Bioinformatics, 31(16), 2015, 2639–2645
doi: 10.1093/bioinformatics/btv212
Advance Access Publication Date: 20 April 2015
Original Paper

OXFORD

Sequence analysis

MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier

Xiao Wang[1,*], Weiwei Zhang[1], Qiuwen Zhang[1] and Guo-Zheng Li[2,*]

[1]School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China and [2]Department of Control Science and Engineering, Tongji University, Shanghai 201804, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on December 19, 2014; revised on March 29, 2015; accepted on April 13, 2015

Abstract

Motivation: Identifying protein subchloroplast localization in chloroplast organelle is very helpful for understanding the function of chloroplast proteins. There have existed a few computational prediction methods for protein subchloroplast localization. However, these existing works have ignored proteins with multiple subchloroplast locations when constructing prediction models, so that they can predict only one of all subchloroplast locations of this kind of multilabel proteins.

Results: To address this problem, through utilizing label-specific features and label correlations simultaneously, a novel multilabel classifier was developed for predicting protein subchloroplast location(s) with both single and multiple location sites. As an initial study, the overall accuracy of our proposed algorithm reaches 55.52%, which is quite high to be able to become a promising tool for further studies.

Availability and implementation: An online web server for our proposed algorithm named MultiP-SChlo was developed, which are freely accessible at http://biomed.zzuli.edu.cn/bioinfo/multip-schlo/.

Contact: pandaxiaoxi@gmail.com or gzli@tongji.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Chloroplasts are organelles in most green plant cells, and also exists in some eukaryotic organisms, such as seaweed. Chloroplast's main function is to conduct photosynthesis, where they capture and store the energy from the sunlight, transform it to chemical energy, and finally release oxygen from water. In addition to the important photosynthesis, they are also responsible for carrying out a lot of other functions, including fatty acid synthesis and the immune response in plants. Chloroplast proteins play different roles in the biological processes mentioned above. Knowing the functions of these proteins is of significant value. Due to the very close relationship between the functions and localizations of chloroplast proteins, identifying the subchloroplast localizations of these proteins in chloroplast organelle is very helpful for understanding the function of chloroplast proteins.

With in-depth study of cell organelles, the researchers have found a number of substructures in cell organelles, such as nuclear chromatin, heterochromatin, nuclear envelope, nucleolus in the nucleus, inner membrane, outer membrane in the mitochondria, stroma, thylakoid in the chloroplast and so on. In order to more deeply understand the function of these proteins, it is necessary to identify the subchloroplast localizations of these proteins in the organelle level. As can be seen from the recently released UniProtKB/Swiss-Prot database (release 2013_05), there are a total of 14 408

chloroplast proteins, where 7367 chloroplast proteins have subchloroplast localization annotations, accounting for about 7367/14408 = 51% of all chloroplast proteins. However, of the 7367 chloroplast proteins with subchloroplast localization annotations, 6955 are annotated with experimental confidence and 412 annotated with uncertain terms such as 'by similarity', 'probable' and 'potential'. Therefore, there are only 6955 chloroplast proteins with determined subchloroplast localization annotations, which are about 6955/14408 = 48% of all chloroplast proteins. In other words, subchloroplast localizations of more than half of the chloroplast proteins need to be further identified either by computational methods or biological experiments. Since the organelle is a more microscopic structure unit relative to the cell, it is more difficult and time-consuming to determine protein sub-subcellular localizations by biological experiments. With explosively increasing of protein sequences in the post-genomic era, the gap between the sequence number and function annotations of chloroplast proteins will become more and more wide. To fill the gap, it is very necessary to develop new computational prediction methods for predicting protein subchloroplast localizations.

In recent years, there have existed a few computational prediction methods for protein sub-subcellular localizations, such as, subnuclear localization prediction (Kumar *et al.*, 2014; Shen and Chou, 2005), submitochondria localization prediction (Du and Li, 2006; Lin *et al.*, 2013a) and subchloroplast localization prediction (Du *et al.*, 2009; Saravanan and Lakshmi, 2013) concerned in this article. In the field of subchloroplast localization prediction, the first pioneering work is completed by Du *et al.* (2009). They constructed the first public benchmark dataset in the field, and developed a new predictor based on pseudoamino acid composition (PseAAC) and the evidence-theoretic K-nearest neighbor (ET-KNN) classifier. Since then, other researchers have also proposed a few prediction methods for protein subchloroplast localization. However, these existing works have the following drawbacks: (i) only considering the small number of subchloroplast locations, reducing practicality of prediction methods, (ii) the homologous bias of the benchmark dataset adopted is too large to accurately evaluate the performance of prediction methods, for instance, the dataset constructed by Du *et al.* has as high as 60% protein sequence similarity, (iii) more importantly, ignoring proteins with multiple subchloroplast locations when constructing prediction models, so that these existing works will not be able to completely predict all the subchloroplast locations of these multilabel proteins.

It has been observed that more and more proteins have multiple subcellular or sub-subcellular locations. Actually, multilabel proteins should not be ignored because they may have some special biological functions worthy of our in-depth notices. Recently, some multilabel predictors have been developed for predicting subcellular localization of proteins with both single and multiple sites, such as, iLoc-Plant (Wu *et al.*, 2011), iLoc-Animal (Lin *et al.*, 2013b), PlantmPLoc (Chou and Shen, 2010) and MLPred-Euk (Wang and Li, 2013). A review article (Chou, 2013) has concluded recent advances in predicting multilabel attributes in molecular biosystems.

However, none of the existing multilabel predictors was specially developed for chloroplast proteins. In this article, we aim to develop a novel computational method that can predict single and multiple subchloroplast locations for chloroplast proteins simultaneously by overcoming the above three problems. Firstly, we constructed a new benchmark dataset that contains chloroplast proteins with both single and multiple subchloroplast locations. To enlarge the prediction scope, the number of subchloroplast locations concerned is increased from 4 to 5. To avoid redundancy and homology bias, protein sequence similarity in the new-constructed dataset is controlled under 40%. Secondly, through selecting features that are most related to each subchloroplast location, respectively, and taking into account correlations between different subchloroplast locations simultaneously, we proposed a novel multilabel algorithm that can accurately predict multilabel protein subchloroplast localizations. Finally, we developed an online web server for our proposed algorithm called MultiP-SChlo freely accessible at http://biomed.zzuli.edu.cn/bioinfo/multip-schlo/.

2 Materials and Methods

2.1 Dataset

In this article, the chloroplast protein sequences used here were collected from the UniProtKB/Swiss-Prot database (release 2013_05). To obtain a high quality benchmark dataset, we preprocessed the raw data in the database according to the following steps:

1. Only those proteins with single or multiple subchloroplast localization annotations were collected. In this article, the following five subchloroplast locations, Envelope, Stroma, Thylakoid lumen, Thylakoid membrane and Plastoglobule, were used to search the UniProtKB/Swiss-Prot database (release 2013_05). Among them, Plastoglobule is newly added in this study.
2. Ambiguous annotated proteins, such as 'by similarity', 'probable' and 'potential' were excluded, because they might introduce noisy data for decreasing prediction performance.
3. Those proteins annotated as 'fragment' or containing less than 50 amino acid residues were excluded.
4. Proteins containing ambiguous letters, like 'B', 'X' or 'Z', were excluded.
5. To avoid redundancy and homology bias, protein sequence similarity in the new-constructed dataset was controlled fewer than 40% using the CD-HIT program (Fu *et al.*, 2012). Due to shortage of available chloroplast protein sequences, we do not choose a smaller sequence similarity cutoff value. When there are more chloroplast proteins available in future, we will further lower the cutoff value for obtaining more rigorous benchmark dataset.

Through the above process, we obtained the benchmark dataset MSchlo578 including 578 chloroplast proteins distributed in the following 5 main subchloroplast locations, Envelope, Stroma, Thylakoid lumen, Thylakoid membrane and Plastoglobule, of which 556 proteins are associated with one subchloroplast location, 21 with two subchloroplast locations, one with three subchloroplast locations, and none with four and more subchloroplast locations. The number of chloroplast proteins belonging to each subchloroplast location are presented in Table 1. The number of chloroplast proteins with different number of subchloroplast locations can be found in Figure 1. The benchmark dataset is provided in Supplementary Data or can be downloaded from our online web server (http://biomed.zzuli.edu.cn/bioinfo/multip-schlo/).

Table 1. Breakdown of the chloroplast protein benchmark dataset MSchlo578

Order	Compartment	Number of proteins
1	Envelope	199
2	Stroma	105
3	Thylakoid lumen	34
4	Thylakoid membrane	233
5	Plastoglobule	30

2.2 Feature extraction

In various kinds of protein attributes prediction based on machine-learning techniques, including protein subchloroplast localization prediction of course, an important task is to extract discriminative features for building a powerful classifier. Of the feature extraction models for a protein, the simplest one is its amino acid (AA) composition or AAC (Nakashima *et al.*, 1986). This model extracts 20 discrete numbers to represent the occurrence frequencies of the native amino acids in a protein. So far, many protein attributes prediction methods were based on the AAC model (Jahandideh *et al.*, 2007; Nakashima and Nishikawa, 1994; Reinhardt and Hubbard, 1998; Zhou and Doctor, 2003). However, if using the ACC model to represent a protein, all its sequence order effects would be lost, and hence the prediction quality might be considerably limited. To avoid losing the sequence order effects hidden in protein sequences, the PseAAC model was proposed (Chou, 2001, 2005) to replace AAC for vectorizing protein sequences. Since the concept of PseAAC was proposed in 2001, it has rapidly penetrated into many fields of protein attribute prediction, such as predicting allergenic proteins (Mohabatkar *et al.*, 2013), predicting supersecondary structure (Zou *et al.*, 2011), predicting DNA-binding proteins (Fang *et al.*, 2008), predicting enzyme family and sub-family classes (Qiu *et al.*, 2010; Zhou *et al.*, 2007), among many others. Due to the extensive use of PseAAC, several powerful softwares or web servers have been established for conveniently generating various modes of PseAAC, such as the web-server PseAAC (Shen and Chou, 2008), the tools PseAAC-Builder (Du *et al.*, 2012), propy (Cao *et al.*, 2013) and PseAAC-General (Du *et al.*, 2014).

In this article, we also applied the PseAAC model. PseAAC vectorizes a protein into a $(20+\xi\cdot\lambda)$-D feature vector, in which the first 20 components are the conventional AAC and the next $\xi\cdot\lambda$ components reflect the sequence-order effects between the amino acids in the protein. The number of features in the PseAAC vector is controlled by two important parameters: the number of amino acid indices selected (ξ) and the maximum number of correlation tiers (λ) along a protein sequence. In this article, the following six amino acid indices ($\xi=6$) are used to calculate the correlation factors between the amino acids along the protein sequence: (i) hydrophobicity, (ii) hydrophilicity, (iii) mass, (iv) pK (alpha-COOH), (v) pK (NH3) and (vi) pI (at 25°C). It should be noted that λ must be smaller than the length of the shortest protein sequence in the training set. In the extreme case where $\lambda=0$, the PseAAC is degenerated to the conventional AAC. In this study, the maximum number of correlation tiers is set to be 50 because the length of the shortest protein sequence in the training set is 51. Thus the dimensionality of PseAAC feature vector is $20+6\cdot 50=320$. The reason for setting λ to the maximum available value in this study is that our proposed prediction algorithm can automatically select the optimal feature subset with no need for tuning in advance. Next, we will introduce our proposed multilabel prediction algorithm in detail.

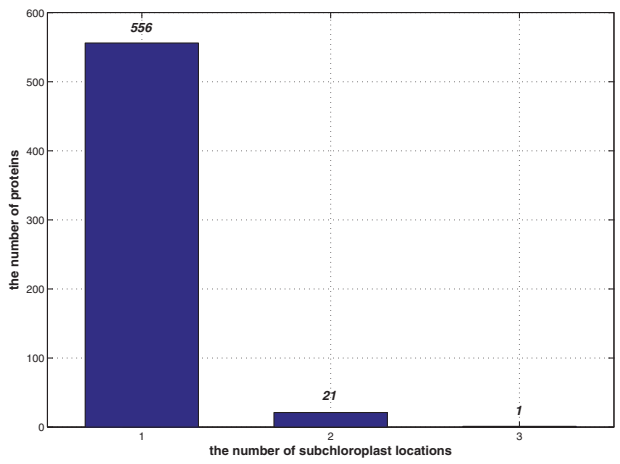

Fig. 1. The distribution of proteins with different number of subchloroplast locations

2.3 The proposed prediction algorithm

As reported in Li *et al.* (2012), exploiting label correlations and leveraging label-specific features can, respectively, obtain good prediction performance for predicting protein subcellular locations with both single and multiple sites. In this article, to achieve much better performance for predicting protein subchloroplast locations with both single and multiple sites, a novel multilabel algorithm is proposed, in which both label correlations and label-specific features are taken into account.

Suppose a training set T contains N proteins classified into M subchloroplast locations. According to the different subchloroplast locations, the N proteins can be further grouped into M subsets, i.e. $T=T_1\cup T_2\cup\cdots\cup T_i\cup\cdots\cup T_M$, where $T_i(i=1,2,\cdots,M)$ is the subset containing N_i proteins belonging to the same ith subchloroplast location. $N=578$ and $M=5$ in this study. It is noteworthy that $N\leq N_1+N_2+\cdots+N_M$ because proteins may belong to multiple subchloroplast locations as elaborated in (Chou *et al.*, 2011). Roughly speaking, our proposed multilabel algorithm is a two-level prediction algorithm, where each level contains M binary classifier, each corresponding to one of the M subchloroplast locations. It is important to note that these binary classifier can be constructed by any common machine learning algorithm and each binary classifier can choose different algorithm for training. In this article, for keeping our algorithm simple, we apply support vector machine (SVM) as the base learner for training all the binary classifier. The first level is denoted as L_1 and the second level L_2, i.e.

$$\begin{cases} L_1=\{SVM_1^1,SVM_1^2,\cdots,SVM_1^M\} \\ L_2=\{SVM_2^1,SVM_2^2,\cdots,SVM_2^M\} \end{cases} \quad (1)$$

where SVM_1^1 and SVM_2^1 are the prediction classifiers for the first subchloroplast location, SVM_1^2 and SVM_2^2 the second and so forth. Classifiers in the first level play an auxiliary role in the entire prediction process, while classifiers in the second level are responsible for making final prediction by using intermediate results provided by the first level.

For building these classifiers in the two levels, we should first construct their training sets. For the ith subchloroplast location, its training set Tr_i is composed of two subsets, as follows:

$$Tr_i=Tr_i^+\cup Tr_i^- \quad (2)$$

where Tr_i^+ is the set of proteins belonging to the ith subchloroplast location, and Tr_i^- is the set of proteins not belonging to the ith subchloroplast location. Tr_i^+ and Tr_i^- can be constructed by the following equations,

$$\begin{cases} Tr_i^+=\{(p,+1)|p\in T_i\} \\ Tr_i^-=\{(p,-1)|p\notin T_i\} \end{cases} \quad (3)$$

where $(p,+1)$ represents protein p belonging to the ith subchloroplast location, and $(p,-1)$ not belonging to the ith subchloroplast location.

According to the PseAAC feature extraction method, the kth protein p_k in the training set Tr_i can be denoted as

$$p_k = \left[p_k^1, p_k^2, \cdots, p_k^D\right]^T \tag{4}$$

where D is the dimensionality of PseAAC feature vector, and $D = 20 + 6 \times 50 = 320$ in this study. For classifiers in the first level, they can be constructed simply by training SVM on the corresponding training set. For the second level, the constructing method is relatively more complicated for classifiers in the second level. We first append all class labels except the current one to be predicted into the original feature space as additional features, and thus the augmented protein p_k can be denoted as

$$p_k = \left[p_k^1, p_k^2, \cdots, p_k^D, y_k^1, y_k^2, \cdots, y_k^{i-1}, y_k^{i+1}, \cdots, y_k^M\right]^T \tag{5}$$

where the former D features are the PseAAC features of the kth protein p_k, and the latter $M - 1$ features represent whether the kth protein p_k belongs to the other subchloroplast locations except the ith location or not, that is, if the protein p_k belongs to the j subchloroplast location $(j = 1, 2, \cdots, i-1, i+1, \cdots, M)$, then $y_k^j = 1$, otherwise $y_k^j = -1$. After that, we choose the most relevant PseAAC features and class labels from the augmented feature space for the ith subchloroplast location by using feature selection techniques. In this study, genetic algorithm is used for feature selection. Then the kth protein p_k can be redenoted as

$$p_k = \left[p_k^{\lambda_1}, p_k^{\lambda_2}, \cdots, p_k^{\lambda_d}, y_k^{\rho_1}, y_k^{\rho_2}, \cdots, y_k^{\rho_m}\right]^T \tag{6}$$

where $[\lambda_1, \lambda_2, \cdots, \lambda_d] \subseteq [1, 2, \cdots, D]$, $[\rho_1, \rho_2, \cdots, \rho_m] \subseteq [1, 2, \cdots, i-1, i+1, \cdots, M]$, the former d features are the optimal PseACC feature subset selected, and the latter m features the optimal subchloroplast location subset selected. Finally, classifiers in the second level can be constructed by training SVM based on the optimal feature subset and location subset. Because we not only utilize the most relevant features for each subchloroplast location, but also mix in correlations among different subchloroplast locations at the same time, our proposed algorithm ought to achieve better prediction performance.

Representation of individuals and evaluation of the fitness function are two important steps in applying the genetic algorithm for feature subset selection. Firstly, each individual in the population is represented by a n-dimensional binary vector f. In this study, $n = (320 + 4) \times 5$. Specifically, each n-dimensional binary vector f is divided into five groups, where each group corresponds to a different label (subchloroplast location in this study). Each group also consists of two parts, where the first part represents the PseAAC features ($320D$) and the other part represents all class label features except the corresponding one ($4D$). The corresponding feature p_i is excluded if $f_i = 0$, and is included otherwise. According to the feature selection information provided by each individual f, the selected features and labels for each label (subchloroplast location in this study) are retained. Secondly, the value of the fitness function for each individual f is computed as the overall accuracy [ACC as defined in Equation (10)] of the jackknife test.

Given a query protein p^*, its subchloroplast locations can be predicted according to the following steps:

1. Using PseACC model to extract the feature vector from the query protein p^*, just like in Equation (4).
2. Inputting the protein p^* into the classifiers in the first level, and then M prediction outputs for p^* will be obtained
$$\{y_1, y_2, \ldots, y_M\} \in \{+1, -1\} \tag{7}$$
where y_i represents whether p^* is predicted belonging to the ith subchloroplast location or not.
3. Using Equation (5) to augment p^*'s feature space. From Equation (5), we know subchloroplast locations for p^* are needed to append to p^*'s feature space as additional features. But we don't know which subchloroplast locations it belongs to. Actually, those are exactly what we want to predict. Therefore, we take the intermediate outputs generated in step 2 as its location estimation, and then augment p^*'s feature space by using its location estimation. This is why we construct classifiers in the first level.
4. According to Equation (6), selecting out the optimal feature subset of the query protein p^* for each subchloroplast location, respectively.
5. These optimal feature subsets from step 4 are put into classifiers in the second level accordingly, and then M final prediction outputs for p^* will be obtained
$$\{y_1^*, y_2^*, \ldots, y_M^*\} \in \{+1, -1\} \tag{8}$$

According to Equation (8), the predicted subchloroplast locations to which the query protein p^* belongs can be represented as

$$pred\_set(p^*) = \{j | y_j^* = +1, j = 1, 2, \ldots, M\} \tag{9}$$

2.4 Performance measures

Predicting subchloroplast localizations for chloroplast proteins with both single and multiple sites belongs to the case of multilabel classification. In machine learning community, it is well-known that performance evaluation of multilabel classification differs from that of traditional single-label classification because each example may have more than one labels simultaneously. Therefore, the following five novel measures are used to evaluate the performance of our proposed method from multiple aspects more exactly. These evaluation measures are defined as follows (Tsoumakas *et al.*, 2010):

$$\begin{cases} mlACC = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \\ mlPRE = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Z_i|} \\ mlREC = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Y_i|} \\ mlF1 = \dfrac{2 \cdot mlREC \cdot mlPRE}{mlREC + mlPRE} \\ ACC = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} 1(Y_i \equiv Z_i) \end{cases} \tag{10}$$

where Y_i is the set of true labels of each example, Z_i the set of predicted labels of each one, m the number of test examples, $|\cdot|$ the operator acting on the set to count the number of its elements. $1(Y_i \equiv Z_i)$ equals 1 if true labels are entirely identical to predicted labels, 0 otherwise. mlF1 is the harmonic mean of multilabel recall (mlREC) and multilabel precision (mlPRE). Note that for these above measures, the bigger the measure value, the better the

performance. As mentioned above, it is much more complicated to evaluate the performance of a classifier on a multilabel system. To understand the measures more easily, the readers should refer to the review article (Chou, 2013) where the meanings of these measures have been explained more intuitively.

3 Results and discussion

3.1 Evaluating our proposed algorithm on the benchmark dataset MSchlo578

Since the jackknife test is thought to be the most objective and rigorous method for cross validation, it has been increasingly used by investigators to examine the accuracy of various predictors (Guo *et al.*, 2014; Lin *et al.*, 2013b; Liu *et al.*, 2014; Wang and Li, 2012; Wang *et al.*, 2013; Wu *et al.*, 2011). Accordingly, the jackknife test was also adopted to examine the performance of the model proposed in this study. We applied the most commonly used SVM software libSVM (Chang and Lin, 2011) and optimized the parameters of each SVM classifier as defined in Equation (1) with a grid search strategy. The multilabel prediction outcomes obtained by our proposed prediction algorithm on the benchmark dataset MSchlo578 for the five performance measures as defined in Equation (10) are as follows, 63.26% (mlACC), 64.10% (mlPRE), 71.06% (mlREC), 67.38% (mlF1) and 55.52% (ACC). From its definition as defined in Equation (10), the overall accuracy (ACC) is the most strict one of the five measures without any overprediction or underprediction. So it is difficult to obtain very high overall accuracy (ACC). This indicates our proposed prediction algorithm achieves sufficiently high overall accuracy (ACC) and could become quite a promising multilabel predictor. To demonstrate the power of our algorithm, we carried out a fair comparison between our algorithm and the state-of-the-art AL-KNN (also called ML-KNN) algorithm commonly used in predicting multilabel protein subcellular localization (Chou *et al.*, 2011, 2012; Lin *et al.*, 2013b). We use the same features and the same jackknife test method to evaluate the performance of the AL-KNN classifier on the same benchmark dataset MSchlo578. The parameter K of the AL-KNN algorithm was optimized. The detailed results of the prediction performance are presented in Table 2. As can be seen from the table, our proposed algorithm significantly outperforms the state-of-the-art AL-KNN algorithm in terms of the five performance measures.

It is noteworthy that obtaining and comparing the accuracy of each subchloroplast location is meaningless in a multilabel prediction task just like the current study. Therefore, we give the overall accuracies (ACCs) for proteins with different number of labels (subchloroplast locations in this study) by our algorithm in Table 3, rather than the accuracy of each subchloroplast location. Moreover, for comparison, the overall accuracies (ACCs) by the AL-KNN algorithm are also shown in Table 3. As can be seen from the table, the overall accuracies (ACCs) by our algorithm are significantly higher than those by the AL-KNN algorithm. Particularly, for proteins with two subchloroplast locations, our algorithm performs better than the AL-KNN algorithm with a more than 42% big improvement. This indicates our algorithm is more capable of dealing with the multilabel prediction tasks than the state-of-the-art AL-KNN algorithm.

To further evaluate the prediction performance of our algorithm, we carried out an independent dataset test. In the independent dataset test, 20% of samples were randomly selected as the testing dataset. The remaining samples were used to train our algorithm and the AL-KNN algorithm. The prediction performance would be tested on the testing dataset. The procedure would be repeated 10 times to see whether the performance would vary significantly for different random selections. The averages and the standard deviations of the performances in ten repeats would be reported for the five measures. The performance details obtained by our algorithm and the AL-KNN algorithm can be found in Table 4. Obviously, our algorithm still performs significantly better than the AL-KNN algorithm. Moreover, the standard deviations of the performances are small, approximate 4%, for both our algorithm and the AL-KNN algorithm. This indicates both our algorithm and the AL-KNN algorithm are not over-estimated.

Table 2. Performance comparison of our proposed algorithm with AL-KNN on the benchmark dataset MSchlo578 by the jackknife test

Measure	Our algorithm (%)	AL-KNN (%)
mlACC	63.26	45.21
mlPRE	64.10	46.63
mlREC	71.06	45.30
mlF1	67.38	45.95
ACC	55.52	43.77

Table 3. A comparison of the overall accuracies (ACCs) by our proposed algorithm and AL-KNN for proteins with different number of subchloroplast locations

		The overall accuracy (ACC)	
Number of locations	Number of proteins	Our algorithm (%)	AL-KNN (%)
1	556	56.12	45.50
2	21	42.86	0
3	1	0	0

3.2 Comparison with the existing single-label predictors

As mentioned in the Introduction section, all the existing methods can only be used to identify the single subchloroplast location of a query protein, none of the existing methods can be used to deal with proteins with multiple subchloroplast locations. Nevertheless, it is still interesting to see if our algorithm could work better than the existing methods for the same single-label problem that the existing methods have tackled. We compare our proposed algorithm with several recent state-of-the-art predictors, SubChlo (Du *et al.*, 2009), ChloroRF (Tung *et al.*, 2010), SubIdent (Shi *et al.*, 2011) and BS-KNN (Hu and Yan, 2012). All of them are the sequence-based predictors. SubChlo was based on the ET-KNN (evidence theoretic K-nearest neighbor) algorithm and PseAAC, ChloroRF was based on the Random Forest algorithm and the physicochemical properties of protein amino acid sequences, SubIdent was based on the Support Vector Machine and the discrete wavelet transform feature extraction method, and BS-KNN was based on the bit-score weighted K-nearest neighbor (BS-KNN) algorithm and the selected PseAAC. To obtain a comparable result with these predictors, we also evaluate the prediction performance of our algorithm on the same dataset S60 as used in these predictors by using the jackknife test method. For the dataset S60, it contains 262 proteins classified into four different subchloroplast locations, in which 40 proteins belong to the 'Envelope' location site, 49 to the 'Stroma', 44 to the 'Thylakoid lumen' and 129 to the 'Thylakoid membrane'. The performances of

Table 4. Performance comparison of our proposed algorithm with AL-KNN on the benchmark dataset MSchlo578 by the independent dataset test

Measure	Our algorithm (%)	AL-KNN (%)
mlACC	63.36 (4.02)[a]	45.20 (3.60)
mlPRE	64.57 (4.00)	46.52 (3.66)
mlREC	65.55 (3.74)	45.61 (3.57)
mlF1	65.05 (3.86)	46.06 (3.60)
ACC	60.00 (4.48)	43.48 (3.71)

[a]Enclosed in parentheses are the standard deviations of the performances for the five measures.

Table 5. Single-label prediction performance comparison of our proposed algorithm with the existing single-label subchloroplast localization predictors on the S60 dataset by the jackknife test

Location	Accuracy (%)				
	Our algorithm	SubChlo	ChloroRF	SubIdent	BS-KNN
Envelope	72.5	40.0	47.5	80.0	47.5
Stroma	95.9	67.4	57.1	85.7	73.9
Thylakoid lumen	61.4	43.2	38.6	64.4	77.5
Thylakoid membrane	100	83.7	87.5	98.2	85.0
Overall	88.6	67.2	67.4	89.3	75.9

these compared methods are presented in Table 5. As we can see from the table, the overall accuracy of our algorithm is over 88%, which is significantly better than all the compared predictors except SubIdent. Compared with SubIdent, our algorithm obtains a comparable overall accuracy (88.6% versus 89.3%). Particularly, the accuracy of our algorithm is superior to SubIdent in the 'Stroma' location site, and inferior to SubIdent in the 'Envelope' location site, and the similar results are obtained by our algorithm and SubIdent in the other two location sites. This observation indicates that although it is specially designed for predicting multilabel protein subchloroplast locations, our algorithm also performs well in predicting single-label protein subchloroplast locations.

4 Web server

Based on the above prediction algorithm, we developed an online web server for providing multilabel protein subchloroplast localization prediction service, called MultiP-SChlo freely accessible at http://biomed.zzuli.edu.cn/bioinfo/multip-schlo/, by which biologists can easily obtain their desired results with no need for professional math and computer knowledge.

5 Conclusion

In this article, we further study sub-structures of the chloroplast organelle, and try to identify subchloroplast locations of proteins with both single and multiple sites. All the existing predictors have the following drawbacks: (i) only considering the small number of subchloroplast locations, (ii) homologous bias of the benchmark dataset adopted is too large, (iii) more importantly, ignoring proteins with multiple subchloroplast locations when constructing prediction models. In view of this, we propose a novel multilabel prediction algorithm for identifying multilabel protein subchloroplast locations. The main contributions of this article is as follows:

1. A new benchmark dataset is constructed that contains chloroplast proteins with both single and multiple subchloroplast locations. It not only covers much more subchloroplast locations but also has much less homology bias.
2. A novel multilabel algorithm is proposed through combining label-specific features with label correlations, which can predict multilabel protein subchloroplast locations accurately.
3. An online web server for our proposed algorithm is established called MultiP-SChlo freely accessible at http://biomed.zzuli.edu.cn/bioinfo/multip-schlo/.

Acknowledgement

We thank the anonymous reviewers for suggestions and comments, which helped us improving the quality of this article.

Funding

This work was supported by the National Natural Science Foundation of China (61402422, 61403349, 61302118 and 61273305), Key Project of Science and Technology Research of the Education Department of Henan Province (14A520063), Doctoral Research Fund of Zhengzhou University of Light Industry (2013BSJJ082) and Open Fund of MOE Key Laboratory of Embedded System and Service Computing of Tongji University (ESSCKF201308).

Conflict of Interest: none declared.

References

Cao,D.-S. *et al.* (2013). propy: a tool to generate various modes of chou's pseaac. *Bioinformatics,* **29**, 960–962.

Chang,C.-C. and Lin,C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.,* **2**, **27**, 1–27.

Chou,K.-C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins,* **43**, 246–255.

Chou,K.-C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics,* **21**, 10–19.

Chou,K.-C. (2013). Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.,* **9**, 1092–1100.

Chou,K.-C. and Shen,H.-B. (2010). Plant-mPLoc: a Top-Down strategy to augment the power for predicting plant protein subcellular localization. *PLoS One,* **5**, e11335.

Chou,K.-C. *et al.* (2011). iLoc-Euk: a Multi-Label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One,* **6**, e18258.

Chou,K.-C. *et al.* (2012). iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.,* **8**, 629.

Du,P. and Li,Y. (2006). Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinform.,* **7**, 518.

Du,P. *et al.* (2009). SubChlo: Predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic k-nearest neighbor (ET-KNN) algorithm. *J. Theor. Biol.,* **261**, 330–335.

Du,P. *et al.* (2012). Pseaac-builder: A cross-platform stand-alone program for generating various special chou's pseudo-amino acid compositions. *Anal. Biochem.,* **425**, 117–119.

Du,P. *et al.* (2014). Pseaac-general: Fast building various modes of general form of chou's pseudo-amino acid composition for large-scale protein datasets. *Int. J. Mol. Sci.,* **15**, 3495–3506.

Fang,Y. *et al.* (2008). Predicting dna-binding proteins: approached from chou's pseudo amino acid composition and other specific sequence features. *Amino Acids,* **34**, 103–109.

Fu,L. *et al.* (2012). CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics,* **28**, 3150–3152.

Guo,S.-H. *et al.* (2014). inuc-pseknc: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics,* **30**, 1522–1529.

Hu,J. and Yan,X. (2012). BS-KNN: an effective algorithm for predicting protein subchloroplast localization. *Evol. Bioinform.,* **8**, 79–87.

Jahandideh,S. *et al.* (2007). Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. *Biophys. Chem.,* **128**, 87–93.

Kumar,R. *et al.* (2014). Protein sub-nuclear localization prediction using svm and pfam domain information. *PLoS One,* **9**, e98345.

Li,G.-Z. *et al.* (2012). Multilabel learning for protein subcellular location prediction. *IEEE Trans. NanoBiosci.,* **11**, 237–243.

Lin,H. *et al.* (2013a). Using over-represented tetrapeptides to predict protein submitochondria locations. *Acta Biotheoretica,* **61**, 259–268.

Lin,W.-Z. *et al.* (2013b). iloc-animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. Biosyst.,* **9**, 634–644.

Liu,B. *et al.* (2014). Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics,* **30**, 472–479.

Mohabatkar,H. *et al.* (2013). Prediction of allergenic proteins by means of the concept of chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.,* **9**, 133–137.

Nakashima,H. and Nishikawa,K. (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.,* **238**, 54–61.

Nakashima,H. *et al.* (1986). The folding type of a protein is relevant to the amino acid composition. *J. Biochem.,* **99**, 153–162.

Qiu,J.-D. *et al.* (2010). Using the concept of chou's pseudo amino acid composition to predict enzyme family classes: An approach with support vector machine based on discrete wavelet transform. *Protein Pept. Lett.,* **17**, 715–722.

Reinhardt,A. and Hubbard,T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.,* **26**, 2230–2236.

Saravanan,V. and Lakshmi,P. (2013). Sclap: An adaptive boosting method for predicting subchloroplast localization of plant proteins. *OMICS,* **17**, 106–115.

Shen,H.-B. and Chou,K.-C. (2005). Predicting protein subnuclear location with optimized evidence-theoretic k-nearest classifier and pseudo amino acid composition. *Biochem. Biophys. Res. Commun.,* **337**, 752–756.

Shen,H.-B. and Chou,K.-C. (2008). Pseaac: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.,* **373**, 386–388.

Shi,S.-P. *et al.* (2011). Identify submitochondria and subchloroplast locations with pseudo amino acid composition: Approach from the strategy of discrete wavelet transform feature extraction. *Biochimica et Biophysica Acta,* **1813**, 424–430.

Tsoumakas,G. *et al.* (2010). Mining multi-label data. In: *Data Mining and Knowledge Discovery Handbook*. Springer, Boston, MA, pp. 667–685.

Tung,C.-W. *et al.* (2010). Prediction of protein subchloroplast locations using random forests. In: *Proceeding of World Academy of Science, Engineering and Technology*. Tokyo, Japan, pp. 699–703.

Wang,X. and Li,G.-Z. (2012). A Multi-Label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins. *PLoS One,* **7**, e36317.

Wang,X. and Li,G.-Z. (2013). Multilabel learning via random label selection for protein subcellular multilocations prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.,* **10**, 436–446.

Wang,X. *et al.* (2013). Virus-ECC-mPLoc: a Multi-Label predictor for predicting the subcellular localization of virus proteins with both single and multiple sites based on a general form of chou's pseudo amino acid composition. *Protein Pept. Lett.,* **20**, 309–317.

Wu,Z.-C. *et al.* (2011). iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. Biosyst.,* **7**, 3287.

Zhou,G.-P. and Doctor,K. (2003). Subcellular location prediction of apoptosis proteins. *Proteins,* **50**, 44–48.

Zhou,X.-B. *et al.* (2007). Using chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.,* **248**, 546–551.

Zou,D. *et al.* (2011). Supersecondary structure prediction using chou's pseudo amino acid composition. *J. Comput. Chem.,* **32**, 271–278.