

Speeding up Subcellular Localization by Extracting Informative Regions of Protein Sequences for Profile Alignment

Wei Wang, Man-Wai Mak and Sun-Yuan Kung

Abstract—The functions of proteins are closely related to their subcellular locations. In the post-proteomics era, the amount of gene and protein data grows exponentially, which necessitates the prediction of subcellular localization by computational means. This paper proposes mitigating the computation burden of alignment-based approaches to subcellular localization prediction by using the information provided by the N-terminal sorting signals. To this end, a cascaded fusion of cleavage site prediction and profile alignment is proposed. Specifically, the informative segments of protein sequences are identified by a cleavage site predictor. Then, only the informative segments are applied to a homology-based classifier for predicting the subcellular locations. Experimental results on a newly constructed dataset show that the method can make use of the best property of both approaches and can attain an accuracy higher than using the full-length sequences. Moreover, the method can reduce the computation time by 20 folds. We advocate that the method will be important for biologists to conduct large-scale protein annotation or for bioinformaticians to perform preliminary investigations on new algorithms that involve pairwise alignments.

Index Terms—Subcellular localization; cleavage sites prediction; profiles alignment; protein sequences; support vector machines.

I. INTRODUCTION

A. Motivation of Subcellular Localization Prediction

Prediction of subcellular localization, which involves the computational prediction of where a protein resides in a cell, is a challenging task. Accurate prediction of subcellular locations can assist the prioritization of proteins for downstream analysis and the identification of drug targets. Because of the rapid increase in the number of sequenced genomes, it is highly desirable to develop effective prediction methods so that the newly found proteins can be effectively used in drug development. A number of approaches to solving this problem have been proposed in the literature. These methods can be generally divided into four categories, including predictions based on sorting signals [1], [2], [3], [4], [5], global sequence properties [6], [7], [8], [9], homology [10], [11], [12], and other information in addition to sequences [13], [14].

B. Approaches to Subcellular Localization Prediction

Prediction based on sorting signals determines the localization of proteins via the recognition of their N-terminal

sorting signal. These cleavable peptides contain information that allows the protein to be transported to either the secretory pathway (in which case they are called signal peptides) or to mitochondria and chloroplast (in which case they are called transit peptides). PSORT [1] and its extension WoLF PSORT [2], [3] are some of the early methods that use the N-terminal information. PSORT is a knowledge-based program for predicting protein localization, and WoLF PSORT uses the information contained in sorting signals, amino acid composition and functional motifs to convert amino acid sequences into numerical localization features. More recent predictors such as TargetP [4], [5] use Hidden Markov models and neural networks to learn the relationship between the subcellular locations and amino acid sequences.

The second group of prediction methods is based on the fact that proteins of different subcellular compartments differ in global properties, such as their amino acid composition. One of the early studies that use amino acid composition is SubLoc [6]. This method converts full-length protein sequences into 20-dim amino composition vectors for classification by support vector machines. To incorporate the information of sequence order into the global properties, amino acid composition has been extended to amino-acid pair compositions (dipeptide) [7] and gapped amino-acid pair compositions [8]. One advantage of using global sequence properties is that genomic or EST (Expressed Sequence Tag) sequences without the N-terminus can be handled. It has been found that a simple odds-ratio statistics based on amino-acid composition and residue-pair frequencies can be used to discriminate between soluble intracellular and extracellular proteins [9].

The third group of prediction methods is based on the knowledge that homologs often share the same subcellular compartment. Given a query sequence, these methods use the sequence to search against databases for homologs [10], [11] and predict its subcellular location as the one to which the homologs belong. For example, Mak et al. [12] proposed a predictor called PairProSVM in which the profile of an unknown sequence is aligned with the profile of every training sequence to form a score vector for classification by support vector machines. It was found that profile alignment is more sensitive to the weak similarity between protein families than sequence alignment.

Some predictors not only use peptide sequences as input but also require extra information such as lexical context in database entries [13] or Gene Ontology entries [14]. Although studies have shown that this type of method can outperform sequence-based methods, the performance has

Wei Wang and Man-Wai Mak are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR (email: enmwamak@polyu.edu.hk).

Sun-Yuan Kung is with the Department of Electrical Engineering, Princeton University, USA

This project was in part supported by the RGC of HKSAR project Nos. 5264/09E and 5251/08E.

only been measured on data sets where all sequences have the required additional information. Thus, the applicability is limited by the availability of the extra information.

C. Limitations of Existing Approaches

Among all these methods, the signal-based and homology-based methods have attracted a great deal of attention, primarily because of their biological plausibility and robustness in predicting newly discovered sequences. Comparing these two approaches, the signal-based methods seem to be more direct, because they determine the localization from the sequence segments that contain the localization information. However, this type of method is typically limited to the prediction of a few subcellular locations only. For example, the popular TargetP [4], [5] can only detect three localizations: chloroplast, mitochondria, and secretory pathway. The homology-based methods, on the other hands, can in theory predict as many localizations as available in the training data. The downside, however, is that the whole sequence is used for the homology search or pairwise alignment, without considering the fact that some segments of the sequence are more important or contain more information than the others. Moreover, the computation requirement will be excessive for long sequences. The problem will become intractable for database annotation where hundreds of thousands of proteins are involved.

D. Our Proposal for Addressing the Limitations

The computation burden of homology-based methods is mainly due to the alignment of the whole sequences. Because localization information is not evenly spread over the whole sequence (otherwise the signal-based method will perform poorly), potential computation saving can be achieved by aligning the portion of the sequences that contains most of the localization information. For this, the signal-based methods can provide a good solution because these methods scan the whole sequence to look for the signal peptide (i.e., informative region). In fact, the length of chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP), and secretory pathway signal peptide (SP) is under 100 amino acids only [5], as illustrated in Table I.

This paper proposes using cleavage site prediction to determine the most informative region for alignment. Experiments on a data set extracted from a recent release of Swiss-Prot show that the proposed fusion method not only improves the accuracy of subcellular localization, but also reduces the computation time by 20 folds.

The remainder of the paper is organized as follows. In Section II, the cascaded fusion of cleavage site detection and homology-based approaches for subcellular localization are described. In Sections III, we describe the experiments for analyzing the sensitivity of subcellular localization accuracy with respect to the error in cleavage site detection. In section IV, we compare the performance of localization predictors that use full-length profile alignment with the predictors that use cleaved-profile alignment. Finally, Section V presents our conclusions and outlines directions of future work.

TABLE I: Length of secretory pathway signal peptide (SP), mitochondrial targeting peptide (mTP), and chloroplast transit peptide (cTP).

| Peptide | Length (No. of Amino Acids) |
|---------|-----------------------------|
| SP | 15–30 |
| mTP | 6–85 |
| cTP | 20–100 |

II. SUBCELLULAR LOCALIZATION PREDICTION BY CLEAVAGE SITE PREDICTION AND PROFILE ALIGNMENT

Here, we describe the homology-based approaches for subcellular localization and explain how the computational burden of this approach can be alleviated by the cascaded fusion of cleavage site prediction and profile alignment.

A. Profile-Alignment SVM

Kernel techniques based on profile alignment have been used successfully in detecting remote homologous proteins [15] and in predicting subcellular locations of eukaryotic protein [12]. Instead of extracting feature vectors directly from sequences, this method trains an SVM classifier by using the scores of local profile alignment. A profile is a matrix in which elements in a column specify the frequency of that amino acid appears in the corresponding position. Practically, the profile of a sequence can be obtained by using the sequence as a seed to search against a protein database (e.g., Swiss-Prot) for homologous sequences using the PSI-BLAST program [16]. The homolog information pertaining to the aligned sequences are represented by two matrices (profiles): position-specific scoring matrix (PSSM) and position-specific frequency matrix (PSFM). Each entry of a PSSM represents the log-likelihood of the residue substitutions at the corresponding positions in the query sequence. The PSFM contains the weighted observation frequencies of each position of the aligned sequences.

Let us denote the operation of PSI-BLAST search given the query sequence $S^{(i)}$ of length n_i as,

$$\phi^{(i)} \equiv \phi(S^{(i)}) : S^{(i)} \rightarrow \{P^{(i)}, Q^{(i)}\} \quad (1)$$

where $P^{(i)}$ and $Q^{(i)}$ are the PSSM and PSFM of $S^{(i)}$, respectively. Using the profile alignment algorithm specified in [12], we obtain the profile alignment scores $\rho(\phi(S^{(i)}), \phi(S^{(j)}))$. Then, the following normalized alignment scores are obtained:

$$\zeta(\phi^{(i)}, \phi^{(j)}) = \frac{\rho(\phi(S^{(i)}), \phi(S^{(j)}))}{\sqrt{\rho(\phi(S^{(i)}), \phi(S^{(i)}))\rho(\phi(S^{(j)}), \phi(S^{(j)}))}}. \quad (2)$$

Given N training sequences, the scores $\{\zeta(\phi^{(i)}, \phi^{(j)})\}_{i,j=1}^N$ constitute a symmetric matrix Z whose columns can be considered as N -dimensional vectors:

$$\zeta^{(j)} = [\zeta(\phi^{(1)}, \phi^{(j)}) \quad \dots \quad \zeta(\phi^{(N)}, \phi^{(j)})]^T \quad j = 1, \dots, N. \quad (3)$$

This means that there are N feature vectors with dimension equal to the training set size. The N N -dimensional column

vectors can be used to train M one-vs-rest SVMs for an M -class protein prediction problem:

$$f_m(S) = \sum_{j \in \mathcal{S}_m} y_{m,j} \alpha_{m,j} K(\phi(S), \phi(S^{(j)})) + b_m \quad (4)$$

$$m = 1, \dots, M,$$

where S is an unknown sequence, $y_{m,j} \in \{+1, -1\}$, \mathcal{S}_m contains the indexes of support vectors, $\alpha_{m,j}$ are Lagrange multipliers, and $K(\phi(S), \phi(S^{(j)}))$ is a kernel function. When $K(\cdot)$ is a linear kernel, we have

$$K(\phi(S), \phi(S^{(j)})) = \langle \zeta, \zeta^{(j)} \rangle$$

$$= \sum_{n=1}^N \zeta(\phi(S^{(n)}, \phi(S))) \zeta(\phi(S^{(n)}, \phi(S^{(j)})). \quad (5)$$

During prediction, the class of an unknown sequence S can be obtained by

$$y(S) = \arg \max_{m=1}^M f_m(S). \quad (6)$$

B. Protein Cleavage-Site Prediction

1) *TargetP, SignalP, and ChloroP*: TargetP [4], [5] is one of the most popular signal-based subcellular localization predictors and cleavage site predictors. Given a query sequence, TargetP can determine its subcellular localization and will also invoke SignalP [17], ChloroP [18], or a program specialized for mTP to determine the cleavage site of the sequence. TargetP requires the N-terminal sequence of a protein as input. During prediction, a sliding window scans over a query sequence; for each segment within the window, a numerically encoded vector is presented to a neural network to compute the Y-score of the segment. The cleavage site is determined by finding the position at which the Y-score is maximum. The cleavage site prediction accuracy of SignalP on Eukaryotic proteins is around 70% [19] and that of ChloroP on cTP is 60% (± 2 residues) [18], suggesting that there is room for improvement.

2) *CRF-based Predictors*: Conditional random fields (CRFs) were originally designed for sequence labelling tasks such as Part-of-Speech (POS) tagging. Given a sequence of observations, a CRF predictor finds the most likely label for each of the observations. CRFs have a graphical structure consisting of edges and vertices in which an edge represents the dependency between two random variables (e.g., two amino acids in a protein) and a vertex represents a random variable whose distribution is to be inferred. Therefore, CRFs are undirected graphical models, as opposed to directed graphical models such as HMMs. Also, unlike HMMs, the distribution of each vertex in the graph is conditioned on the whole input sequence.

To use CRFs for cleavage site prediction, the prediction problem is formulated as a sequence labelling task in which amino acid sequences are treated as observations and each amino acid in the sequences is labelled as either ‘‘Signal’’, ‘‘Cleavage’’, or ‘‘Mature’’, e.g., SSSSSCMMMMMM. The cleavage site is located at the transition between C and M.

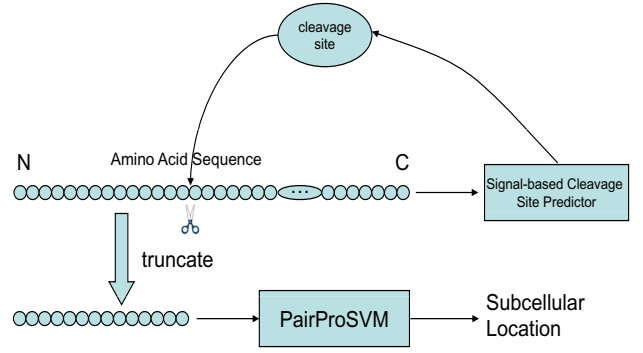


Fig. 1: Cascaded fusion of signal-based and homology-based methods. The signal-based cleavage site predictor, such as TargetP [5] and CSitePred [20], is used as a pre-processor that reduces the sequence length for the computationally expensive homology-based method such as PairProSVM [12].

Since accurate cleavage site prediction is important for the prediction of subcellular localization, we have recently proposed a cleavage site predictor (called CSitePred [20]) that uses conditional random fields (CRFs) [21] and demonstrated that CSitePred outperforms TargetP in predicting the cleavage sites of signal peptides (SP) [22], [23]. This finding has motivated us to use CRFs to find the cleavage sites of mTP and cTP, in addition to SP, in this work. Results of these predictors will be reported in Section IV.

C. Combining Cleavage Site Detection and Profile Alignment

The computation burden of homology-based methods is mainly due to the alignment of the whole sequences. Generally, the length of signal peptide is less than 100 amino acids. Given the fact that the majority of proteins in the Swiss-Prot database have about a few hundred amino acids and that some proteins could have length longer than 5,000 amino acids, tremendous computational saving can be achieved by aligning the pre-sequence region (from the N-terminus to the cleavage site) for those proteins containing a signal or targeting sequence. For profile alignment, this amounts to aligning the pre-profile region, i.e., the PSSM and PSFM in Section II-A are truncated at the column corresponding to the cleavage site before carrying out profile alignment.

The above observation suggests that the computation burden can be largely alleviated by a cascaded fusion of signal-based and homology-based methods. The fusion has three steps (Fig. 1):

- 1) *Cleavage site detection*. The cleavage site (if any) of a query sequence is determined by a signal-based method.
- 2) *Pre-profile selection*. The pre-profile of the query is obtained by selecting from the N-terminus up to the cleavage site.

TABLE II: Breakdown of the eukaryotic dataset used in this work. The data were extracted from Swiss-Prot Release 57.5, with sequence identity less than 25%.

| Subcellular Location | No. of Sequences |
|----------------------|------------------|
| Extracellular | 693 |
| Mitochondria | 167 |
| Chloroplast | 74 |
| Cytoplasm/Nucleus | 1,617 |
| All | 2,552 |

- 3) *Pairwise alignment.* The pre-profile is aligned with each of the training pre-profiles to form an N -dim vector, which is fed to a one-vs-rest SVM classifier for prediction.

During the training phase, N training pre-profiles are obtained by truncating at the columns corresponding to the cleavage sites. Pairwise alignments are then performed to create an $N \times N$ symmetric score matrix whose column vectors are used to train a one-vs-rest SVM classifier.

III. EXPERIMENTS

A. Data Set Construction

Protein sequences with experimentally annotated subcellular locations were collected from Swiss-Prot Release 57.5 according to the following criteria.

- 1) Only the proteins of eukaryotic species are included. In Swiss-Prot, these sequences are annotated with “Eukaryota” in the OC (Organism Classification) field.
- 2) A large number of sequences in Swiss-Prot are annotated with ambiguous words, such as “probable”, “by similarity” and “potential”. These entries were excluded because they lack experimental evidence.
- 3) Sequences annotated with “fragment” were excluded.

The extracted sequences were then further filtered by BLASTclust [24] to produce a dataset with sequence identity less than 25%.

Sequence quality is of primary importance for the development of good prediction methods. To this end, all training sequences should have experimental evidence and should not be inferred by similarity or existing prediction methods. Otherwise, it can lead to circular prediction in which methods reproduce each other’s predictions. For this reason we built a non-redundant dataset comprising proteins sharing less than 25% sequence identity. Table II shows the breakdown of the number of sequences in each class.

Sprenger et al. [25] compare the performance of five subcellular localization methods that are capable of predicting at least nine locations. It was concluded that none of the five methods had a sufficient level of sensitivity that would allow reliable prediction of hypothetical proteins. Therefore, we consider four subcellular compartments shown in Table II: Extracellular, mitochondria, chloroplast and others (including cytoplasm and nucleus). We decided not to predict other locations because the number of annotated proteins with less

than 25% sequence identity is very small, which do not allow us to train a predictor with good generalization capability.

B. Experiments for Performance Evaluation

1) *Effect of Incorrect Cleavage Site Prediction:* To evaluate the effect of incorrect cleavage site prediction on the accuracy of subcellular localization, sensitivity analysis was performed by using the N-terminal signal peptides cleaved at the ground-truth cleavage sites or plus/minus several positions of the ground-truths. The sequence cut-off positions are 16, 8, 2 amino acids upstream or 2, 16, 32, 64 amino acids downstream from the ground-truth cleavage site. For comparison, another experiment was done in which the cleaved-off position was set to 170, i.e., none of the sequences (or profiles) have length exceeding 170.

2) *Performance of Cleavage Site Predictors:* To assess the performance of different cleavage site predictors, TargetP and CSitePred (a CRF-based predictor [22], [20]) were compared for the prediction accuracy of the cleavage site for SP, mTP and cTP. During prediction, the subcellular locations of the test sequences were assumed to be unknown. For TargetP, the subcellular location of a test sequence was first determined by presenting the sequence to TargetP using either the ‘Plant’ or ‘Non-plant’ option of the predictor. Based on the subcellular location, TargetP will then determine the cleavage site of the sequence by invoking SignalP, ChloroP (for plant), or a program specialized in predicting the cleavage sites of mTP. For CSitePred, given a query sequence, the CRF (corresponding to either SP, mTP, or cTP) with the maximum Viterbi-search score is first identified. Then, the cleavage site is obtained from the optimal Viterbi search path of this maximum-scoring CRF.

3) *Performance of Subcellular Localization:* The performance of subcellular localization prediction by the proposed cascaded fusion method was evaluated and compared with two state-of-the-art subcellular localization predictors: SubLoc [6] and TargetP. The performance of SubLoc and TargetP were obtained by presenting the sequences of the dataset to their webserver. We used TargetP and CSitePred for cleavage site detection and used PairProSVM [12] for classification of pre-profiles. Sequences with cleavage site include extracellular, mitochondria, and chloroplast. Sequences without a cleavage site include cytoplasm and nucleus. We measured the time taken to create a $2,552 \times 2,552$ alignment-score matrix and the time taken to perform 5-fold cross validation on the score matrix on an Intel Core 2 Duo 3.16 GHz CPU. For PSI-BLAST, parameters h and j were set to 0.001 and 3, respectively. The Spider Toolbox¹ was used to implement the SVM classifiers, and CRF++² was used for implementing CSitePred.

C. Assessment of Prediction Performance

We used 5-fold cross validation to evaluate the performance. In this technique, the original dataset was divided

¹<http://www.kyb.mpg.de/bs/people/spider/>

²<http://crfpp.sourceforge.net/>

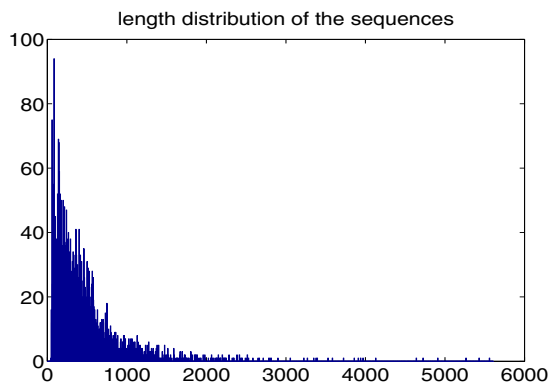


Fig. 2: The histogram of the length of the sequences in our dataset. Vertical axis: number of occurrences; horizontal axis: sequence length.

randomly into 5 sets consisting of nearly equal number of sequences. In each fold, one subset was singled out as a testing set, and the remaining ones were merged as the training set; this process was repeated five times.

The overall prediction accuracy, the accuracy for each subcellular location, and the Matthew’s correlation coefficient (MCC) [26] were used to quantify the prediction performance. MCC allows us to overcome the shortcoming of accuracy (Acc) on unbalanced data [26].

IV. RESULTS AND DISCUSSIONS

A. Histograms of Sequence Length

As shown in Fig. 2, the majority of proteins in the datasets have a few hundred amino acids. The average sequence length is 469 amino acids and some proteins have length up to 5,560 amino acids. Fig. 3 shows the histograms of the length of signal peptides, mitochondrial transit peptides, and chloroplast transit peptides. It is obvious that the lengths of the three types of peptides are rather short (ranging from 6 to 100), with cTP longer than mTP and SP on average. The length distribution of SP has a relatively narrow peak, whereas that of the cTP and mTP spread over a wider range.

B. Sensitivity Analysis

When sequences were cut at the ground-truth cleavage sites (denoted as “ p ” in Table III), the overall accuracy reaches 98.47%. The prediction accuracy for Ext and Cyt/Nuc is above 98%. Despite the relatively weak signal, 80% of chloroplasts were correctly predicted. It is obvious that the localization performance degrades when the cut-off position drifts away from the ground-truth cleavage site. But the overall accuracy can be maintained at above 95% even if the drift is as large as -16 and $+64$ positions from the ground-truth.

Table III also shows that mTP and cTP are more sensitive to the error of cleavage site prediction, which agrees with the fact that the signals of mTP and cTP are weaker. For comparison, another experiment was done in which the cleaved-off

TABLE III: Sensitivity of subcellular localization accuracy with respect to the profile cut-off positions. p is the ground-truth cleavage site. For “Cyt/Nuc” proteins, p is set to 170.

| Seq. Cutoff Position | Accuracy of Individual Class (%) | | | | Overall Accuracy(%) |
|----------------------|----------------------------------|-------|-------|---------|---------------------|
| | Ext | Mit | Chl | Cyt/Nuc | |
| $p - 16$ | 94.95 | 51.50 | 74.32 | 100 | 94.71 |
| $p - 8$ | 98.56 | 86.23 | 77.03 | 99.94 | 98.00 |
| $p - 2$ | 98.70 | 85.63 | 79.73 | 99.94 | 98.08 |
| p | 98.85 | 90.42 | 82.43 | 99.88 | 98.47 |
| $p + 2$ | 98.99 | 88.62 | 85.14 | 99.88 | 98.47 |
| $p + 16$ | 99.28 | 88.62 | 70.27 | 99.69 | 98.00 |
| $p + 32$ | 99.28 | 86.83 | 64.86 | 99.51 | 97.61 |
| $p + 64$ | 98.99 | 77.25 | 54.05 | 99.01 | 96.28 |
| Fix-length(170) | 91.92 | 53.89 | 28.38 | 97.28 | 90.98 |

TABLE IV: Cleavage-site prediction accuracies achieved by TargetP and CSitePred. For TargetP, (P) and (N) means using the ‘Plant’ and ‘Non-plant’ option of the predictor, respectively. TargetP will invoke SignalP, ChloroP, or a program specialized in predicting mTP for cleavage site prediction. CSitePred is based on conditional random fields.

| Cleavage Site Predictor | Cleavage Site Prediction Accuracy (%) | | | |
|-------------------------|---------------------------------------|-------|-------|---------|
| | SP | mTP | cTP | Overall |
| TargetP(P) | 64.55 | 44.04 | 8.82 | 56.48 |
| TargetP(N) | 75.28 | 46.69 | 2.21 | 64.38 |
| CSitePred | 71.81 | 39.74 | 31.62 | 62.89 |

position was set to 170, i.e., none of the sequences (or profiles) have length exceeding 170. The prediction performance using fixed-length pre-profile alignment is shown in the last row of Table III. It is obvious that cutting the profiles at the cleavage sites can achieve a higher accuracy than cutting them at a fixed position.

C. Performance of Cleavage Site Prediction

As demonstrated in the Section IV-B, the accuracy of subcellular localization depends on the positions at which the protein sequences are cut. Therefore, it is imperative to find a good cleavage site predictor, especially for mTP and cTP.

Table IV shows the cleavage site prediction accuracy of TargetP and CSitePred (a CRF-based predictor). Table IV shows that CSitePred is better than TargetP(P) in terms of predicting the cleavage sites of signal peptide (Ext) but is worse than TargetP(N). The results also suggest that while CSitePred is slightly inferior to TargetP in predicting the cleavage sites of mitochondria, it is significantly better than TargetP in predicting the cleavage site of chloroplasts. Note that the overall accuracies depend heavily on the Ext class because of the large number of signal peptides in the dataset (see Table II).

Note that the prediction accuracy of chloroplasts in our experiments is significantly lower than that of [18]. There are two reasons for this difference: (1) our dataset has sequence identity lower than that of [18] and (2) we consider the prediction of the exact ground-truth position as a correct

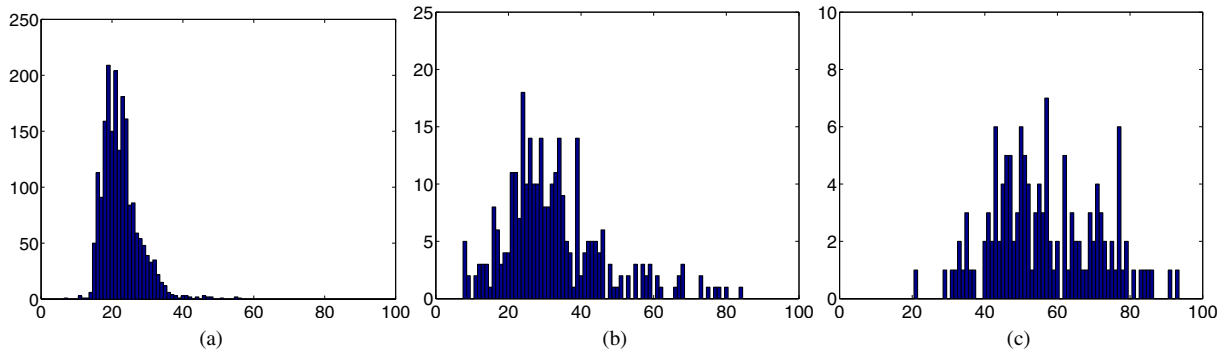


Fig. 3: The histograms of length of (a) secretory pathway signal peptides, (b) mitochondrial targeting peptides, and (c) chloroplast transit peptides. Vertical axis: number of occurrences; horizontal axis: sequence length.

prediction whereas [18] consider a prediction within ± 2 of the ground-truth as a correct prediction. In fact, if we relaxed the criterion of correct prediction to ± 2 ground-truth positions, the prediction accuracy on chloroplasts achieved by TargetP increases to 47.06%.

D. Performance of Cascaded Fusion

Table V shows that the computation time for full-length profile alignment is a striking 116 hours, which suggests that full-length alignment is computationally prohibitive for most practical applications. Therefore, it is imperative to limit the length of the sequences or profiles before alignment. Our method not only leads to nearly a 20 folds reduction in computation time but also boosts the prediction performance considerably. This is because the signal segment can be found in the N-terminus, and removing the amino acids beyond the cleavage site helps the alignment focus on the relevant features in the sequences and disregard noise.

E. Compared with State-of-the-Art Predictors

Table VI shows that the overall accuracy of the proposed method (last row) is 9.6% higher than that of TargetP (3rd row) and is significantly better than that of SubLoc (1st row). One limitation of TargetP is that users need to select either “Plant” or “Non-plant”. If the former is selected, the performance of Ext and Cyt/Nuc degrade significantly, leading to a low overall accuracy; if the latter is selected, none of the chloroplast proteins can be correctly predicted. The cascaded fusion of cleavage site prediction and PairProSVM, on the other hand, can classify all four classes with fairly high accuracy, leading to a higher overall accuracy.

Because ChloroP is weak in predicting the cleavage sites of chloroplasts (see Table IV), it is not a good candidate for assisting PairProSVM. This is evident by the low subcellular localization accuracy of chloroplasts in Table VI when TargetP is used as a cleavage site predictor. However, TargetP is fairly good at predicting the subcellular location of chloroplasts when it is used as a localization predictor.

Among the four classes in Table VI, the subcellular localization accuracies of mitochondria and chloroplasts are generally lower than the extracellular (secretory SP). The

reason may be that these transit peptides are less well characterized and their motifs are less conserved than those of secretory SP [5].

Table VI also suggests that the CRF-based cleavage site predictor is very effective in assisting PairProSVM, leading to the highest prediction accuracy (97.61%) among all subcellular localization predictors. In particular, CSitePred can help PairProSVM to increase the subcellular localization accuracy of chloroplasts from 58% to 63%.

V. CONCLUSIONS

We proposed a novel subcellular-localization-prediction method that is based on the cascaded fusion of signal-based and homology-based methods. Through five-fold cross validation tests on a newly created redundancy-removed data set, we obtained an overall accuracy of 97.61% and an average MCC of 0.97. These values are higher than TargetP and SubLoc – methods based on sorting signals or amino acid composition. We believe that the high accuracy attained by our method indicates that our method can efficiently capture the subtle patterns in signal sequences. The profiles which are computed from multiple sequence alignment can provide evolutionary information of sorting signals. Proteins’ profiles are calculated by searching the Swiss-Prot database using PSI-BLAST. Then the scores of local pairwise profile alignment are computed, which in turn are used to construct the kernel of an SVM classifier. Moreover, the computational burden is greatly alleviated by excluding the uninformative regions in profile alignment. We hope that this in-silico method can be complementary to experimental subcellular localization techniques.

REFERENCES

- [1] K. Nakai and M. Kanehisa, “A knowledge base for predicting protein localization sites in eukaryotic cells,” *Genomics*, vol. 14, pp. 897–911, 1992.
- [2] P. Horton, K. J. Park, T. Obayashi, and K. Nakai, “Protein subcellular localization prediction with WoLF PSORT,” in *Proc. 4th Annual Asia Pacific Bioinformatics Conference (APBC06)*, 2006, pp. 39–48.

TABLE V: Subcellular localization accuracy and computation time for different cut-off positions for sequences with and without cleavage sites. Computation time for alignment is the time taken to create a profile-alignment score matrix. Computation time for classification is the time taken to perform 5-fold cross validation on the score matrix. In the first column, “Full length” means there is no cutoff for sequences, i.e., the whole sequences will be directly processed by PairProSVM. “TargetP(P)” and “TargetP(N)” mean that the cutoff position is determined by TargetP using the “Plant” option and “Non-plant” option, respectively. CSitePred is a cleavage site predictor based on conditional random fields.

| Seq. Cutoff Position | Computation Time | | Accuracy of each Sequence Class (%) | | | | Overall Accuracy(%) |
|--------------------------|------------------|----------------------|-------------------------------------|-------|-------|---------|---------------------|
| | Alignment (hr.) | Classification (hr.) | Ext | Mit | Chl | Cyt/Nuc | |
| Full length | 115.83 | 0.20 | 95.15 | 51.94 | 32.22 | 97.14 | 91.64 |
| 170 | 15.69 | 0.19 | 91.92 | 53.89 | 28.38 | 97.28 | 90.98 |
| Ground-truth | 6.47 | 0.13 | 99.28 | 90.29 | 90.00 | 99.89 | 98.77 |
| Determined by TargetP(P) | 6.23 | 0.19 | 90.48 | 71.86 | 58.11 | 95.98 | 89.08 |
| Determined by TargetP(N) | 5.85 | 0.19 | 97.11 | 69.46 | 41.89 | 96.23 | 93.14 |
| Determined by CSitePred | 6.36 | 0.13 | 98.85 | 86.23 | 63.51 | 99.81 | 97.61 |

TABLE VI: Subcellular localization performance achieved by different combinations of cleavage site predictors and localization predictors. The first column specifies the the cleavage site predictor (if any) in the cascaded fusion. Notice that TargetP can perform both cleavage site prediction and subcellular localization. In the cascaded fusion of TargetP and PairProSVM (the 4th and 5th row), we only used the cleavage site prediction of TargetP. “TargetP(P)” and “TargetP(N)” mean selecting “plant” or “non-plant” option in TargetP. CSitePred is a cleavage site predictor based on conditional random fields. The dataset with sequence identity less than 25% was used in the experiments.

| Cleavage Site Predictor | Localization Predictor | Subcell Localization Accuracy (%) | | | | | Subcell Localization MCC | | | | |
|-------------------------|------------------------|-----------------------------------|--------------|--------------|---------|--------------|--------------------------|-------------|-------------|---------|-------------|
| | | Ext | Mit | Chl | Cyt/Nuc | Overall | Ext | Mit | Chl | Cyt/Nuc | Overall |
| — | SubLoc [6] | 51.44 | 55.83 | — | 77.86 | 66.79 | — | — | — | — | — |
| — | TargetP (P) | 79.08 | 88.02 | 89.19 | 69.57 | 73.93 | 0.79 | 0.49 | 0.79 | 0.64 | 0.65 |
| — | TargetP (N) | 97.40 | 89.22 | 0.00 | 87.82 | 87.97 | 0.93 | 0.58 | 0.00 | 0.81 | 0.84 |
| TargetP(P) | PairProSVM | 90.48 | 71.86 | 58.11 | 95.98 | 89.08 | 0.88 | 0.75 | 0.66 | 0.84 | 0.89 |
| TargetP(N) | PairProSVM | 97.11 | 69.46 | 41.89 | 96.23 | 93.14 | 0.94 | 0.72 | 0.57 | 0.87 | 0.91 |
| CSitePred | PairProSVM | 98.85 | 86.23 | 63.51 | 99.81 | 97.61 | 0.98 | 0.89 | 0.76 | 0.97 | 0.97 |

- [3] P. Horton, K.J. Park, T. Obayashi, N. Fujita, H. Harada, C.J Adams-Collier, and K. Nakai, “WolF PSORT: protein localization predictor,” *Nucleic acids research*, vol. 35, no. Web Server issue, pp. 585–587, 2007.
- [4] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, “Predicting subcellular localization of proteins based on their N-terminal amino acid sequence,” *J. Mol. Biol.*, vol. 300, no. 4, pp. 1005–1016, 2000.
- [5] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, “Locating proteins in the cell using TargetP, SignalP, and related tools,” *Nature Protocols*, vol. 2, no. 4, pp. 953–971, 2007.
- [6] S. J. Hua and Z. R. Sun, “Support vector machine approach for protein subcellular localization prediction,” *Bioinformatics*, vol. 17, pp. 721–728, 2001.
- [7] Y. Huang and Y. D. Li, “Prediction of protein subcellular locations using fuzzy K-NN method,” *Bioinformatics*, vol. 20, no. 1, pp. 21–28, 2004.
- [8] K. J. Park and M. Kanehisa, “Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs,” *Bioinformatics*, vol. 19, no. 13, pp. 1656–1663, 2003.
- [9] H. Nakashima and K. Nishikawa, “Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies,” *J. Mol. Biol.*, vol. 238, pp. 54–61, 1994.
- [10] R. Mott, J. Schultz, P. Bork, and C.P. Ponting, “Predicting protein cellular localization using a domain projection method,” *Genome research*, vol. 12, no. 8, pp. 1168–1174, 2002.
- [11] M.S. Scott, D.Y. Thomas, and M.T. Hallett, “Predicting subcellular localization via protein motif co-occurrence,” *Genome research*, vol. 14, no. 10a, pp. 1957–1966, 2004.
- [12] M. W. Mak, J. Guo, and S. Y. Kung, “PairProSVM: Protein subcellular localization based on local pairwise profile alignment and SVM,” *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, vol. 5, no. 3, pp. 416 – 422, 2008.
- [13] R. Nair and B. Rost, “Sequence conserved for subcellular localization,” *Protein Science*, vol. 11, pp. 2836–2847, 2002.
- [14] K.C. Chou and H.B. Shen, “Recent progress in protein subcellular location prediction,” *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.
- [15] H. Rangwala and G. Karypis, “Profile-based direct kernels for remote homology detection and fold recognition,” *Bioinformatics*, vol. 21, no. 23, pp. 4239–4247, 2005.
- [16] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: A new generation of protein database search programs,” *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, 1997.
- [17] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, “Improved prediction of signal peptides: SignalP 3.0,” *J. Mol. Biol.*, vol. 340, pp. 783–795, 2004.
- [18] O. Emanuelsson, H. Nielsen, and G. von Heijne, “ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites,” *Protein Science*, vol. 8, pp. 978–984, 1999.
- [19] H. Nielsen, S. Brunak, and G. von Heijne, “Machine learning approaches for the prediction of signal peptides and other protein sorting signals,” *Protein Eng.*, vol. 12, no. 1, pp. 3–9, 1999.
- [20] <http://158.132.148.85:8080/CSitePred/faces/Page1.jsp>.
- [21] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in

Proc. 18th Int. Conf. on Machine Learning, 2001.

- [22] M. W. Mak and S. Y. Kung, "Conditional random fields for the prediction of signal peptide cleavage sites," in *Proc. ICASSP*, Taipei, April 2009, pp. 1605–1608.
- [23] M. W. Mak, W. Wang, and S. Y. Kung, "Fusion of conditional random field and SignalP for protein cleavage site prediction," in *Proc. APSIPA'09*, Supporo, Oct. 2009, pp. 716–721.
- [24] <http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html>.
- [25] J. Sprenger, J.L. Fink, and R. Teasdale, "Evaluation and comparison of mammalian subcellular localization prediction methods," *BMC bioinformatics*, vol. 7, no. Suppl 5, pp. S3, 2006.
- [26] B. W. Matthews, "Comparison of predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta*, vol. 405, pp. 442–451, 1975.