Journal of Theoretical Biology 335 (2013) 205–212

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions

Chao Huang *, Jing-Qi Yuan

Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

HIGHLIGHTS

- It is an initial study for prediction of subchloroplast proteins with multiple sites.
- Several efficient approaches of feature extraction are used.
- Using the protein evolution information can improve the prediction performance.
- A better success rate can be received by using the OET_KNN or ET_KNN instead of KNN.

ARTICLE INFO

Article history:
Received 4 February 2013
Received in revised form
26 May 2013
Accepted 29 June 2013
Available online 10 July 2013

Keywords:
Pseudo amino acid composition
ET_KNN
OET_KNN
Sub-subcellular localization
Subchloroplast locations

ABSTRACT

Owing to the fact that location information can indicate important functionalities of proteins, developing computational tools to predict protein subcellular localization is one of the most efficient and meaningful tasks with no doubt. The existence methods dealing with prediction of protein subchloroplast locations can only handle the case of single location site. Therefore, it is meaningful and challenging to make effort in how to deal with the proteins with multiple subchloroplast location sites instead of just excluding them. To solve this problem, new systems for predicting protein subchloroplast localization with single or multiple sites are developed and discussed in the paper. Three different editions of KNN algorithms and four different types of feature extraction were adopted to construct the prediction systems. This is the first effort to predict the proteins with multiple subchloroplast locations. The overall jackknife success rates achieved by the best combination (features+classifier) on three dataset with different levels of homology were 89.08%, 81.29% and 71.11%. The performance of the prediction models indicate that the proposed methods might be applied as a useful and efficient assistant tool for the prediction of sub-subcellular localizations.

1. Introduction

Chloroplasts are crucial organelles existing in eukaryotes organisms and plant cells (Hu and Yan, 2012; Shi and Theg, 2013; Wang et al., 2009). They are surrounded by two layers of membrane, and are believed to play important roles in photosynthesis and cellular metabolism (Kleffmann et al., 2004). Similar to different subcellular locations existing in a cell, the chloroplast can be also divided into several subcellular locations (Du et al., 2009a; Ferro et al., 2003; Hu and Yan, 2012; Shi et al., 2011): (1) stroma, (2) thylakoid lumen, (3) thylakoid membrane, and (4) envelop. Given a particular subchloroplast protein, it is highly desired to know which subchloroplast location it belongs to because this kind of information is highly correlated with its function and its role in some biological processes. Moreover, it is also worth noticing that the location of certain proteins can change along with the biological status change of the cell or organelles, and play different role under these different conditions.

The success of various genome projects created large number of protein sequences storing in public biology databank. Conducting various experiments to identify different locations are unpractical because these types of experiments are both time-consuming and expensive. Hence it is highly desired to develop efficient computational tools for predicting protein subchloroplast localizations.

Although several efforts in dealing with the prediction of proteins with only single subchloroplast location have already been proposed in last decade or so (Du et al., 2009b; Hu and Yan, 2012; Shi et al., 2011), unfortunately, to best of our knowledge, no

* Corresponding author. Tel.: +86 15 821254988.
E-mail address: huangchao_sjtu@aliyun.com (C. Huang).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.06.034

efforts in dealing with the prediction of proteins with multiple subchloroplast locations have ever been made until nowadays.

This study was initiated in an attempt to develop methods to predict the subchloroplast localization of proteins with single or multiple sites.

2. Materials and methods

2.1. Dataset

Protein subchloroplast dataset was received from the (UniprotKB/Swiss-Prot) database at (http://www.ebi.ac.uk/uniprot/) released on November 2012. The detailed procedures are as follows: (1) open the website at http://www.uniprot.org/. (2) Click the button "advanced", followed by selecting "Subcellular Location" for "Fields"; typing in "chloroplast stroma" for "Term"; and selecting "Experimental" for "Confidence". (3) Click the button "Add&Search", select "or", and repeat the step (2). The only difference is typing in one of these terms "chloroplast thylakoid lumen", "chloroplast thylakoid membrane", "chloroplast envelope", once orderly until all of them are used. (4) Click the button "Add&Search"; choose "and"; select "Fragment (yes/no)" for "Field", and choose "no", (5) click the button "Add&Search"; choose "and"; select "Sequence Length" for "Field", and choose the sequence length "≥50".

Three new homology-reduced datasets, the S80 dataset, the S60 dataset and S40 dataset, were constructed with sequence identity cut off value 80%, 60% and 40% by using the CD-HIT software (Huang et al., 2010; Li and Godzik, 2006; Niu et al., 2010). We construct these datasets in order to study the models' performance on dataset with different levels of homology.

The information of benchmark dataset S80, S60 and S40 are listed in Tables 1–3.

2.2. Feature extraction

To avoid completely losing the sequence-order information, the pseudo amino acid composition (PseAAC) was proposed (Chou, 2001a, 2005) to replace the simple amino acid composition (AAC) for representing the sample of a protein. Ever since the concept of PseAAC was proposed in 2001 (Chou, 2001a), it has penetrated into almost all the fields of protein attribute predictions (Chang et al., 2013; Chen et al., 2009; Chen and Li, 2013; Ding et al., 2009; Esmaeili et al., 2010; Georgiou et al., 2009; Gu et al., 2010; Jiang et al., 2008; Khosravian et al., 2013; Li and Li, 2008; Lin, 2008; Lin et al., 2008; Mohabatkar, 2010; Mohabatkar et al., 2011; Mohabatkar et al., 2013; Qiu et al., 2010; Wan et al., 2013; Yu et al., 2010; Zeng et al., 2009; Zhang et al., 2008a; Zhou et al., 2007). Recently, the concept of PseAAC and its general form (Eq. (6) of (Chou, 2011a)) was further extended to represent the feature vectors of DNA and nucleotides (Chen et al., 2013, 2012b) and even various biological samples such as tissues from patients (Huang et al., 2012; Li et al., 2012). Because it has been widely and increasingly used, recently two powerful softwares called PseAAC-Builder (Du et al., 2012) and propy (Cao et al., 2013) were established for generating various special Chou's PseAAC modes, in addition to the web-server PseAAC built in 2008. In this study, we attempted to develop novel classifiers based on three special modes of Chou's PseAAC.

Table 1
Detail of the subchloroplast benchmark dataset S80 derived from Swiss-Prot database according to the procedures described in Section 2.1 and processed by the CD-HIT at the similarity level 80%.

Oder	Location	Number of proteins
1	Stroma	150
2	Thylakoid lumen	58
3	Thylakoid membrane	895
4	Envelope	372
Total number of locative proteins		1475
Total number of different proteins		1440

*Of the 1440 different proteins, 1406 belongs to only 1 location, 33 to 2 locations, 1 to 3 locations, i.e., total 1475 locative proteins.

Table 2
Detail of the subchloroplast benchmark dataset S60 derived from Swiss-Prot database according to the procedures described in Section 2.1 and processed by the CD-HIT at the similarity level 60%.

Oder	Location	Number of proteins
1	Stroma	115
2	Thylakoid lumen	45
3	Thylakoid membrane	407
4	Envelope	272
Total number of locative proteins		839
Total number of different proteins		813

*Of the 813 different proteins, 788 belongs to only 1 location, 24 to 2 locations, 1 to 3 locations, i.e., total 839 locative proteins.

Table 3
Detail of the subchloroplast benchmark dataset S40 derived from Swiss-Prot database according to the procedures described in Section 2.1 and processed by the CD-HIT at the similarity level 40%.

Oder	Location	Number of proteins
1	Stroma	95
2	Thylakoid lumen	36
3	Thylakoid membrane	217
4	Envelope	192
Total number of locative proteins		540
Total number of different proteins		522

*Of the 522 different proteins, 505 belongs to only 1 location, 16 to 2 locations,1 to 3 locations, i.e., total 540 locative proteins.

2.2.1. Amino acid composition (AAC)

Using amino acid composition only is the simplest method of protein representation (Sahu and Panda, 2010). In this method, a 20-dimensional vector is used to represent a protein sample. The feature vector of a protein sample can be represented by

$$V = (1/L)[v_1, v_2, \ldots, v_{20}] \qquad (1)$$

where L is the length of the protein sample, and v_i represents the *ith* residue occurrence frequency in this protein sample.

2.2.2. Pseudo amino acid based features (PseACC)

This is a type of protein descriptor proposed by (Chou, 2001b) which avoid losing sequence ordering information from the protein samples.

Suppose a protein including L amino acid residues can be represented as

$$P = Q_1 Q_2 Q_3 Q_4 \ldots Q_L \qquad (2)$$

where Q_1 is the residue at the first position along the sequence and Q_2 the residue at the second position and so forth.

The sequence-order information can be indirectly represented by the following equations:

$$\delta_\theta = \sum_{i=1}^{L-\theta} \Omega(Q_i, Q_{i+1})/(L-\theta), \quad (\theta = 1, 2, \cdots, n \ \text{ and } \ n < L) \qquad (3)$$

where L denotes the length of the protein and the δ_θ is called the θth correlation factor which harbors the sequence order information between all the θ most contiguous residues. The correlation