Ensemble linear neighborhood propagation for predicting subchloroplast localization of multi-location proteins

Journal:	*Journal of Proteome Research*
Manuscript ID	pr-2016-006869.R1
Manuscript Type:	Technical Note
Date Submitted by the Author:	11-Oct-2016
Complete List of Authors:	Wan, Shibiao; Hong Kong Polytechnic University, Department of Electronic and Information Engineering Mak, Man-Wai; Hong Kong Polytechnic University, Department of Electronic and Information Engineering Kung, Sun-Yuan; Princeton University, Department of Electrical Engineering

Ensemble linear neighborhood propagation for predicting subchloroplast localization of multi-location proteins

Shibiao Wan, *,† Man-Wai Mak, *,† and Sun-Yuan Kung ‡

†*Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China*

‡*Department of Electrical Engineering, Princeton University, New Jersey, USA*

E-mail: shibiao.wan@connect.polyu.hk; enmwmak@polyu.edu.hk

Phone: +852 2766 6257. Fax: +852 2362 8439

Abstract

In the post-genomic era, the number of unreviewed protein sequences is remarkably larger and grows tremendously faster than that of reviewed ones. However, existing methods for protein subchloroplast localization often ignore the information from these unlabelled proteins. This paper proposes a multi-label predictor based on ensemble linear neighborhood propagation (LNP), namely LNP-Chlo, which leverages hybrid sequence-based feature information from both labelled and unlabelled proteins for predicting localization of both single- and multi-label chloroplast proteins. Experimental results on a stringent benchmark dataset and a novel independent dataset suggest that LNP-Chlo performs at least 6% (absolute) better than state-of-the-art predictors. This paper also demonstrates that ensemble LNP significantly outperforms LNP based on individual features. For readers' convenience, the online web-server LNP-Chlo is freely available at `http://bioinfo.eie.polyu.edu.hk/LNPChloServer/`.

Keywords

protein subchloroplast localization; linear neighborhood propagation; multi-label classification; transductive learning; split amino-acid composition

Introduction

As one of the most prominent plant-specific organelles, the chloroplast serves as a specialized subcellular location to conduct photosynthesis, which is arguably the most fundamental biological process maintaining atmospheric oxygen levels and supplying energy and organic compounds for life on Earth [1]. Besides, chloroplast proteins also carry out a series of other molecular functions, such as fatty acid synthesis [2], amino acid biosynthesis [3] and lipid metabolism [4]. Conventionally, the chloroplast can be further divided into a number of microscopic yet intricate structures at the sub-subcellular level, including envelope, thylakoid membrane, thylakoid lumen, stroma and plastoglobule. Knowing where a protein locates in these chloroplast sub-structures can shed light on its biological functions. With the avalanche of novel protein sequences found in the post-genomic era, computational approaches are highly required to assist conventional time-consuming and costly wet-lab techniques for accurate, fast and large-scale prediction of protein subchloroplast localization.

Recent decades have witnessed various *in-silico* approaches applied in protein subcellular localization prediction. These approaches are generally divided into four categories: (1) amino-acid composition-based [5–7], (2) homology-based [8,9], (3) sorting-signals based [10–12] and (4) knowledge-based [13–22]. The first three categories are often regarded as sequence-based methods. Yet, because subchloroplast localization is more microscopic than subcellular localization, not all aforementioned methods that work very well for the former can be readily applied to the latter. To the best of our knowledge, only a few predictors are capable of predicting protein subchloroplast localization which includes BS-KNN [23], SubIdent [24], ChloroRF [25] and SubChlo [26]. Among these predictors, BS-KNN and SubChlo use the K-nearest neighbor (KNN) classifier whereas SubIdent and ChloroRF use more advanced classifiers such as support vector machines (SVM) and random forest (RF). All of these four predictors use amino-acid sequence-based information as features.

However, these subchloroplast-localization predictors become ineffective when dealing with cases where both single- and multi-location chloroplast proteins are involved. This prob-

lem becomes a grave concern when more and more chloroplast proteins are found to co-locate in more than one subchloroplast compartments. For example, Ferredoxin-NADP reductase (leaf isozyme 2)[27] is found to co-reside in both chloroplast stroma and thylakoid membrane; glyceraldehyde phosphate dehydrogenase[28] can co-locate in both chloroplast envelope and stroma. Recently, two multi-label subchloroplast-localization predictors have been proposed, namely MultiP-SChlo[29] and AL-KNN[14].[a] Both predictors can predict single- and multi-label chloroplast proteins, and they use pseudo amino-acid composition (PseAA)[5] as features followed by a genetic algorithm for feature selection. In terms of classification, MultiP-SChlo uses a multi-label SVM classifier while AL-KNN uses a multi-label KNN classifier. The former is found to outperform the latter[29]. Nevertheless, the performance of both predictors is still far from satisfactory. Moreover, previous studies have suggested that the evolutionary background of plant proteins is correlated with their subcellular localization[30] and that predictors not considering the N-terminal modifications of proteins have a higher chance of making false predictions of chloroplast localization[31]. Therefore, evolutionary based features and N-terminal features should be considered for reliable protein subchloroplast localization.

Actually, all of the aforementioned computational approaches (no matter single-label or multi-label) predict the subchloroplast localization of proteins by extracting feature information from the training proteins (or reviewed[b]/labelled proteins) only. They often ignore the information from those unreviewed/unlabelled proteins. In fact, recent advances in high-throughput genome sequencing projects lead to a larger number of novel yet unreviewed protein sequences than that of reviewed ones. Moreover, the former increase at a much faster pace than the latter. For example, the numbers of reviewed and unreviewed protein sequences on 02-Feb-2004 are 137,916 and 895,002, respectively, whereas those numbers on 17-May-2016 become 551,193 and 62,148,086, respectively. This means that the ratio of the number of reviewed/unreviewed protein sequences has been remarkably widen from 1:6 to

[a] Note that AL-KNN was implemented in[29].

[b] The *reviewed* proteins should be those proteins that are manually annotated, whereas the *unreviewed* proteins are those that are not manually annotated.

1:112. Therefore, it is unwise to ignore the information from the unlabelled proteins.

Our recent finding[32] suggests that a transductive-learning approach that exploits information from both labelled and unlabelled proteins can achieve a much higher prediction accuracy than the conventional approaches that only rely on labelled proteins. In [32], a multi-label multi-class predictor called EnTrans-Chlo was proposed. EnTrans-Chlo uses ensemble features comprising PseAA sequence information and profile-based evolutionary information from both labelled and unlabelled proteins, which are classified by a multi-label transductive algorithm based on least squares and nearest neighbors. However, EnTrans-Chlo has the following drawbacks: (1) it uses a similarity-based feature-vector construction method, which restricts the feature vectors to be pairwise-similarity based only; (2) it adopts a classification scheme that minimizes the least squared error between the predicted score vectors and their nearest neighbors with their pairwise-similarity weighting applied to the nearest neighbors, meaning that the weights of the nearest neighbors are probably not optimized; and (3) it uses the PseAA features which are found to perform poorly and should be replaced by better features.

To address these problems, this paper proposes a multi-label predictor based on ensemble linear neighborhood propagation (LNP), namely LNP-Chlo, for predicting subchloroplast localization of both single- and multi-location proteins. Compared to conventional multi-label predictors, LNP-Chlo can leverage information from both labelled and unlabelled proteins. Compared to EnTrans-Chlo, LNP-Chlo adopts a multi-label classifier based on ensemble LNP, which allows various kinds of input features with different dimensions, and at the same time adopts a quadratic programming method to optimize the weights of nearest neighbors. In addition, LNP-Chlo uses the split amino-acid composition (SAAC) features to replace the PseAA features to improve the performance. Experiential results demonstrate that LNP-Chlo performs significantly better than state-of-the-art multi-label predictors. Moreover, this paper also found that the SAAC features and profile-alignment features are complementary with each other for protein subchloroplast localization prediction.

Feature Extraction

In this paper, we extract two kinds of sequence-based features from amino acid sequences: split amino-acid composition features and profile-alignment features.

Split Amino-Acid Composition Features

Previous studies[11,12] have indicated that sorting signals may exist in the short segment of amino acid sequences around the N-terminus, particularly for chloroplast proteins. Therefore, more specific information can be extracted from different regions of a protein sequence independently. To this end, a method named split amino-acid composition (SAAC) was proposed[33,34].

Given a protein, its sequence is first split into three mutually exclusive regions: N-terminal, middle and C-terminal. Then, the frequencies of occurrences of the 20 amino acids in each of these three segments are counted,[c] which are then uniformly normalized by the length of the whole sequence. Mathematically, given a query protein $\mathbb{Q}_i$ with sequence length L_i, its SAAC feature vector is:

$$\mathbf{q}_i^{\mathrm{SAAC}} = \frac{1}{L_i}\left[\underbrace{f_{i,1}^N,\ldots,f_{i,20}^N}_{\text{N-terminal region}},\underbrace{f_{i,1}^M,\ldots,f_{i,20}^M}_{\text{middle region}},\underbrace{f_{i,1}^C,\ldots,f_{i,20}^C}_{\text{C-terminal region}}\right]^{\mathsf{T}}, \quad (1)$$

where $L_i = \sum_{u=1}^{20}(f_{i,u}^N + f_{i,u}^M + f_{i,u}^C)$. In Eq. 1, $\{f_{i,u}^N\}_{u=1}^{20}$, $\{f_{i,1}^M\}_{u=1}^{20}$ and $\{f_{i,1}^C\}_{u=1}^{20}$ are the frequencies of occurrences of the u-th amino acid in the N-terminal region, middle region and C-terminal region, respectively, of the i-th protein. For simplicity and according to previous studies[35], we set the lengths of both N-terminal and C-terminal regions to 25.[d]

Because of its simplicity and efficiency, SAAC has been widely applied to various domains, including multi-functional enzyme classification[35], mitochondrial protein identification[33] and membrane protein prediction[34].

[c] We ignore those non-standard amino acid residues.
[d] Note that all proteins of interest have more than 50 amino acid residues.

Profile-Alignment Features

The profile of a protein contains its sequence evolutionary information, which is usually represented by two matrices: a position-specific scoring matrix (PSSM) and a position-specific frequency matrix (PSFM). The columns of PSSM and PSFM correspond to the position of residues along the protein sequence. For each column in a PSSM, the entries represent the log-likelihood of residue substitutions at that position. Each column of a PSFM contains the weighted observation frequencies of amino acid residues at the corresponding position of the aligned sequences. Both PSSM and PSFM can be obtained from performing multiple sequence alignments on a large protein database (e.g., Swiss-Prot) using PSI-BLAST (position-specific iterative BLAST)[36]. PSI-BLAST involves an iterative search process in which the profile of a query protein is searched against the database to iteratively update itself to detect distant relationships between protein families. Thus, the profile of a protein encapsulates the information of its homologs. Typically, the E-value cutoff and the number of iterations for PSI-BLAST are set to 0.001 and 3, respectively.

The similarity score between a known and an unknown protein sequence can be computed by aligning the profile of the known sequence with that of the unknown sequence[37]. Given a query protein $\mathbb{Q}_i$, we align its profile with the profile of every protein in a dataset of interest to form an alignment score vector $\mathbf{q}_i$. Then, the profile-alignment (PA) feature vector for the i-th protein is computed as:

$$\mathbf{q}_i^{\mathrm{PA}} = [q_{i,1}^{(g)}, \ldots, q_{i,j}^{(g)}, \ldots, q_{i,N}^{(g)}]^{\mathsf{T}}, \tag{2}$$

where $q_{i,j}^{(g)} = \frac{q_{i,j}}{\sqrt{q_{i,i}q_{j,j}}}$, T is the transpose operator, N is the number of proteins in the dataset, and $q_{i,j}$ is the j-th element of $\mathbf{q}_i$. Details of obtaining the profiles and profile alignment can be found in[38].

Over the years, the profile-based evolutionary features have been extensively used in many bioinformatics domains, such as protein disorder prediction[39], protein subcellular localization

prediction [40] and RNA binding sites prediction [41].

Multi-Label Classification

Multi-Label Linear Neighborhood Propagation

Linear neighborhood propagation (LNP) [42] is a powerful semi-supervised learning method. Essentially, LNP assumes that each instance in a classification problem can be linearly reconstructed by its neighboring instances (either labelled or unlabelled). LNP has been successfully applied to various classification topics, including protein function prediction [43], video annotation [44] and image retrieval [45].

In this work, we extended LNP to multi-label classification and applied it to subchloroplast localization. Without loss of generality, given a dataset of N chloroplast proteins distributed in M subchloroplast locations, the first L proteins are with known subchloroplast location(s) (i.e., the training part), and the localization of the remaining $T(= N-L)$ proteins are to be predicted (i.e., the test part). Denote $\{\mathcal{Y}_i, \mathbf{q}_i\}_{i=1}^{N}$, where $\mathcal{Y}_i \subset \{1, 2, \ldots, M\}$ and $\mathbf{q}_i \in \mathcal{R}^d$ as the label set and the feature vectors, respectively, of this dataset. By using the concept of *transformed labels* [46], the label set of the i-th protein can be converted to a label vector $\mathbf{y}_i = [y_{i,1}, \ldots, y_{i,m}, \ldots, y_{i,M}]^{\mathsf{T}}$, where $y_{i,m} \in \{0, 1\}$. Because this is a multi-label classification problem, for multi-location proteins, $\sum_{m=1}^{M} y_{i,m} > 1$; for single-location proteins, $\sum_{m=1}^{M} y_{i,m} = 1$. For a training protein $(0 < i \leq L)$, $y_{i,m} = 1$ if the i-th protein is located in the m-th subchloroplast location; otherwise, $y_{i,m} = 0$. For a test protein $((L+1) < i \leq N)$, because initially we do not know to which of these M locations the protein belongs, we assume that $y_{i,m} = 0$, $1 \leq m \leq M$.

Then, given the i-th protein $\mathbb{Q}_i$, its feature vector $\mathbf{q}_i$ can be reconstructed from a set of

neighboring proteins, which leads to the following objective function for optimization:

$$\{\widehat{w}_{i,k}\}_{i=1}^{N} = \arg\min_{\substack{\{w_{i,k}\}_{i=1}^{N} \\ k\in\mathcal{K}(i)}} \sum_{i=1}^{N} \left\| \mathbf{q}_i - \sum_{k\in\mathcal{K}(i)} w_{i,k}\mathbf{q}_k \right\|^2, \tag{3}$$

where $\sum_{k\in\mathcal{K}(i)} w_{i,k} = 1, w_{i,k} \geq 0$ is the contribution of $\mathbf{q}_k$ in constructing $\mathbf{q}_i$, and $\mathcal{K}(i)$ is a set of neighbors to the i-th protein. In this work, $\mathcal{K}(i)$ is a set comprising the top-K nearest neighbors.

After some mathematical manipulations, Eq. 3 is equivalent to solving the following quadratic programming problem:

$$\begin{aligned} &\min \sum_{i=1}^{N} \sum_{k,r\in\mathcal{K}(i)} w_{i,k}(\mathbf{q}_i - \mathbf{q}_k)^{\mathsf{T}}(\mathbf{q}_i - \mathbf{q}_k)w_{i,r}, \\ &\text{s.}t. \sum_{k\in\mathcal{K}(i)} w_{i,k} = 1, w_{i,k} \geq 0, i = 1, \ldots, N. \end{aligned} \tag{4}$$

After Eq. 4 is solved, an optimized weight matrix $\mathbf{W}$ can be obtained, whose (i,k)-th entry is $\widehat{w}_{i,k}$.

According to Wang and Zhang [42], predicted score vectors can be determined by propagating the labels of labelled instances to unlabelled instances via an iterative procedure. Let $\{\mathbf{s}_i^t\}_{t=0}^{\infty} \in \mathcal{R}^M$ as the predicted score vector of the i-th protein at the t-th iteration, whose m-th ($m \in \{1, \ldots, M\}$) element $s_{i,m}^t$ represents the score in the m-th class at the t-th iteration. We set the initial $\mathbf{s}_i^0 = \mathbf{y}_i$. Then, the predicted score vector of the i-th protein at the $(t+1)$-th iteration is given by:

$$\mathbf{s}_i^{t+1} = \alpha\mathbf{W}\mathbf{s}_i^t + (1-\alpha)\mathbf{y}_i, \tag{5}$$

where $\alpha \in (0,1)$ is a parameter controlling the amount of label information from the neighboring data for updating the score vector. When Eq. 5 converges [42], we obtain the i-th predicted score vector, which is denoted as $\widehat{\mathbf{s}}_i$, i.e., $\widehat{\mathbf{s}}_i = \lim_{t\to\infty} \mathbf{s}_i^t$. Note that because $\mathbf{W}$ in

Eq. 5 incorporates information from both labelled (training) and unlabelled (test) proteins (see Eq. 4), the way to obtain $\widehat{\mathbf{s}}_i$ is a typical transductive-learning method.

Ensemble LNP

In this work, we adopted a classifier ensemble scheme to incorporate both SAAC features and PA features in our proposed predictor. Denote $\widehat{\mathbf{s}}_i^{\mathrm{SAAC}}$ and $\widehat{\mathbf{s}}_i^{\mathrm{PA}}$ as the LNP scores obtained from Eq. 5 by using the SAAC features and PA features, respectively. Then, the ensemble score can be obtained as follows:

$$\widehat{\mathbf{s}}_i^{\mathrm{en}} = \beta\widehat{\mathbf{s}}_i^{\mathrm{SAAC}} + (1-\beta)\widehat{\mathbf{s}}_i^{\mathrm{PA}}, \tag{6}$$

where $\beta \in [0,1]$ is a parameter controlling the influence of SAAC features and PA features.

To predict proteins with both single- and multi-label locations, a decision scheme for multi-label classification should be used. In this work, we used a decision scheme similar to our previous studies[47–49]. Specifically, the predicted subchloroplast location(s) of the i-th query protein $\mathbb{Q}_i$ are given by:

$$\mathcal{M}^*(\mathbb{Q}_i) = \begin{cases} \bigcup_{m=1}^{M}\left\{m : \widehat{s}_{i,m}^{\mathrm{en}} \geq \min\left(0.5, \theta\widehat{s}_{i,\max}^{\mathrm{en}}\right)\right\}, \\ \qquad\qquad\qquad\qquad \text{where } \exists\ \widehat{s}_{i,m}^{\mathrm{en}} > 0; \\ \arg\max_{m=1}^{M}\widehat{s}_{i,m}^{\mathrm{en}}, \qquad\qquad \text{otherwise}, \end{cases} \tag{7}$$

where

$$\widehat{s}_{i,\max}^{\mathrm{en}} = \max_{m=1}^{M}\widehat{s}_{i,m}^{\mathrm{en}},$$

$\min(\cdot)$ is the minimum operator, and $\widehat{s}_{i,m}^{\mathrm{en}}$ is the (i,m)-th entry of $\widehat{\mathbf{s}}_i^{\mathrm{en}}$ given by Eq. 6. In Eq. 7, $\theta \in (0.0, 1.0]$ is a parameter controlling the ratios of multi-label predictions. A larger θ leads to a stringent criteria; and vice versa.

For ease of reference, we refer to the proposed predictor as LNP-Chlo. The flowchart of LNP-Chlo is shown in Fig. 1.

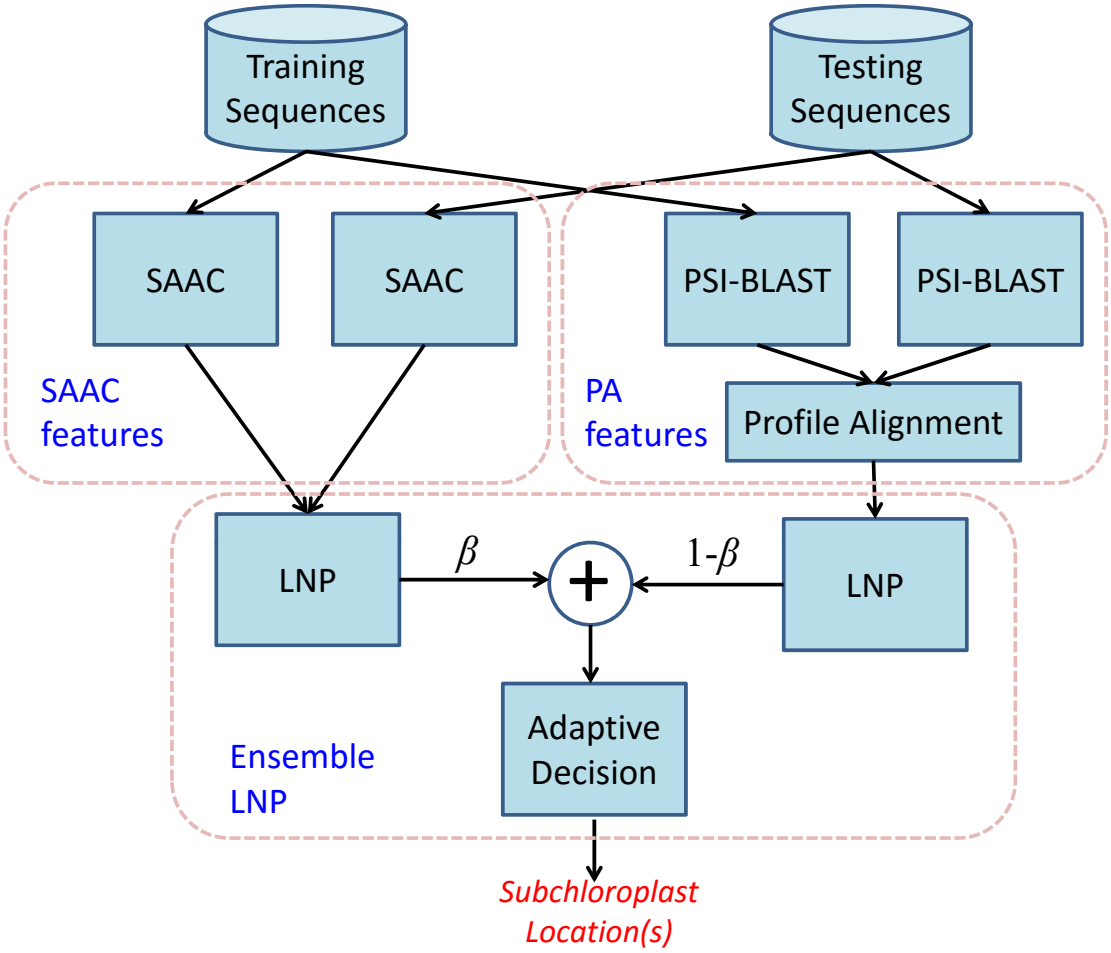


Figure 1: The flowchart of LNP-Chlo. Training sequences: proteins $\{\mathbb{Q}_i\}_{i=1}^{L}$; testing sequences: proteins $\{\mathbb{Q}_i\}_{i=L+1}^{N}$; SAAC: split amino-acid composition; PA: profile-alignment; LNP: linear neighborhood propagation; adaptive decision: the decision scheme given in Eq. 7.

Datasets and Performance Metrics

In this paper, a recent stringent benchmark dataset [29] and a novel independent dataset [13] were used to evaluate the performance of LNP-Chlo. All proteins in the benchmark dataset were added to the Swiss-Prot database before 31-May-2013, whereas those of the novel dataset were added to Swiss-Prot from 1-Jun-2013 and 11-Nov-2015. This guarantees that the novel dataset contains the latest chloroplast proteins that have never been used by other studies and researchers. The sequence identity of the benchmark dataset was cut off to 25%, whereas we did not cut off the sequence similarity of the novel dataset due to the limited number of novel proteins. The benchmark dataset contains 578 actual proteins [13], of which 556 belong to one subchloroplast location, 21 to two locations, 1 to three locations and none to four or more locations. The novel dataset contains 122 actual proteins, of which 113 and 9 are single-location proteins and two-location proteins, respectively. The 578 actual proteins in the benchmark dataset correspond to 601 ($= 556 \times 1 + 21 \times 2 + 1 \times 3$) locative proteins [50],[e] whereas 122 actual proteins in the novel dataset correspond to 131 ($= 113 \times 1 + 9 \times 2$) locative proteins. The specific breakdown of both datasets are shown in Table 1. As can be seen, the majority ($> 70\%$) of proteins in both datasets are located in envelope and thylakoid membrane, while proteins located in the other 3 subchloroplast locations account for less than 30%. This means that both datasets are very imbalanced. Both datasets can be downloadable from the links of the LNP-Chlo web-server.

To facilitate comparison between LNP-Chlo and other multi-label predictors, some popular multi-label measures were used, including *Overall Actual Accuracy (OAA)* [51], *Accuracy, Precision, Recall*, and *F1-score (F1)* [52,53]. For all performance measures, the higher the values, the better the prediction performance. Particularly, *OAA* is the most stringent and objective among these five measures because it requires 'exact-match' of a predicted label set and the corresponding ground-truth label set [54]. Detailed analysis on these metrics can

[e] The number of locative proteins for an actual protein is the number of subchloroplast compartments where the actual protein co-locates.

Table 1: Breakdowns of the benchmark and novel datasets. All proteins of the benchmark dataset were added to Swiss-Prot before 31-May-2013, whereas those of the novel dataset were added to Swiss-Prot from 1-Jun-2013 to 11-Nov-2015. *: no chloroplast plastoglobule proteins were found when the novel proteins were retrieved from Swiss-Prot.

Label	Location	No. of Proteins	
		Benchmark	Novel
1	Envelope	199	61
2	Stroma	105	26
3	Thylakoid lumen	34	5
4	Thylakoid membrane	233	39
5	Plastoglobule	30	0*
Number of locative proteins		601	131
Number of actual proteins		578	122

be found in supplementary materials of the LNP-Chlo web-server.

To strike a good balance among all of the performance measures, we propose a new measure, namely *GrandMean*, which is defined as:

$$GrandMean = \frac{1}{5}(OAA + Accuracy + Precision + Recall + F1). \quad (8)$$

Obviously, the higher the *GrandMean*, the better the prediction performance. Since *GrandMean* incorporates all of the aforementioned performance measures, we used it as the criteria for parameter optimization in our algorithm.

We used both leave-one-out cross-validation (LOOCV) and independent tests for evaluating classifiers' performance. These statistical methods were used because LOOCV is regarded as the most rigorous and bias-free procedure [55] and independent tests can demonstrate the generalization capabilities of classifiers [38].

Table 2: Comparing LNP-Chlo with state-of-the-art multi-label predictors on the benchmark dataset based on leave-one-out cross-validation (LOOCV) tests. The results of AL-KNN reported here were extracted from [29].

Measures	Predictors			
	AL-KNN [14]	MultiP-SChlo [29]	EnTrans-Chlo [32]	LNP-Chlo
OAA	0.4377	0.5552	0.6003	**0.6609**
Accuracy	0.4521	0.6326	0.6600	**0.7085**
Precision	0.4663	0.6410	0.6730	**0.7226**
Recall	0.4530	0.7106	0.7106	**0.7437**
F1	0.4595	0.6738	0.6804	**0.7249**
GrandMean	0.4537	0.6426	0.6649	**0.7121**

Results and Discussion

Comparing with State-of-the-Art Predictors

Table 2 compares LNP-Chlo against several state-of-the-art multi-label chloroplast predictors on the benchmark dataset based on LOOCV. As far as we know, only three existing multi-label predictors, namely MultiP-SChlo [29], AL-KNN [14] and EnTrans-Chlo [32], are designed to predict both single- and multi-location chloroplast proteins. Note that AL-KNN was implemented in [29]. From the perspective of feature extraction, both AL-KNN and MultiP-SChlo use pseudo amino-acid composition (PseAA) features followed by a genetic algorithm for feature selection, whereas EnTrans-Chlo uses features derived from PseAA and profile-alignment. From the perspective of classification, the former two use a multi-label SVM classifier and a multi-label KNN classifier, respectively, whereas the latter uses a multi-label classifier based on least squares and nearest neighbors. Our proposed predictor LNP-Chlo uses profile-alignment features and SAAC features and adopts an ensemble LNP-based multi-label classifier. In addition, EnTrans-Chlo and LNP-Chlo can exploit features from both labelled and unlabelled data, whereas the other two were trained on labelled data only.

As shown in Table 2, LNP-Chlo significantly outperforms the other three predictors in

terms of all performance measures. Particularly, for the most stringent and object criteria *OAA*, LNP-Chlo outperforms EnTrans-Chlo, Multi-SChlo and AL-KNN by 6% (absolute), 11% (absolute) and 23% (absolute), respectively. We noticed that the performance of LNP-Chlo and EnTrans-Chlo surpasses that of the other two by a large margin. This is possibly because transductive-learning based predictors are more powerful than conventional inductive-learning based predictors, which will be confirmed by the in-depth analysis in Section **'Transductive versus Non-Transductive'**.

To confirm that the improvement of LNP-Chlo over state-of-the-art predictors is statistically significant, we performed McNemar's tests [56,57] on the prediction scores of LNP-Chlo and the existing top-performing predictor EnTrans-Chlo. We found that the p-value between the *OAA* of LNP-Chlo and EnTrans-Chlo is 1.1225×10^{-6} ($\ll 0.05$), suggesting that, statistically speaking, the performance of LNP-Chlo is significantly better that of EnTrans-Chlo.

To further demonstrate the superiority of LNP-Chlo over state-of-the-art predictors, we performed independent tests for all of the four aforementioned predictors. Specifically, 20% of the benchmark dataset were randomly chosen as the test dataset and the remaining samples were used to train the four predictors. This procedure was repeated ten times to test the robustness of the predictors for different random selections. The performance comparisons are shown in Fig. 2. As can be seen, the same conclusion as in Table 2 can be drawn from Fig. 2: LNP-Chlo impressively outperforms the other three predictors in terms of all performance metrics. Besides, the performance of the top two predictors (LNP-Chlo and EnTrans-Chlo) is still remarkably superior to the other predictors, which further demonstrate the effectiveness of transductive-learning models.

Transductive versus Non-Transductive

To unravel the advantages of the proposed transductive model, we compared our proposed predictor LNP-Chlo with a state-of-the-art non-transductive predictor. We selected the multi-label SVM (ML-SVM) as the non-transductive model due to its superior performance

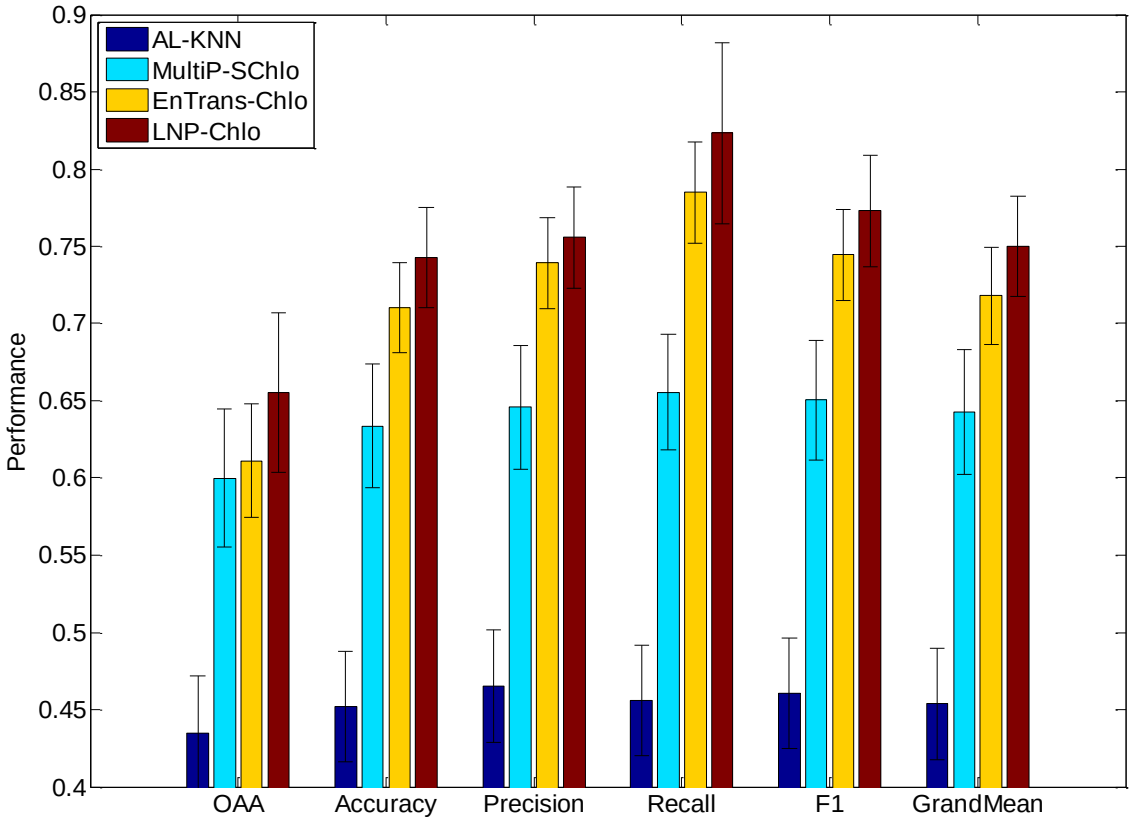


Figure 2: Comparing LNP-Chlo with state-of-the-art multi-label predictors on the benchmark dataset based on independent tests. Bars and errorbars denote the mean and standard deviation of different performance measures. AL-KNN, MultiP-SChlo and EnTrans-Chlo are from [14], [29] and [32], respectively. The results of AL-KNN reported here were extracted from [29].

Table 3: Comparing LNP-Chlo with a non-transductive predictor (ML-SVM) based on leave-one-out cross-validation.

Measures	Predictors	
	ML-SVM [46]	LNP-Chlo
OAA	0.6194	**0.6609**
Accuracy	0.6332	**0.7085**
Precision	0.6401	**0.7226**
Recall	0.6401	**0.7437**
F1	0.6378	**0.7249**
GrandMean	0.6341	**0.7121**

demonstrated in various bioinformatics domains, including subchloroplast localization prediction (e.g., MultiP-Schlo [29]) and protein subcellular localization (e.g., mGOASVM [46]). Both ML-SVM and LNP-Chlo use the same features (both PA features and SAAC features) and adopt the same ensemble scheme (see Section **'Ensemble LNP'**) for classification.

As can be seen from Table 3, the performance of LNP-Chlo is impressively superior to that of ML-SVM in terms of all performance measures, suggesting that using transductive models is better than non-transductive models for predicting protein subchloroplast localization.

Ensemble LNP versus Individual LNP

Table 4: Comparing ensemble LNP against LNP with individual features on the benchmark dataset based on LOOCV. Pse-LNP-Chlo, Pro-LNP-Chlo and SAAC-LNP-Chlo use pseudo amino-acid composition features, profile-alignment (PA) features and SAAC features, respectively, whereas LNP-Chlo uses both PA and SAAC features.

Measures	Predictors		
	Pro-LNP-Chlo	SAAC-LNP-Chlo	LNP-Chlo
OAA	0.6228	0.6453	**0.6609**
Accuracy	0.6560	0.6597	**0.7085**
Precision	0.6684	0.6747	**0.7226**
Recall	0.6796	0.6597	**0.7437**
F1	0.6675	0.6646	**0.7249**
GrandMean	0.6589	0.6608	**0.7121**

To investigate the benefits of using ensemble LNP, we compared the ensemble LNP (LNP-Chlo) against LNP that uses individual features on the benchmark dataset based on LOOCV tests. We named LNP with PA features and SAAC features as PA-LNP-Chlo and SAAC-LNP-Chlo, respectively. Table 4 shows the results, which demonstrates that using the ensemble LNP performs better than PA-LNP-Chlo and SAAC-LNP-Chlo in terms of all performance measures. Particularly, the *OAA* of the former is around 2% (absolute) and 4% (absolute), respectively, better than those of the latter two. The results suggest that the PA features and SAAC features are complementary with each other for predicting protein

subchloroplast localization. Besides, we noticed that the performance difference of SAAC-LNP-Chlo and PA-LNP-Chlo in all performance metrics is by no means considerable. We conjecture that this is usually a basic precondition for successful ensemble classification.

To verify the above conjecture, we have also investigated the performance of LNP with the PseAA features, which we name as Pse-LNP-Chlo. Totally, we have three different features: PseAA, profile-alignment (PA) and SAAC. Based on these individual features, besides LNP-Chlo, we constructed three more ensemble LNP with different combinations of features, namely PseAA + Pro, PseAA + SAAC and PseAA + PA + SAAC, which we name as PsePro-LNP-Chlo, PseSAAC-LNP-Chlo and All-LNP-Chlo, respectively. The results of these seven predictors are shown in Fig. 3. As can be seen, in term of individual features, the performance of Pse-LNP-Chlo is much worse than that of Pro-LNP-Chlo and SAAC-LNP-Chlo, suggesting that the latter is probably suitable for being combined with other features. In terms of the ensemble LNP with hybrid features, we found that our proposed predictor LNP-Chlo performs the best. Particularly, LNP-Chlo performs better than All-LNP-Chlo–the predictor that uses all of the three features. This is probably because PseAA features contribute negatively to the final performance of All-LNP-Chlo, leading to poorer performance. Therefore, we dropped the PseAA features, and adopted only PA features and SAAC features.

Predicting Novel Proteins

A powerful bioinformatics predictor should possess good generalization capabilities, which can be directly reflected by predicting novel independent tests. To further demonstrate the good generalization capabilities of LNP-Chlo, we created a novel and independent dataset (See Table 1). To guarantee the strict objectivity of the independent tests, all of the proteins in the novel dataset were added to Swiss-Prot later than those in the benchmark dataset, which are used as the training set. This novel dataset contains all the proteins added to Swiss-Prot between 1-Jun-2013 and 11-Nov-2015, and no similarity cutoff technique is adopted due

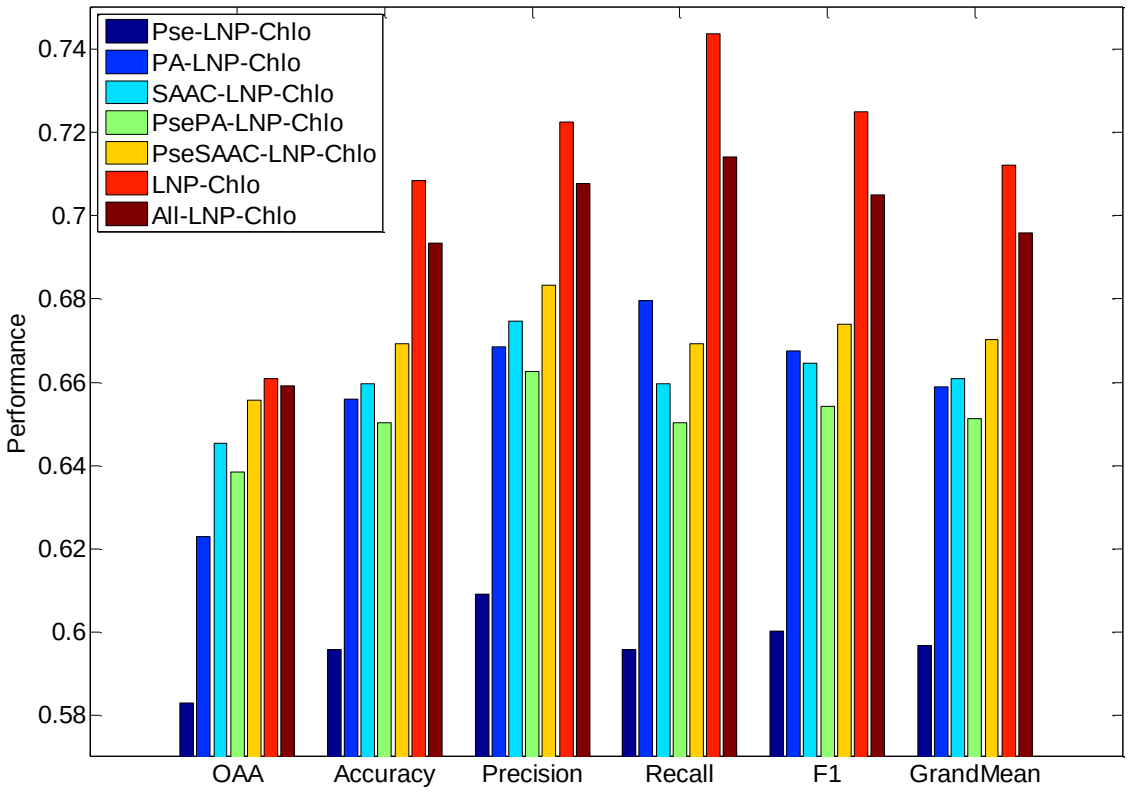


Figure 3: Comparing LNP-Chlo against LNP with individual features and ensemble LNP with various features on the benchmark dataset based on LOOCV. Pse-LNP-Chlo, Pro-LNP-Chlo and SAAC-LNP-Chlo use pseudo amino-acid composition features, profile-alignment (PA) features and SAAC features, respectively, whereas PsePro-LNP-Chlo, PseSAAC-LNP-Chlo, LNP-Chlo and All-LNP-Chlo use features of PseAA + PA, PseAA + SAAC, PA + SAAC and PseAA + PA + SAAC, respectively.

Table 5: Comparing LNP-Chlo with state-of-the-art multi-label predictors on the novel dataset based on independent tests. The benchmark dataset was used as the training set for all predictors. The results of MultiP-SChlo and EnTrans-Chlo were obtained from their web-servers.

Measures	Predictors		
	MultiP-SChlo[29]	EnTrans-Chlo[32]	LNP-Chlo
OAA	0.2705	0.3607	**0.5492**
Accuracy	0.3279	0.4631	**0.5738**
Precision	0.3525	0.4850	**0.5984**
Recall	0.3607	0.5492	**0.5738**
F1	0.3470	0.4986	**0.5820**
GrandMean	0.3317	0.4713	**0.5754**

to the limited number of novel proteins.

Table 5 compares LNP-Chlo against state-of-the-art multi-label predictors by using independent tests on the novel dataset. The benchmark dataset was used for training. The performance of MultiP-SChlo is based on the results of its web-server. As can be seen, LNP-Chlo outperforms both MultiP-SChlo and EnTrans-Chlo by at least 10% (absolute) in terms of all performance measures except *Recall*, for which the former is 3% (absolute) and 21% (absolute) better than EnTrans-Chlo and MultiP-SChlo, respectively. The results suggest that LNP-Chlo is more capable of predicting novel proteins than MultiP-SChlo and EnTrans-Chlo.

Moreover, the specific prediction results of LNP-Chlo and EnTrans-Chlo on the novel dataset are shown in Section S4 of the supplementary materials. Generally speaking, LNP-Chlo can correctly predict proteins in *envelope* and *thylakoid membrane* with higher accuracies than those in other locations. This is understandable because in the training benchmark dataset, these two subchloroplast locations constitute the major part of the whole dataset, making LNP-Chlo better trained in these two locations. Actually, most machine learning based predictors (e.g., EnTrans-Chlo) suffer from the insufficient-data problem. The prediction performance of LNP-Chlo in other locations can be improved when more and more chloroplast proteins in other locations are available for training.

Conclusion

In this paper, we propose an ensemble LNP based predictor called LNP-Chlo, which can exploit information from both labelled and unlabelled data for predicting localization of chloroplast proteins at the sub-subcellular level. Specifically, LNP-Chlo first extracts composition-based sequence information and profile-based evolutionary information, which are respectively used to train an LNP-based multi-label classifier. Subsequently, the scores for these two LNP classifiers are combined to make the final decisions. Experimental results on a

stringent benchmark dataset and a novel dataset demonstrate the superiority of LNP-Chlo over state-of-the-art predictors. The main contributions of this paper are summarized below:

1. The proposed LNP-Chlo outperforms state-of-the-art subchloroplast-localization predictors.
2. LNP-Chlo leverages information from both labelled and unlabelled proteins.
3. The proposed ensemble LNP performs remarkably better than the LNP based on individual features as well as the ensemble LNP with other hybrid features.
4. Profile-alignment features and SAAC features are complementary with each other for predicting protein subchloroplast localization.

Acknowledgement

The author thanks the anonymous reviewers for their comments and suggestions which have helped strengthen the presentation of this article. This work was in part supported by the RGC of Hong Kong SAR Grant No. PolyU152068/15E, and the Brandeis Program of the Defense Advanced Research Project Agency (DARPA) and Space and Naval Warfare System Center Pacific (SSC Pacific) under Contract No. 66001-15-C-4068.

References

(1) Bryant, D. A.; Frigaard, N. U. Prokaryotic photosynthesis and phototrophy illuminated. *Trends in Microbiology* **2006**, *14*, 488–496.

(2) Post-Beittenmiller, D.; Roughan, G.; Ohlrogge, J. B. Regulation of plant fatty acid biosynthesis analysis of acyl-coenzyme A and acyl-acyl carrier protein substrate pools in spinach and pea chloroplasts. *Plant Physiology* **1992**, *100*, 923–930.

(3) Kirk, P. R.; Leech, R. M. Amino acid biosynthesis by isolated chloroplasts during photosynthesis. *Plant Physiology* **1972**, *50*, 228–234.

(4) Wang, Z.; Benning, C. Chloroplast lipid synthesis and lipid trafficking through ER-plastid membrane contact sites. *Biochemical Society Transactions* **2012**, *40*, 457.

(5) Chou, K. C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Structure, Function, and Genetics* **2001**, *43*, 246–255.

(6) Fan, G. L.; Li, Q. Z. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. *Journal of Theoretical Biology* **2012**, *304*, 88–95.

(7) Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *PROTEINS: Structure, Function, and Genetics* **2003**, *50*, 44–48.

(8) Mott, R.; Schultz, J.; Bork, P.; Ponting, C. Predicting protein cellular localization using a domain projection method. *Genome research* **2002**, *12*, 1168–1174.

(9) Mak, M. W.; Guo, J.; Kung, S. Y. PairProSVM: Protein Subcellular Localization Based on Local Pairwise Profile Alignment and SVM. *IEEE/ACM Trans. on Computational Biology and Bioinformatics* **2008**, *5*, 416 – 422.

(10) Emanuelsson, O.; Nielsen, H.; Brunak, S.; von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **2000**, *300*, 1005–1016.

(11) Nakai, K. Protein sorting signals and prediction of subcellular localization. *Advances in Protein Chemistry* **2000**, *54*, 277–344.

(12) Emanuelsson, O. Predicting protein subcellular localisation from amino acid sequence information. *Briefings in Bioinformatics* **2002**, *3*, 361–376.

(13) Wan, S.; Mak, M. W.; Kung, S. Y. Sparse regressions for predicting and interpreting subcellular localization of multi-label proteins. *BMC Bioinformatics* **2016**, *17*, 97.

(14) Lin, W. Z.; Fang, J. A.; Xiao, X.; Chou, K. C. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Molecular BioSystems* **2013**, *9*, 634–644.

(15) Chou, K. C.; Wu, Z. C.; Xiao, X. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Molecular BioSystems* **2012**, *8*, 629–641.

(16) Wan, S.; Mak, M. W.; Kung, S. Y. mLASSO-Hum: A LASSO-based interpretable human-protein subcellular localization predictor. *Journal of Theoretical Biology* **2015**, *382*, 223–234.

(17) Mei, S. Multi-label multi-kernel transfer learning for human protein subcellular localization. *PLoS ONE* **2012**, *7*, e37716.

(18) Wan, S.; Mak, M. W.; Kung, S. Y. R3P-Loc: A compact multi-label predictor using ridge regression and random projection for protein subcellular localization. *Journal of Theoretical Biology* **2014**, *360*, 34–45.

(19) Chou, K. C.; Shen, H. B. Predicting Eukaryotic Protein Subcellular Location by Fusing Optimized Evidence-Theoretic K-Nearest Neighbor Classifiers. *J. of Proteome Research* **2006**, *5*, 1888–1897.

(20) Wan, S.; Mak, M. W.; Kung, S. Y. Semantic Similarity over Gene Ontology for Multi-Label Protein Subcellular Localization. *Engineering* **2013**, *5*, 68–72.

(21) Chou, K. C.; Cai, Y. D. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem. Biophys. Res. Commun.* **2004**, *320*, 1236–1239.

(22) Wan, S.; Mak, M. W.; Kung, S. Y. GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. *Journal of Theoretical Biology* **2013**, *323*, 40–48.

(23) Hu, J.; Yan, X. BS-KNN: An effective algorithm for predicting protein subchloroplast localization. *Evolutionary Bioinformatics* **2012**, *8*, 79–87.

(24) Shi, S.-P.; Qiu, J.-D.; Sun, X.-Y.; Huang, J.-H.; Huang, S.-Y.; Suo, S.-B.; Liang, R.-P.; Zhang, L. Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction. *Biochimica et Biophysica Acta* **2011**, *1813*, 424–430.

(25) Tung, C.-W.; Liaw, C.; Ho, S.-J.; Ho, S.-Y. Prediction of protein subchloroplast locations using random forests. Proceeding of World Academy of Science, Engineering and Technology. 2010; pp 699–703.

(26) 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. *Journal of Theoretical Biology* **2009**, *261*, 330–335.

(27) Hanke, G. T.; Okutani, S.; Satomi, Y.; Takao, T.; Suzuki, A.; Hase, T. Multiple iso-proteins of FNR in Arabidopsis: evidence for different contributions to chloroplast function and nitrogen assimilation. *Plant, Cell and Environment* **2005**, *28*, 1146–1157.

(28) Ferro, M.; Salvi, D.; Brugière, S.; Miras, S.; Kowalski, S.; Louwagie, M.; Garin, J.; Joyard, J.; Rolland, N. Proteomics of the chloroplast envelope membranes from Arabidopsis thaliana. *Molecular and Cellular Proteomics* **2003**, *2*, 325–345.

(29) Wang, X.; Zhang, W.; Zhang, Q.; Li, G. Z. MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. *Bioinformatics* **2015**, *31*, 2639–2645.

(30) Bayer, R. G.; Kostler, T.; Jain, A.; Stael, S.; Ebersberger, I.; Teige, M. Higher plant proteins of cyanobacterial origin: are they or are they not preferentially targeted to chloroplasts? *Molecular Plant* **2014**, *7*, 1797–1800.

(31) Stael, S.; Bayer, R. G.; Mehlmer, N.; Teige, M. Protein N-acylation overrides differing targeting signals. *FEBS letters* **2011**, *585*, 517–522.

(32) Wan, S.; Mak, M. W.; Kung, S. Y. Transductive learning for multi-label protein subchloroplast localization prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2016**, *to appear*, 1–13.

(33) Verma, R.; Varshney, G. C.; Raghava, G. P. S. Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. *Amino Acids* **2010**, *39*, 101–110.

(34) Hayat, M.; Khan, A.; Yeasin, M. Prediction of membrane proteins using split amino acid and ensemble classification. *Amino acids* **2012**, *42*, 2447–2460.

(35) Zou, H. L.; Xiao, X. Classifying multifunctional enzymes by incorporating three different models into Chou's general pseudo amino acid composition. *The Journal of Membrane Biology* **2016**, *249*, 551–557.

(36) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

(37) Rangwala, H.; Karypis, G. Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics* **2005**, *21*, 4239–4247.

(38) Wan, S.; Mak, M. W. *Machine learning for protein subcellular localization prediction*; De Gruyter: Berlin, Germany, 2015.

(39) Su, C. T.; Chen, C. Y.; Ou, Y. Y. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics* **2006**, *7*, 319.

(40) Zhu, L.; Yang, J.; Shen, H.-B. Multi label learning for prediction of human protein subcellular localizations. *The Protein Journal* **2009**, *28*, 384–390.

(41) Kumar, M.; Gromiha, M. M.; Raghava, G. P. S. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *Journal of Molecular Recognition* **2011**, *24*, 303–313.

(42) Wang, F.; Zhang, C. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering* **2008**, *20*, 55–67.

(43) Yu, G.; Zhu, H.; Domeniconi, C.; Guo, M. Integrating multiple networks for protein function prediction. *BMC Systems Biology* **2015**, *9*, S3.

(44) Tang, J.; Hua, X. S.; Wang, M.; Gu, Z.; Qi, G. J.; Wu, X. Correlative linear neighborhood propagation for video annotation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **2009**, *39*, 409–416.

(45) Li, F.; Dai, Q.; Xu, W.; Er, G. Multilabel neighborhood propagation for region-based image retrieval. *IEEE Transactions on Multimedia* **2008**, *10*, 1592–1604.

(46) Wan, S.; Mak, M. W.; Kung, S. Y. mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinformatics* **2012**, *13*, 290.

(47) Wan, S.; Mak, M. W.; Kung, S. Y. mPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. *Analytical Biochemistry* **2015**, *473*, 14–27.

(48) Wan, S.; Mak, M. W. Predicting subcellular localization of multi-location proteins by

improving support vector machines with an adaptive-decision scheme. *International Journal of Machine Learning and Cybernetics* **2015**, *to appear*, 1–13.

(49) Wan, S.; Mak, M. W.; Kung, S. Y. Mem-ADSVM: A two-layer multi-label predictor for identifying multi-functional types of membrane proteins. *Journal of Theoretical Biology* **2016**, *398*, 32–42.

(50) Xiao, X.; Wu, Z. C.; Chou, K. C. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *Journal of Theoretical Biology* **2011**, *284*, 42–51.

(51) Wan, S.; Mak, M. W.; Kung, S. Y. Mem-mEN: Predicting multi-functional types of membrane proteins by interpretable elastic nets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2015**, *13*, 706–718.

(52) Tsoumakas, G.; Katakis, I.; Vlahavas, I. Mining multi-label data. Data Mining and Knowledge Discovery Handbook, O. Maimon, l. Rokach (Ed.). Springer, 2nd edition. 2010; pp 667–685.

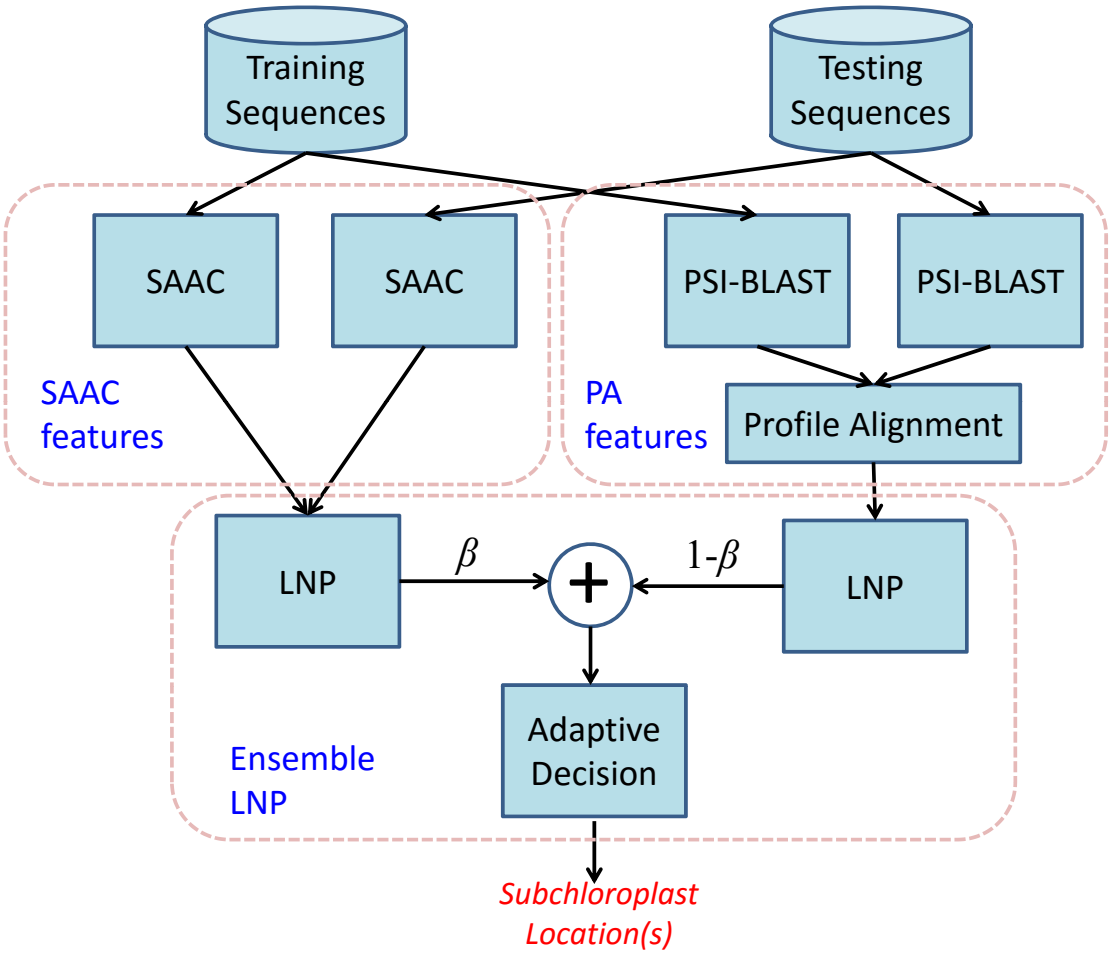
(53) Schapire, R. E.; Singer, Y. Boostexter: A boosting-based system for text categorization. *Machine Learning* **2000**, *39*, 135–168.

(54) Wan, S.; Mak, M. W.; Kung, S. Y. HybridGO-Loc: Mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins. *PLoS ONE* **2014**, *9*, e89545.

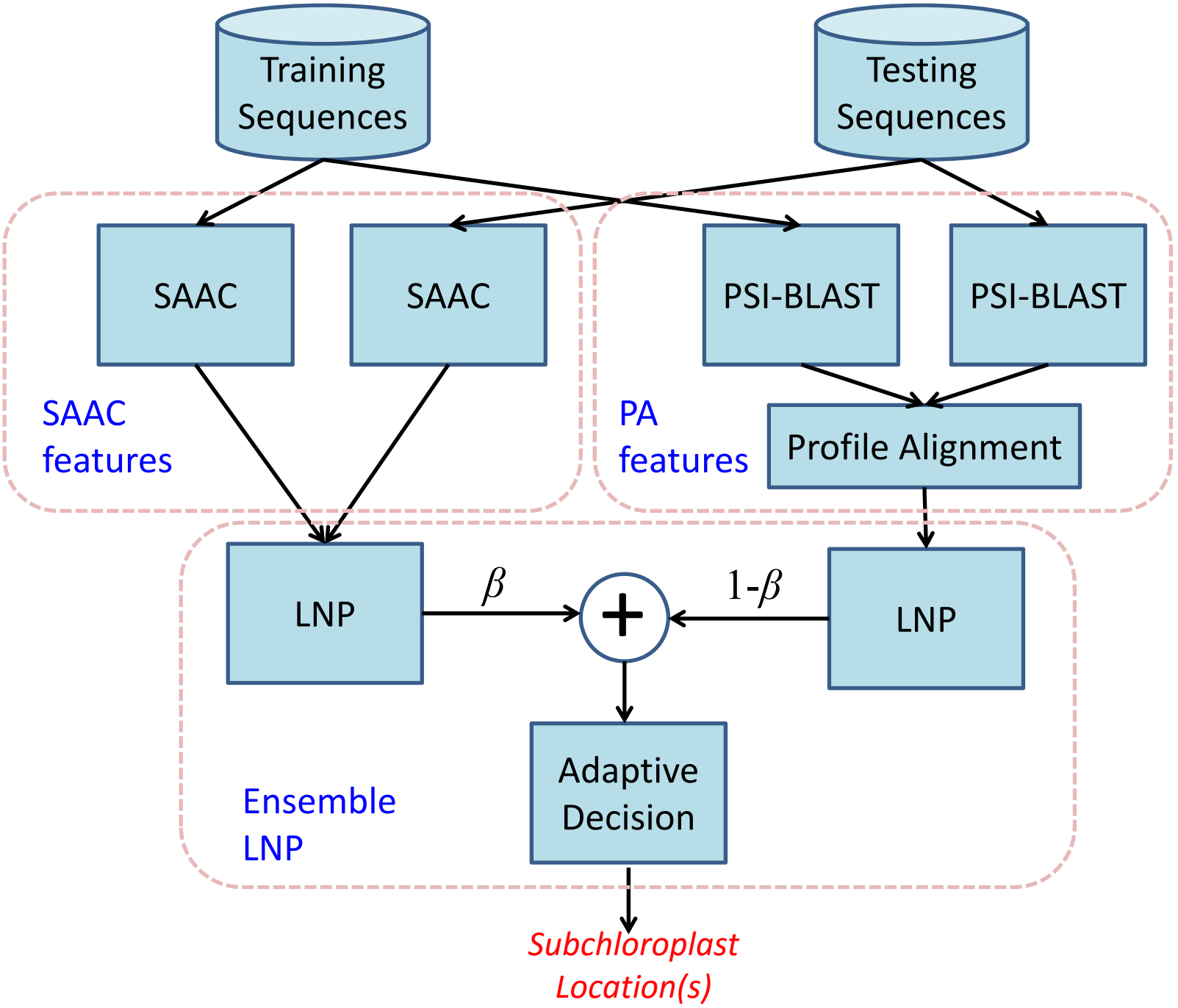
(55) Hastie, T.; Tibshirani, R.; Friedman, J. *The element of statistical learning*; Springer-Verlag: Berlin, Germany, 2001.

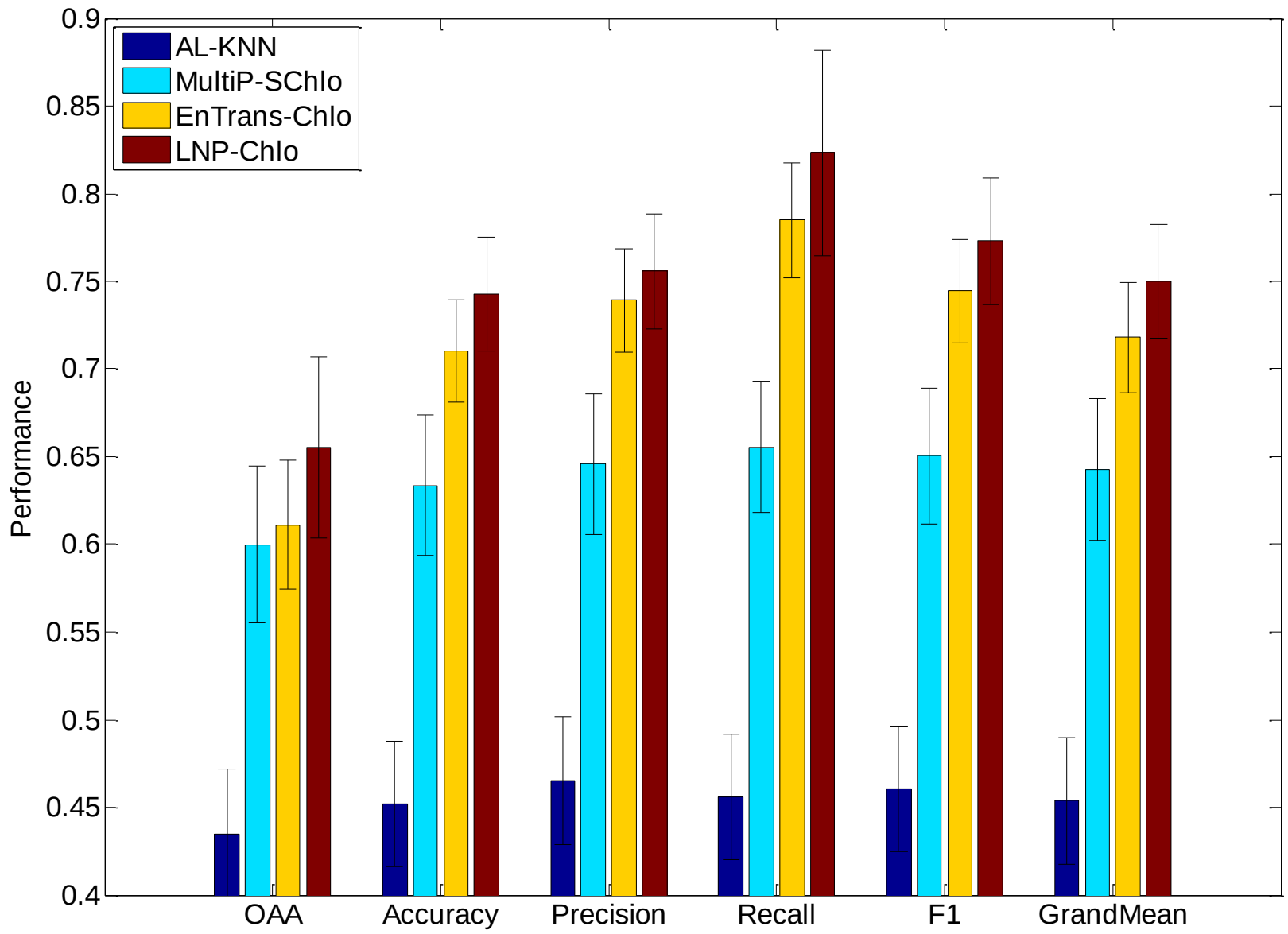
(56) Gillick, L.; Cox, S. J. Some statistical issues in the comparison of speech recognition algorithms. 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'89). 1989; pp 532–535.

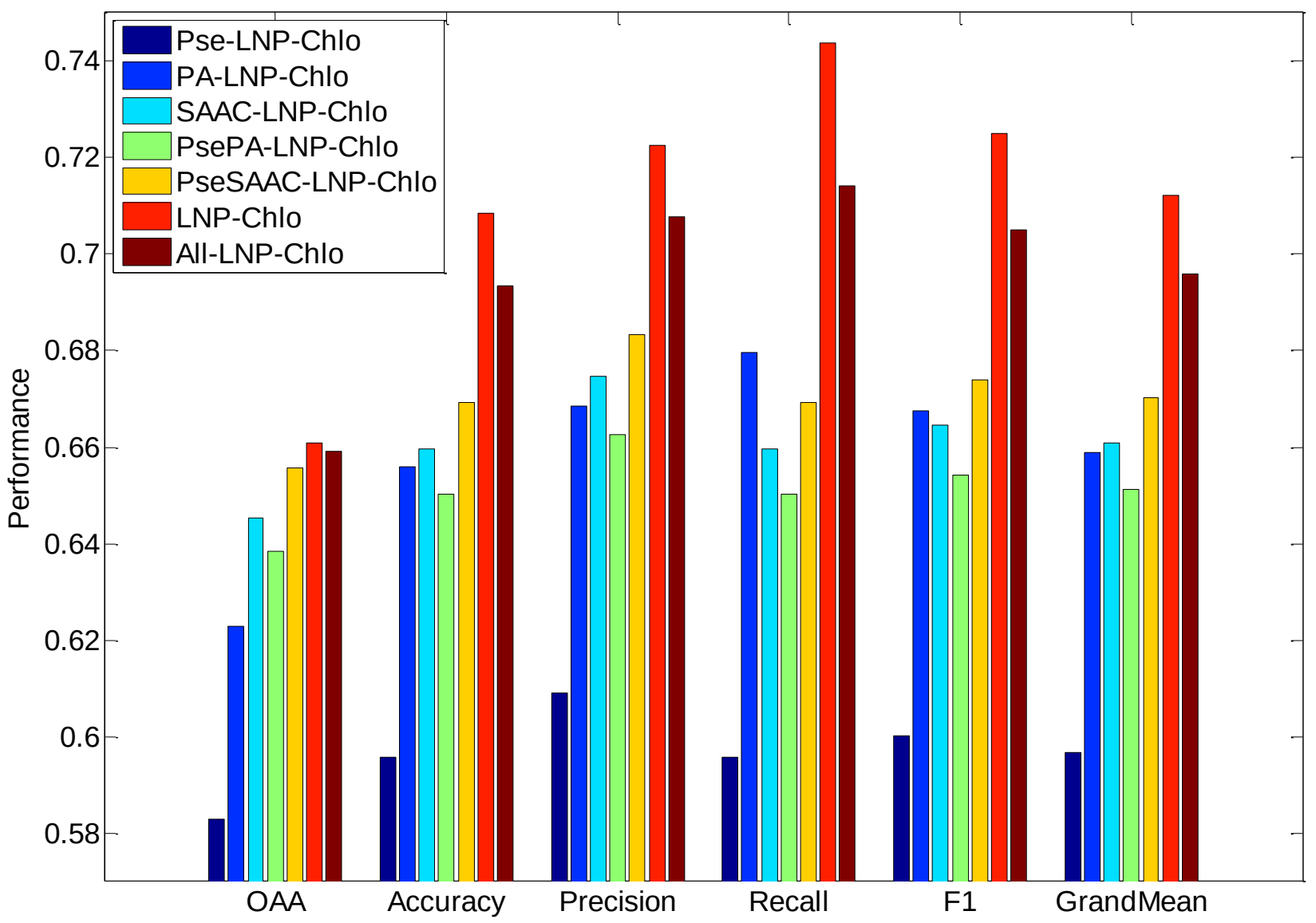
(57) McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157.


Training Sequences
Testing Sequences
SAAC
SAAC
PSI-BLAST
PSI-BLAST
SAAC features
PA features
Profile Alignment
LNP
β
1-β
LNP
Adaptive Decision
Ensemble LNP
Subchloroplast Location(s)

For TOC only


Training Sequences
Testing Sequences
SAAC
SAAC
PSI-BLAST
PSI-BLAST
SAAC features
PA features
Profile Alignment
LNP
β
1-β
LNP
Adaptive Decision
Ensemble LNP
Subchloroplast Location(s)


AL-KNN
MultiP-SChlo
EnTrans-Chlo
LNP-Chlo
Performance
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
OAA
Accuracy
Precision
Recall
F1
GrandMean


Pse-LNP-Chlo
PA-LNP-Chlo
SAAC-LNP-Chlo
PsePA-LNP-Chlo
PseSAAC-LNP-Chlo
LNP-Chlo
All-LNP-Chlo
Performance
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
OAA
Accuracy
Precision
Recall
F1
GrandMean