gives competitive performance compared with other complex machine learning methods. Nearest neighbor methods have been used for the prediction of protein secondary structure [45] and protein β-turn [46]. Cai

Table 1
The number of protein sequences in each subcellular location

Organism	Subcellular location	Number of proteins
Eukaryotic	Cytoplasmic	684
	Extracellular	325
	Mitochondrial	321
	Nuclear	1097
Prokaryotic	Cytoplasmic	688
	Extracellular	107
	Periplasmic	202

Table 2
The 10 physicochemical properties for amino acids (derived from [40])

Property	Reference
Refractivity	Jones (1975)
Flexibility	Bhaskaran and Ponnuswamy (1988)
Volume	Chothia (1984)
Transfer free energy to surface	Bull and Breese (1974)
Electron–ion interaction potential	Cosic (1994)
pK of side chain	White et al. (1978)
Hydrophilicity	Hopp and Woods (1981)
Polarity	Ponnuswamy et al. (1980)
Hydrophobicity	Eisenberg et al. (1984)
Isoelectric point	Zimmerman et al. (1968)