Evolutionary Bioinformatics

METHODOLOGY

BS-KNN: An Effective Algorithm for Predicting Protein Subchloroplast Localization

Jing Hu[1] and Xianghe Yan[2]

[1]Department of Mathematics and Computer Science, Franklin & Marshall College, P.O. Box 3003, Lancaster, PA 17604, USA. [2]Eastern Regional Research Center, Agricultural Research Service, US Department of Agriculture, 600 E. Mermaid Lane, Wyndmoor, PA 19038, USA. Corresponding author email: jing.hu@fandm.edu

Abstract: Chloroplasts are organelles found in cells of green plants and eukaryotic algae that conduct photosynthesis. Knowing a protein's subchloroplast location provides in-depth insights about the protein's function and the microenvironment where it interacts with other molecules. In this paper, we present BS-KNN, a bit-score weighted K-nearest neighbor method for predicting proteins' subchloroplast locations. The method makes predictions based on the bit-score weighted Euclidean distance calculated from the composition of selected pseudo-amino acids. Our method achieved 76.4% overall accuracy in assigning proteins to 4 subchloroplast locations in cross-validation. When tested on an independent set that was not seen by the method during the training and feature selection, the method achieved a consistent overall accuracy of 76.0%. The method was also applied to predict subchloroplast locations of proteins in the chloroplast proteome and validated against proteins in Arabidopsis thaliana. The software and datasets of the proposed method are available at https://edisk.fandm.edu/jing.hu/bsknn/bsknn.html.

Keywords: subchloroplast localization, bit-score weighted K-nearest neighbor method, pseudo-amino acids, feature selection

Evolutionary Bioinformatics 2012:8 79–87

doi: 10.4137/EBO.S8681

This article is available from http://www.la-press.com.

Introduction

Chloroplasts are organelles found in cells of green plants and eukaryotic algae. They are believed to have originated from cyanobacteria through endosymbiosis. Chloroplasts play important functional roles in many biological processes such as photosynthesis and cellular metabolism. Similar to a cell that can be divided into several subcellular locations, the chloroplast is also subdivided into multiple subchloroplast locations. Knowing a protein's subchloroplast location information provides in-depth biological insights about the protein's roles in these biological processes.

Recent developments in high-throughput genome sequencing projects have resulted in an increasing number of raw chloroplast protein sequences stored in public databases. For the majority of these proteins, little knowledge is known about their subchloroplast locations. Therefore, computational methods that can predict subchloroplast localizations of chloroplast proteins are needed. However, despite the chloroplast proteome projects[1–4] and various computational approaches[5–8] to identify chloroplast proteins in proteomic scale, there are only a limited number of methods as to our knowledge for predicting protein subchloroplast locations. SubChlo[9] is the first method for predicting the subchloroplast locations of chloroplast proteins. The method is based on the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm.[10,11] Using pseudo-amino acid composition[12] as the feature set, the method achieved 67.2% overall accuracy in predicting proteins' subchloroplast locations on a dataset consisting of chloroplast proteins with less than 60% sequence similarities. ChloroRF[13] predicts subchloroplast locations using a feature vector of 531 physicochemical properties obtained from AAindex database.[14] Applying the Random Forest algorithm, ChloroRF achieved a comparable accuracy, 67.4%, as that of SubChlo on the same dataset. An extra benefit of ChloroRF is that it utilizes human-interpretable physicochemical properties, which can provide meaningful information for analyzing the mechanisms of protein subchloroplast localizations. Recently, a method called Subldent[15] was developed to identify submitochondria and subchloroplast locations. The method first created numerical series of hydrophobicity and polarity values from protein's amino acid sequence, and then applied a discrete wavelet transform to formulate them into a different representation of pseudo-amino acid composition. These features were then used to train a support vector machine classification model to predict protein's subchloroplast locations. For a complete list of subchloroplast localization methods and their details, please see Supplementary Table 1.

In this paper, we present BS-KNN, a bit-score weighted K-Nearest neighbor method for predicting proteins' subchloroplast locations. The method makes predictions based on a bit-score weighted Euclidean distance (BS-WED) computed from residue composition. For each subchloroplast location, it finds its K nearest neighbors (ie, proteins) to the query protein based on the BS-WED. Then the average BS-WED of the query protein to the K proteins is used as the distance of the query protein to this location. The query protein is then predicted to be in a subchloroplast location to which its distance is the smallest. The method achieved 68.0% overall accuracy in assigning proteins to four subchloroplast locations in cross-validation using amino acid composition. The method was then further improved by applying a heuristic feature selection process to choose pseudo-amino acid composition. The final method achieved 76.4% overall prediction accuracy in cross-validation. When tested on an independent set that was not seen by the method during the training and feature selection, the method still achieved 76.0% overall accuracy.

Table 1. The performance of BS-KNN based on selected pseudo-amino acid composition.

Subchloroplast location	4-fold cross-validation (S60_A)		Independent test (S60_B)	
	Recall	Precision	Recall	Precision
Thylakoid lumen	75.0%	77.4%	50.0%	66.7%
Stroma	64.9%	63.2%	88.9%	57.1%
Thylakoid membrane	85.3%	84.5%	84.0%	91.3%
Envelope	62.5%	64.5%	62.5%	71.4%
Overall accuracy	**76.4%**		**76.0%**	

Materials and Methods

Dataset

We used the benchmark dataset that was used in the study of SubChlo,[9] ChloroRF[13] and ChloroRF.[15] The dataset can be downloaded from the website of SubChlo (http://bioinfo.au.tsinghua.edu.cn/subchlo/). The original dataset contained 736 proteins extracted from Swiss-Prot database release 56.2.[16] After discarding three proteins with incomplete Swiss-Prot IDs and removal of redundant proteins by CD-HIT[17] using a sequence similarity threshold of 60%, there were 253 proteins left. The final dataset (S60) contained 40 envelope proteins, 46 stroma proteins, 127 thylakoid membrane proteins and 40 thylakoid lumen proteins. The dataset used in this study was slightly different from the reported S60 datasets of SubChlo (262 proteins) and ChloroRF (261 proteins). This could be due to different versions of the CD-HIT program.

In most studies, a much stricter sequence similarity threshold (25% or 30%) was used to remove the redundancy among protein sequences. However, using such thresholds for this small dataset would greatly reduce the number of proteins in the final dataset. This would lead to insufficient training, and the performance evaluation would have little significance. Therefore we accepted a relaxed threshold (60%), which was also used in previous studies (ie, SubChlo, ChloroRF and Subldent). To show that our method does not suffer from generalization problems due to comparatively high sequence similarity among proteins in the dataset, we also compared our method with prediction based solely on similarity search.

Feature set

There were two feature sets investigated in this study, which were amino acid composition and pseudo-amino acid composition of protein sequences.

Amino acid Composition: The amino acid composition of a protein was calculated as:

$$x_i = n_i / \sum_j n_j \qquad (1)$$

where n_i and n_j are the numbers of amino acid i and j in the protein sequence.

Pseudo-Amino Acid Composition: The model of pseudo-amino acid composition was first proposed by Chou[12] to predict protein cellular attributes. Unlike the classical amino acid composition that consists of only 20 discrete numbers, the pseudo amino acid composition consists of 20+λ discrete numbers, among which the first 20 numbers represent the occurrence frequencies of 20 amino acids in a protein, and the remainders represent different ranks of sequence-order correlation factors. The model was then extended to include two sets of sequence-order correlation factors: delta-function set (λ discrete numbers) and hydrophobicity set (μ discrete numbers). The new model has been successfully applied to predict proteins' subcellular locations by Chou and Cai.[18]

In this study, we investigated two sets of sequence-order correlation factors.[18] Suppose there is a protein X with a sequence of L amino acid residues: $R_1R_2R_3R_4 \ldots R_L$, where R_1 represents the residue at sequence position 1, R_2 the residue at position 2, and so on. The first set, delta-function set, consists of λ sequence-order-correlated factors, which are given by:

$$\delta_i = \frac{1}{L-i}\sum_{j=1}^{L-i}\Delta_{j,j+i} \qquad (2)$$

where $i = 1, 2, 3 \ldots \lambda$, $\lambda < L$, and $\Delta_{j,j+i} = \Delta(R_j, R_{j+1})$ if $R_j = R_{j+i}$, 0 otherwise. The second set is physicochemical set. It consists of μ sequence-order-correlated factors, which are given by:

$$h_i = \frac{1}{L-i}\sum_{j=1}^{L-i}H_{j,j+i} \qquad (3)$$

where $i = 1, 2, 3 \ldots \mu$, $\mu < L$, and $H_{i,j} = H(R_i) \cdot H(R_j)$. In this study, $H(R_i)$ and $H(R_j)$ are the normalized hydrophilicity values of R_i and R_j respectively. Note that the original hydrophilicity values are normalized by formula (4) before applying formula (3):

$$H_i = (H_i^0 - \overline{H^0}) \Big/ \sqrt{\left\{\sum_{j=1}^{20}(H_j^0 - \overline{H^0})^2\right\}/20} \qquad (4)$$

where $i = 1, 2, 3, \ldots, 20$. H_1^0 is the original hydrophilicity value of amino acid i, and $\overline{H^0}$ is the average hydrophilicity value of 20 amino acids. Therefore, the pseudo-amino acid

composition consists of 20 (classic amino acid composition)+λ(delta-functionset)+μ(hydrophilicity set) numbers. In this study, λ and μ were both set to be 20 since most polypeptide sequence lengths are greater than 20.

Bit-Score Weighted Euclidean Distance

For each query protein *t*, its distance to a training protein *T* is calculated as:

$$D = \sqrt{\sum_i (t_i - T_i)^2} \Big/ BS(t,T) \quad (5)$$

where t_i is the *i*th composition of the query protein *t*, and T_i is the *i*th composition of the training protein *T*. *BS(t,T)* is the bit score computed by *blastp* program of Blast package[19] for comparing the sequence similarity between protein sequence *t* and *T*. A higher bit score indicates that two protein sequences are more similar, and vice versa. Therefore, the more similar the two protein sequences are, the smaller the numerator will be, the bigger the denominator will be, and the smaller the distance will be. Notice that $\sqrt{\sum_i (t_i - T_i)^2}$ gives the Euclidean distance between the query protein and the training protein. Here, the distance is weighted by a factor of 1/*BS(t,T)*. Therefore, the distance is referred to as bit-score weighted Euclidean distance (BS-WED).

Bit-Score Weighted K-Nearest Neighbor Method

For each test protein, its BS-WED to every training protein in the training set was calculated. Then for each subchloroplast location, K shortest distances were chosen. For example, let $D_{env\text{-}1}$, $D_{env\text{-}2}$, …, $D_{env\text{-}k}$ be the K shortest distances between the test protein and proteins that locate at the envelope. Then, the distance between the test protein and the location of envelope was given by:

$$\overline{D_{env}} = \sum_{i=1}^{K} D_{env\text{-}i} \Big/ K \quad (6)$$

The distances between the test protein and every subchloroplast location were calculated separately. Then the test protein was assigned to a location to which its distance is the shortest.

Cross-validation, Independent Test, Self-consistency Test, and jackknife Test

The dataset was randomly split into 5 subsets. Four subsets (referred to as S60_A) were used to perform four-fold cross-validation and feature selection. In each round of experiments, three subsets were used as a training set, and the remaining subset was used as a test set. This procedure was repeated four times with each subset being used as a test set once. The overall performance was calculated. The fifth subset (independent set, and referred to as S60_B) served as the test set in the independent test stage, in which the classifier was trained using the four subsets (S60_A) and then tested on the independent set (S60_B). Note that the algorithm did not see the independent set during the feature selection stage and the training of the classifier.

Self-consistency test and jackknife test have been used by previous studies[9,13,18,20] to evaluate the multiclass classification performance of protein localizations. In this study, we also evaluated the final BS-KNN method on S60 dataset using self-consistency test and jackknife test. In self-consistency test, proteins in the dataset were predicted using the classification model trained on the same dataset. Therefore, self-consistency test gives the most optimal estimation of the classification performance. In jackknife test, each protein in the dataset was used as the test protein once, and the remaining proteins were used to train the classifier. Therefore, jackknife test provides a more reliable estimation of the classification performance, especially when the dataset is small.

Performance Measurement

Performances were measured using accuracy/recall (RC) and precision (PR) for each subchloroplast location *i*:

$$RC_i = TP_i / N_i \quad (7)$$

$$PR_i = TP_i / (TP_i + FP_i) \quad (8)$$

where TP_i, TN_i, FP_i, FN_i, and N_i were the numbers of true positives, true negatives, false positives, false negatives, and total number of proteins for location *i*.

Besides these, the overall accuracy (OA) of the four subchloroplast locations was also used:

$$OA = \sum_{i=1}^{4} TP_i \Big/ \sum_{i=1}^{4} N_i \quad (9)$$

Heuristic Feature Selection

We further extended the BS-KNN method by using pseudo-amino acid composition to calculate the BS-WED. However, not all the features were useful for the prediction of subchloroplast localizations. Also, some features might be correlated with each other, which could impair the prediction performance. We used a greedy feature selection method to select the most relevant features. The greedy search started with a feature set that included the composition of 20 amino acids. Let n be the size of the feature set. Then $n = 20$ at the beginning. The algorithm was divided into three stages: reduction, growth_I and growth_II. In the reduction stage, the size of the feature set was gradually reduced. First, one amino acid was removed, and the composition of the remaining n–1 amino acids was used to calculate the BS-WED. Four-fold cross-validation was used to evaluate the performance of the method by optimizing the overall accuracy. This step was repeated n times, so that every combination of n-1 amino acids was tried. The combination that improved the performance most was chosen. Thus, the size of the feature set was reduced from n to n–1. This reduction process was continued until removing any amino acid from the feature set would reduce the performance. At the end of the reduction stage, we reached a feature set that included the composition of N amino acids ($N \leq 20$). Next, we entered the growth_I stage to increase the size of the feature set by adding the delta-function factors. One delta-function factor was temporarily added into the feature set, and the resulting feature set was used to calculate BS-WED. Four-fold cross-validation was used to evaluate the performance of the method. This step was repeated λ (ie, 20) times, so that every delta-function factor was tested. The delta-function factor that yielded the greatest improvement in performance was chosen and added to the feature set. Thus, the size of feature set was increased to N+1. The growth_I stage for including delta-function factor was continued until adding more delta-function factor would decrease the performance. Then we entered the growth_II stage for including hydrophilicity factors. The growth_II stage was processed similarly to the growth_I stage.

Results

Prediction performance using only amino acid composition

The BS-WED was developed as the distance measurement in the proposed BS-KNN algorithm. Only the composition of 20 amino acids was used to calculate the distance between the test protein and training proteins. Four-fold cross-validation was used to evaluate the performance. Various K values ranging from 1 to 15 were tried. As can be seen from Figure 1, the best performance was achieved when K = 2. The proposed BS-WED achieved 68.0% overall accuracy in assigning proteins' subchloroplast locations. Notice that this had already outperformed the prediction performance of SubChlo (67.2%) and ChloroRF (67.4%). For comparison purposes, we also showed the prediction performances of the K-NN method using standard Euclidean distance. As can be seen from Figure 1, the BS-WED developed in this study is a better distance measurement than standard Euclidean distance in predicting subchloroplast locations for K values from 1 to 15.

Prediction performance using selected pseudo-amino acid composition

In the above experiment, only the composition of amino acids was used to calculate the BS-WED.

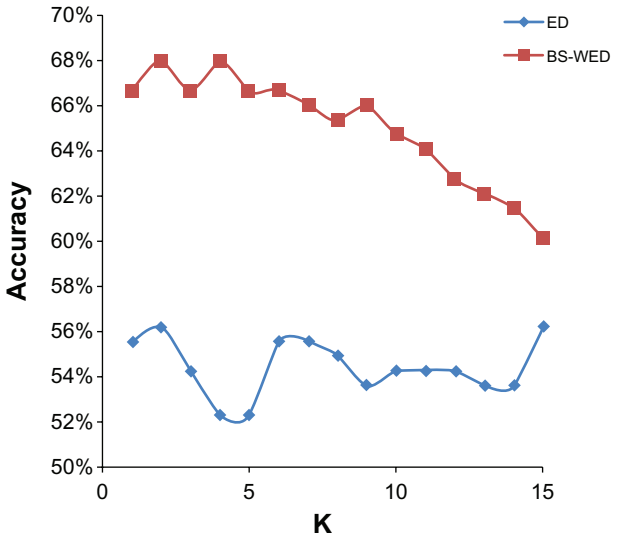

Figure 1. Prediction accuracies of K-NN for various K values (1–15) based on Euclidean distance (ED) vs. bit-score weighted Euclidean distance (BS-WED).

We tried to improve the prediction performance by using the composition of pseudo-amino acids. We applied the heuristic feature selection process as described in Materials and Methods to search for a combination of features that could be used to predict subchloroplast localizations. In the end, a set of features that includes the composition of 9 amino acids {A, D, Q, E, I, S, T, W, Y} and 2 delta-function factors $\{\delta_9, \delta_{20}\}$ were selected. As can been seen from Table 1, using selected features, the final method reached an overall accuracy of 76.4% by four-fold cross-validation. To make sure the selected features do not over-fit the dataset, we also evaluated the final method using an independent test, in which the classifier was trained using the four subsets (S60_A) and tested on an independent set (S60_B). Note that the independent set was not seen by the prediction method during the feature-selection stage and the training of the classifier. The results (Table 1) show that the method achieved 76.0% overall accuracy, which indicates that the proposed method does not suffer from over-fitting problem.

The method was also evaluated by self-consistency test and jackknife test. As can be seen from Table 2, the final BS-KNN achieved 98.8% overall accuracy by self-consistency test and 75.9% overall accuracy by jackknife test. The overall accuracy of 98.8% is the most optimal accuracy our method can achieve. Comparison of the performance of jackknife test with those of cross-validation and independent test showed that our method achieved consistent and robust performance.

Comparison with prediction solely based on similarity search

Similarity searches have been widely used to infer protein functions. If two proteins are highly similar in sequence, then they might share similar functions, structures, and evolutionary origin. For each test protein, we conducted a homologous search on the training set using the BLAST program.[19] The test protein was predicted to be at the same location (ie, thylakoid lumen, stroma, thylakoid membrane, or envelope) as that of the most homologous protein. Using the same dataset partition, the similarity search only achieved 65.0% overall accuracy when evaluated by four-fold cross-valuation, which was much lower than the proposed BS-KNN method.

Comparison with previously published methods

The proposed BS-KNN method was compared with previously published methods (ie, SubChlo, ChloroRF, Subldent) on the same dataset (S60). The prediction results of SubChlo, ChloroRF and Subldent were directly obtained from their reports. In their studies, jackknife test was used to evaluate the prediction performance. To make direct comparison, we also evaluated BS-KNN on the same dataset using jackknife test. Table 3 shows that the proposed BS-KNN method outperformed SubChlo and ChloroRF by approximately 8% of overall accuracy in classifying proteins into four subchloroplast locations. The overall prediction accuracy of Subldent is higher than ours. However, the selection of most optimal parameters (ie, wavelet functions, decomposition scales and SVM parameters) and the evaluation of Subldent on the same dataset may introduce overfitting problem, and therefore overestimated performance evaluation. On the contrary, we searched for the best feature subset on one dataset (S60_A) and evaluated BS-KNN on another dataset (S60_B), and observed consistent performance. It is also worth mentioning that the performance of our method and

Table 2. The performance of BS-KNN using selected pseudo-amino acid composition on S60 dataset by self-consistency test and jackknife test.

Subchloroplast location	Self-consistency test		Jackknife test	
	Recall	Precision	Recall	Precision
Thylakoid lumen	97.5%	97.5%	77.5%	79.5%
Stroma	100%	100%	73.9%	61.8%
Thylakoid membrane	99.2%	98.4%	85.0%	83.7%
Envelope	97.5%	100%	47.5%	63.3%
Overall accuracy	**98.8%**		**75.9%**	

Table 3. Comparison of BS-KNN with previously published subchloroplast localization methods on S60 dataset by jackknife test.

Location	Accuracy			
	SubChlo[9]	ChloroRF[13]	SubIdent[15]	BS-KNN
Thylakoid lumen	43.2%	38.6%	64.4%	**77.5%**
Stroma	67.4%	57.1%	85.7%	**73.9%**
Thylakoid membrane	83.7%	87.5%	98.2%	**85.0%**
Envelope	40.0%	47.5%	80.0%	**47.5%**
Overall accuracy	67.2%	67.4%	89.3%	**75.9%**

that of SubIdent are complementary to each other (ie, the performance of BS-KNN is higher than that of SubIdent in lumen location, and lower in other 3 locations). This suggests that a possible ensemble method with higher prediction capacity could be developed in the future by integrating BS-KNN with SubIdent, and possibly the other two methods (ie, SubChlo and ChloroRF).

Proteome scan

The proposed BS-KNN method was also applied to scan the chloroplast proteomes downloaded from plprot.[21] The plprot contains 690 chloroplast proteins of Arabidopsis thaliana as of March 2011, among which 258 (37.4%) proteins were predicted to be in the stroma, 139 (20.1%) proteins were predicted to be in the envelope, 99 (14.3%) proteins were predicted to be in the thylakoid lumen, and 194 (28.1%) proteins were predicted to be in the thylakoid membrane. Though some of these predictions would need to be validated, they could provide suggestive information to the future chloroplast proteome projects.

In this paper, we also validate our method against proteins in *Arabidopsis thaliana* downloaded from PPDB (http://ppdb.tc.cornell.edu/default.aspx), a Plant Proteome DataBase for *Arabidopsis thaliana* and maize (Zea mays). The extracted proteins with annotation of subchloroplast locations of envelope, stroma, thylakoid lumen side (lumen), and thylakoid-integral (thylakoid membrane) used in this study were based on their subcellular proteomes data set. There were 958 proteins in total, of which 345 (36.0%) proteins were predicted to be in the stroma, 208 (21.7%) proteins were predicted to be in the envelope, 127 (13.2%) were predicted to be in the thylakoid lumen, and 278 (29.0%) proteins predicted to be in the thylakoid membrane. In total, 52.9% of these proteins (experimental verified or predicted by previous studies) have been correctly predicted. For comparison, we also applied SubChlo to predict proteins in *Arabidopsis thaliana*, and it achieved 49.3% accuracy.

The proteins in plprot and *Arabidopsis thaliana* and their prediction using BS-KNN can be found at https://edisk.fandm.edu/jing.hu/bsknn/genomeScan/.

Discussion

In this paper, we present a BS-KNN algorithm for predicting protein subchloroplast locations. For each query protein, BS-WED was used as the distance measurement to find its K nearest neighbors from each location. To further improve the method's prediction performance, we investigated the pseudo-amino acid composition. Compared with the classical amino acid composition, the pseudo-amino acid composition provided more sequential and physicochemical information at different orders. By applying a heuristic feature selection process, the final method achieved 76.4% overall accuracy in classifying proteins into 4 subchloroplast locations using selected pseudo-amino acid composition by four-fold cross-validation. When evaluated on an independent test dataset, the method achieved a consistent accuracy of 76.0%. The method also achieved 75.9% overall accuracy by jackknife test. This shows that our method does not suffer from generalization problem and it has consistent prediction performance. We also applied our method to annotate proteins in the chloroplast proteome and validated the method against proteins in *Arabidopsis thaliana*.

The proposed BS-KNN method used a bit-score weighted Euclidean distance (ie, $\sqrt{\sum_i (t_i - T_i)^2} / BS(t,T)$ as the distance measurement. Compared with the standard Euclidean distance, the bit-scored weighted

Euclidean distance (BS-WED) is a better measurement to evaluate the distance between proteins. BS-WED accounts both sequence similarity and residue compositions. For example, assume we have two pairs of imaginary protein sequences, eg, pair 1 (APAPAPAP vs. AAAAPPPP) and pair 2 (APAPAPAP vs. PAPAPAPA). Two pairs of protein sequences have the same Euclidean distances because they have the same composition. However, pair 2 has a much higher bit score than pair 1 does (protein sequences in pair 2 have higher sequence similarity, therefore a higher bit score.), therefore when weighted by bit scores, the BS-WED of pair 2 is smaller than that of pair 1, which follows the biological sense.

In conclusion, the proposed bit-score weighted K-nearest neighbor algorithm is an effective method for predicting the subchloroplat location of proteins.

Acknowledgments

The project is partially supported by the grant from Howard Hughes Medical Institute awarded to Franklin & Marshall College.

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Ferro M, et al. Integral membrane proteins of the chloroplast envelope: identification and subcellular localization of new transporters. *Proc Natl Acad Sci U S A*. 2002;99:11487–92.
2. Ferro M, et al. Proteomics of the chloroplast envelope membranes from Arabidopsis thaliana. *Mol Cell Proteomics*. 2003;2:325–45.
3. Peltier JB, et al. Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell*. 2000;12:319–41.
4. Peltier JB, et al. Central functions of the lumenal and peripheral thylakoid proteome of Arabidopsis determined by experimentation and genome-wide prediction. *Plant Cell*. 2002;14:211–36.
5. Abdallah F, Salamini F, Leister D. A prediction of the size and evolutionary origin of the proteome of chloroplasts of Arabidopsis. *Trends Plant Sci*. 2000;5:141–2.
6. Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci*. 1999;8:978–84.
7. Leister D. Chloroplast research in the genomic age. *Trends Genet*. 2003;19:47–56.
8. Martin W, et al. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci U S A*. 2002;99:12246–51.
9. Du P, Cao S, Li Y. SubChlo: predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm. *J Theor Biol*. 2009;261:330–5.
10. Guo X, Gao X. A novel hierarchical ensemble classifier for protein fold recognition. *Protein Eng Des Sel*. 2008;21:659–64.
11. Nanni L, Lumini A. Particle swarm optimization for ensembling generation for evidential k-nearest-neighbour classifier. *Neural Comput Appli*. 2009;18:105–8.
12. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*. 2001;43:246–55.
13. Tung CW, Liaw C, Ho SJ, Ho SY. Prediction of protein subchloroplast locations using Random Forests. In: *Proceeding of World Academy of Science, Engineering and Technology*, Tokyo, Japan, May 26–28, 2010, WASET'10, WASET, 903–7.
14. Kawashima S, et al. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*. 2008; Database issue: D202–5.
15. Shi SP, et al. Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction. *Biochim Biophy Acta*. 2011;1813:424–30.
16. UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*. 2009; Database issue: D169–74.
17. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*. 2001;17:282–3.
18. Chou KC, Cai YD. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J Cell Biochem*. 2003;90:1250–60.
19. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–402.
20. Kandaswamy KK, et al. Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. *Protein Pept Lett*. 2010;17:1473–9.
21. Kleffmann T, Hirsch-Hoffmann M, Gruissem W, Baginsky S. plprot: a comprehensive proteome database for different plastid types. *Plant Cell Physiol*. 2006;47:432–6.

Supplementary Table

Table S1. List of computational methods for protein subchloroplast localization.

Method	Details	Application
SubChlo[9]	The method was based on the ET-KNN (evidence theoretic K-nearest neighbor). Using pseudo-amino acid compositions, the method achieved 67.2% overall prediction accuracy.	Predicting subchloroplast localizations of chloroplast proteins from protein sequence.
ChloroRF[13]	The method was based on Random Forest. Using 531 physicochemical properties obtained from AAindex dataset, the method achieved 67.4% overall prediction accuracy.	Predicting subchloroplast localizations of chloroplast proteins from protein sequence.
SubIdent[15]	The method was based on Support Vector Machine. Using features extracted by discreate wavelet transform (DWT) from amino acids' hydrophobicity and polarity values, the method achieved 89.3% overall prediction accuracy in subchloroplast location.	The method can be applied to classify whether a protein is mitochondria or chloroplast protein. If the protein is in mitochondria, then the method can predict its submitochondria location; otherwise predictes its subchloroplast location.
BS-KNN	The method was based on a bit-score weighted K-nearest neighbor (BS-KNN) method for predicting protein subchloroplast locations. The method makes prediction based on the bit-score weighted Euclidean distance calculated from the composition of selected pseudo-amino acids. It achieved about 76% overall prediction accuracy.	Predicting subchloroplast localizations of chloroplast proteins from protein sequence.