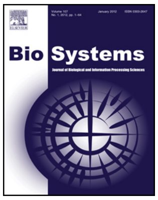
BioSystems 139 (2016) 37–45

Contents lists available at ScienceDirect

BioSystems

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



A multiple information fusion method for predicting subcellular locations of two different types of bacterial protein simultaneously

Jing Chen[a,1], Huimin Xu[a,1], Ping-an He[b], Qi Dai[a], Yuhua Yao[a,*]

[a] College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
[b] College of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China

ARTICLE INFO

Article history:
Received 26 May 2015
Received in revised form 8 October 2015
Accepted 10 December 2015
Available online 24 December 2015

Keywords:
Physicochemical properties
Position-specific score matrix
Gene ontology
Principal component analysis
Support vector machine

ABSTRACT

Subcellular localization prediction of bacterial protein is an important component of bioinformatics, which has great importance for drug design and other applications. For the prediction of protein subcellular localization, as we all know, lots of computational tools have been developed in the recent decades. In this study, we firstly introduce three kinds of protein sequences encoding schemes: physicochemical-based, evolutionary-based, and GO-based. The original and consensus sequences were combined with physicochemical properties. And elements information of different rows and columns in position-specific scoring matrix were taken into consideration simultaneously for more core and essence information. Computational methods based on gene ontology (GO) have been demonstrated to be superior to methods based on other features. Then principal component analysis (PCA) is applied for feature selection and reduced vectors are input to a support vector machine (SVM) to predict protein subcellular localization. The proposed method can achieve a prediction accuracy of 98.28% and 97.87% on a stringent Gram-positive (Gpos) and Gram-negative (Gneg) dataset with Jackknife test, respectively. At last, we calculate "absolute true overall accuracy (ATOA)", which is stricter than overall accuracy. The ATOA obtained from the proposed method is also up to 97.32% and 93.06% for Gpos and Gneg. From both the rationality of testing procedure and the success rates of test results, the current method can improve the prediction quality of protein subcellular localization.

1. Introduction

Determination of protein subcellular localization can provide valuable information in elucidating the interactions between different proteins and other molecules, and understanding the mechanisms of human disease (Xiao et al., 2011a,b). Among all the proteins, bacterial proteins are special, because of the wide range of both harmful and useful roles they play in biological interactions. Bacterial can be divided into two groups: Gram-positive (Gpos) and Gram-negative (Gneg). For those secreted proteins, especially released from gram-negative bacteria, they are known to be a potential cause of a disease. As an example, the proteins located at the outer membrane of a *Leptospira interrogans* cell or those secreted from the cell are likely to stimulate the Leptospirosis disease (Viratyosin et al., 2008). The importance of bacteria, regardless of being Gpos and Gneg, is because they are the active elements on many useful biological interactions; meanwhile, they are the source of many diseases which makes it crucially important to determine their functions especially for drug and vaccine design (Gardy and Brinkman, 2006).

Currently, a multitude of protein sequences are increasingly identified and piled up into public biology databanks, which results from the development of high-throughput technology, in the post-genomic era. However, experimentally determining the subcellular localization of a protein is a laborious and time consuming task. Facing such an avalanche of new protein sequences, it is both challenging and indispensable to develop an automated method for fast and accurately annotating the subcellular attributes of uncharacterized proteins. Through the development of new approaches in computer science, coupled with an increased dataset of proteins of known localization, computational tools can now provide fast and accurate localization predictions for many organisms. Therefore, subcellular localization prediction is becoming more and more challenging. Meanwhile, this problem could be solved by bioinformatics better.

In the process of prediction, the most crucial steps include: (1) the construction of the benchmark dataset; (2) protein feature

[*] Corresponding author.
E-mail address: yaoyuhua2288@163.com (Y. Yao).
[1] These authors contributed equally to this work as co-first authors.

http://dx.doi.org/10.1016/j.biosystems.2015.12.002
0303-2647/© 2015 Elsevier Ireland Ltd. All rights reserved.

representation; (3) a powerful classification algorithm. According to the extracted feature type, there are two kinds of features representation. (a) Sequence-based is derived from protein sequences, which includes amino acid compositions (Nakashima and Nishikawa, 1994), N-terminal amino acid sequences (Nakai and Kanehisa, 1991; Emanuelsson et al., 2000; Horton et al., 2006), pseudo-amino acid compositions (Chou, 2001, 2005; Chou and Cai, 2003; Zuo et al., 2014; Mandal et al., 2015; Zhu et al., 2015), physicochemical-based (Cai et al., 2010; Wang and Li, 2013) and evolutionary-based (PSSM) (Jeong and Lin, 2011; Wu et al., 2011; Huang and Yuan, 2013; Nanni et al., 2014; Zhang et al., 2012). (b) The other is annotation-based methods, which make use of the correlation between the annotations (usually the functional annotations) of a protein and its subcellular localization. Among them, methods based on gene ontology (GO) information are more attractive (Wan et al., 2012). GO-based methods (Shen et al., 2007; Chou and Shen, 2007a,b; Chou et al., 2011) make use of the well-organized biological knowledge about genes and gene products in the GO databases. Actually, the essence of why using GO approach to represent protein samples can significantly improve the prediction quality is due to the fact that proteins mapped into the GO database space would be clustered in a way better reflecting their subcellular locations, thus to significantly enhances the success rate of prediction for those proteins that do not have significant sequence homology to proteins with known locations (Chou and Shen, 2007a,b). In recent years, GO similarity based approach of gene similarity has been put forward to predict subcellular localization, functional similarity and other many biological problems. Since the relationship between GO terms could reflect the association between different gene products (Wan et al., 2014), so the semantic similarity between GO terms is used to represent the similarity between different protein sequences. The higher the degree of similarity, the greater the likelihood of the two sequences located in the same location.

Besides the benchmark dataset, not only protein sequence information but also prediction algorithms could affect the accuracy of the subcellular localization prediction. In the past decades, a wide range of classification techniques have been used, such as the hidden Markov models (HMM) (Lin et al., 2011), neural network (Zou et al., 2007), K-nearest neighbor (Shen and Chou, 2006, 2010a,b; Xiao et al., 2011a,b), Random forest (Breiman, 2001) and support vector machine (Qiu et al., 2010; Wan et al., 2015). Among them, SVM is particularly attractive for prediction analysis due to its computational efficiency in processing multidimensional datasets with complex relationships among the data elements (Dou et al., 2014). Moreover, SVM is readily adaptable to new data, allowing model updates in parallel with the continuing increase of biological databases. In order to show the powerful application of SVM, SVM and Random forest were chose as the prediction algorithm in our model.

In this study, a protein sequence can be represented by fusing the sequence information, evolution information and GO information to represent a protein sample. First, a protein sequence and its consensus sequence amino acid composition were combined by the value of a given physicochemical property for an amino acid. The second feature, which is extracted from position-specific scoring matrix (PSSM), is the improvement of auto covariance transformation (*PSSM-AC* model) (Dehzanqi et al., 2015a,b; Liu et al., 2012). The third representation is based on GO, because the most significant enhancement for protein subcellular localization prediction accuracy has been achieved by using gene ontology (Lin et al., 2013). After this, the above three feature representations were fused. Then, before put these vectors into the classifier, we apply the principal component analysis (PCA) algorithm to extract the essential features (low-dimensional vectors) from the original high-dimensional vectors. Finally, SVM is used to classify and Jackknife cross-validation tests were employed to validate the results.

2. Materials and methods

2.1. Datasets

In this work, two benchmarks that have been widely used in the literature for Gpos and Gneg subcellular localizations were adopted. The first dataset, Gpos523 (Chou and Shen, 2006, 2008; Shen and Chou, 2010a,b), consists of 523 Gpos bacterial protein sequences and has less than 25% pairwise sequence similarity within each subcellular localization. The second dataset is Gneg1456, including 1456 protein. As described in many references (Shen and Chou, 2007; Chou and Shen, 2009; Wu et al., 2012) homology bias but meanwhile cover as many locations as possible, we conduct a sequence identify cutoff procedure to make sure none of the proteins has 25% pairwise sequence identity to any other in a same subset. The information of the benchmark dataset is listed in Table 1.

2.2. Feature extraction

Ever since Chou et al. (Chou, 2001) put forward the concept of pseudo amino acid composition, it has been widely used to study various problems in proteins and protein-related systems, such as references (Karakasidis et al., 2009; Mohabatkar, 2010; Yu et al., 2010; Mei, 2012; Zhang et al., 2014). Regardless of what descriptors were used, the final input must be a vector containing a set of discrete components. According to a comprehensive review (Chou, 2011), the general form of PseAAC for a protein sequence P is

$$P = (\varphi_1\varphi_2 \ldots \varphi_u \ldots \varphi_\Omega)^T \quad (1)$$

where Ω is the fixed length of the descriptor and depends on how to extract the information from the amino acid sequence.

To extract the evolutionary information, the profile of each protein sequence is generated by running *PSI-BLAST* (Schaffer et al., 2001) program against the SWISS-PROT database with parameters h and j set to 0.001 and 3, respectively, where h and j denote the

Table 1
The detailed information of the two datasets.

Gpos523		Gneg1456	
Subcellular location	Number of proteins	Subcellular location	Number of proteins
Cell membrane	174	Cell inner membrane	557
Cell wall	18	Cell outer membrane	124
Cytoplasm	208	Cytoplasm	410
Extracell	123	Extracellular	133
		Fimbrium	32
		Flagellum	32
		Nucleoid	8
		Periplasm	180
Total	523	Total	1456