Sequence driven features for prediction of subcellular localization of proteins

Jong Kyoung Kim Sung-Yang Bang Seungjin Choi

Department of Computer Science

Pohang University of Science and Technology

San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea

Email: blkimjk@postech.ac.kr, sybang@postech.ac.kr, seungjin@postech.ac.kr

ABSTRACT: Predicting the cellular location of an unknown protein gives a valuable information for inferring the possible function of the protein. For more accurate prediction system, we need a good feature extraction method that transforms the raw sequence data into the numerical feature vector, minimizing information loss. In this paper, we propose new methods of extracting underlying features only from the sequence data by computing pairwise sequence alignment scores. In addition, we use composition based features to improve prediction accuracy. To construct an SVM ensemble from separately trained SVM classifiers, we propose specificity based weighted majority voting. The overall prediction accuracy evaluated by the 5-fold cross-validation reached 88.53% for the eukaryotic animal data set. By comparing the prediction accuracy of various feature extraction methods, we could get the biological insight on the location of targeting information. Our numerical experiments confirm that our new feature extraction methods are very useful for predicting subcellular localization of proteins.

1 INTRODUCTION

In a eukaryotic animal cell, nuclear-encoded proteins are synthesized by ribosomes in the cytosol, and delivered to their proper cellular organelles for the co-operational execution of a common biological function. The delivery of a newly synthesized protein in the cytosol to its correct location is referred to as protein sorting or subcellular localization. The major protein sorting processes can be divided into secretory and non-secretory pathway. In the secretory pathway, all proteins are delivered to the ER as a first step, and directed to their final destinations. All proteins that contain no ER signal sequences are delivered through the non-secretory pathway. The targeting information of proteins directing them to their correct cellular destinations is stored either in the signal sequences or in the form of post-translational modifications. The proteins directed to the ER and the mitochondrion have an N-terminal signal sequence. For targeting to the peroxisome, proteins have a signal sequence which is located at the N-terminus or C-terminus. The signal sequences directing to nuclear can be present anywhere in the protein. In the secretory pathway, proteins are sorted to their final locations by several targeting features such as signal sequences, topogenic sequences, and post-translational modifications. The location of these features in the protein sequence cannot be restricted to the subsequences [1].

Predicting the cellular location of an unknown protein gives a valuable information for inferring the possible function of the protein. To achieve a good prediction result, we need effective feature extraction methods that transform the raw sequence data into the numerical feature vector, minimizing information loss. Most of prediction methods can be divided into two classes, depending on their ways of feature extraction: (1) features based on protein sequence data; (2) features based on ontology data. In the protein sequence based approach, two different methods are popular. These involve the recognition of N-terminal sorting signals or the detection of amino acids composition from an entire sequence. The former has the strong biological implication since proteins delivered to the ER, the mitochondrion, or the peroxisome (partially) have an N-terminal signal sequence [2]. However, it is difficult to recognize underlying features from a highly diverse signal sequence and to vectorize them. The latter approach partially overcomes these difficulties but lose the context information stored in the sequence data [3]. The ontology-based approach has received mush attention recently because of its high prediction accuracy [4]. This approach extracts text information of homologous sequences of a query sequence by searching biological databases and vectorize the extracted information. It is not surprising for this approach to show good performance because it utilized various extra information derived from several sources. In addition, it cannot give biological insight and interpretation on factors specifying cellular locations of proteins. Although much work has been done on improving the prediction accuracy of subcellular localization, little research has been conducted on feature extraction methods relying solely on amino acid sequence properties.

In this paper, we propose new methods of extracting underlying features only from the sequence data to predict subcellular localization. To this end, we introduce various pairwise sequence alignment methods so that a protein sequence is represented as a numerical vector of pairwise sequence alignment scores. Additionally, we use amino acids composition based features to improve prediction accuracy. For classification, we use a SVM ensemble to combine mixed type of features. Our numerical experiments confirm that our proposed methods considerably improve the prediction accuracy and give biological insight into the position of targeting information in the protein sequence.

2 FEATURE EXTRACTION

Recent studies of the feature extraction methods based on amino acid sequence properties have tended to center around amino acid composition. Although amino acid composition and subcellular localization are related, composition based methods have critical limitations in terms of its discriminative power and location coverage. We proposed, in our previous work, a new feature extraction method representing a protein sequence as the scores of dynamic global sequence alignment [5]. Despite its very high prediction accuracy, its time complexity was relatively higher than that of composition based method. Additionally, its location coverage was limited to some proteins whose signal sequences are located at the N-terminus. To overcome these limitations, we present three different methods extracting features from signal sequences in the N-terminus, or anywhere in the sequence. We also use two composition based methods to improve the prediction accuracy.

2.1 Preprocessing

In our previous work, we converted a protein sequence into the corresponding feature vector by computing the scores of the Needleman-Wunsch algorithm between the sequence and all sequences in the training set [6]. Therefore, if the size of the training set is very large, the vectorization step takes too much time. We select, in this study, representative sequences in the training set in order to decrease the time complexity. For this purpose, we cluster all sequences in the training set by using a constructed phylogenetic tree. The overall framework clustering the training sequences for each class can be described as follows. First, all sequences are truncated after first 40 residues to consider only the N-terminus. We did not use the entire sequence because it makes very long average distance between pairs of sequences within each cluster. Second, we calculate the Jukes-Cantor distance between each pair of sequences. The Jukes-Cantor distance, which is the maximum likelihood estimate of the number of substitutions between two sequences, can be calculated after aligning all pairs of sequences by using the Needleman-Wunsch algorithm. In the next step, we construct the phylogenetic tree with the UPGMA clustering method, which stands for unweighted pair group method using arithmetic averages [7]. After constructing the tree, we can make clusters by selecting some parent nodes because all the leaf nodes sharing the parent node can be considered as a cluster. We select the parent nodes by considering the distance between leaf nodes and the parent node, and the number of leaf nodes. If the distance exceeds the maximum allowable distance, which is 3.0, the child node is considered. Also, if the number of leaf nodes is too large or small, we consider the child node of the node. The maximum and minimum number of leaf nodes are different for each class, in accordance with the class size.

2.2 N-terminal profile Hidden Markov Model

After the above preprocessing step, we have clusters of sequences for each class. The natural choice is to build profile Hidden Markov models (HMMs) which are well suited to statistically model patterns in multiply aligned sequences [7]. The constructed profile HMMs for each cluster represent the N-terminal sequence families. Before building the profile HMMs, we did multiple sequence alignment of the N-terminal truncated sequences for each cluster by using CLUSTALX 1.83 [8]. Then, we built the profile HMMs with HMMer 2.3.1 [9]. To align a protein sequence with the profile HMMs, we should consider only the first 40 residues of the protein sequence. The truncated sequence is then converted into the corresponding feature vector by computing the log-odds scores between the sequence and all the constructed profile HMMs. A d-dimensional feature vector $\mathbf{x}_k$ for the kth protein sequence has the form

$$\mathbf{x}_k = [x_{k1}, x_{k2}, \cdots, x_{kd}]^T, \quad (1)$$

where T denotes the matrix or vector transpose operator and x_{ki} is the log-odds score between sequence k and the ith profile HMM. Note that d is equal to the total number of constructed profile HMMs.

2.3 N-terminal global pairwise sequence alignment

Instead of using all the sequences in the training set, selecting representative sequences reduces the time complexity of calculating the scores of pairwise sequence alignment. For this purpose, we randomly select a protein sequence from each cluster. The minimum length of the chosen sequences is 80 and the first residue should be methionine, which means the first synthesized residue from the start codon. Since we want to compare the N-terminal region, a protein sequence is truncated after first 80 residues. The processed sequence is then converted into the corresponding feature vector by computing the scores of the Needleman-Wunsch algorithm between the sequence and all the selected representative sequences. The gap penalty is -8 and the substitution matrix is BLOSUM 50.

2.4 Full sequence local pairwise sequence alignment

So far we have assumed that the signal sequences are located at the N-terminal, and that we are looking for the global match between two N-terminal regions. A much more common situation is that the signal sequences or the targeting information are located anywhere in the protein. In this situation, the most sensitive way to detect the internal targeting signals is to use the algorithm for finding optimal local alignments, which are referred to as the Smith-Waterman algorithm [10]. The general procedures of this feature extraction method is almost same to the above one using the Needleman-Wunsch algorithm. There are two differences. First, the N-terminal regions of all the proteins are not truncated. The second change is that the Smith-Waterman algorithm is used to find optimal local alignments.

2.5 Full sequence dipeptide composition

It is known that amino acid composition and subcellular localization is related [11]. But, the predictive power of the composition based approach is not enough to discriminate all

proteins. The dipeptide composition is the extension of amino acid composition adding the information on the local order of amino acids. In practice, it is proved that the predictive power of dipeptide composition is superior to amino acid composition. The compositional fraction of the ith dipeptide $f_{dc}(i)$ is defined by

$$f_{dc}(i) = \frac{N(i)}{\sum_{j=1}^{400} N(j)} \tag{2}$$

where $N(i)$ is the total count of ith dipeptide in the protein sequence. Then, the feature vector $\mathbf{x}$ is given by

$$\mathbf{x} = [f_{dc}(1), f_{dc}(2), \cdots, f_{dc}(400)]^T \tag{3}$$

Note that the dimension of the feature vector is 400.

2.6 Full sequence physico-chemical properties

It is generally thought that the factors determining the cellular destination are physico-chemical properties such as hydrophobicity or the position of charged amino acids since the signal sequences are not well conserved. The 121 physico-chemical properties, whose list is available at here[1], were used to represent a protein sequence as a 121 dimensional feature vector based on amino acid composition. We used the AAindex database to get the values of physico-chemical properties for all 20 amino acids, which are thought to be related to protein function [12]. To be expressed in comparable units, the values are normalized by subtracting the mean and dividing by the standard deviation. The average value of the ith physico-chemical property is defined by

$$g(i) = \sum_{j=1}^{20} A_i(j) f_{ac}(j) \tag{4}$$

where $A_i(j)$ is the normalized value of the jth amino acid of the ith physico-chemical property and $f_{ac}(j)$ is the compositional fraction of the jth amino acid. The feature vector $\mathbf{x}$ is given by

$$\mathbf{x} = [g(1), g(2), \cdots, g(121)]^T \tag{5}$$

3 CLASSIFICATION

3.1 Support vector machine classifier

SVM classifiers receive their popularity from the fact that they are based on the concept of statistical learning theory, or VC (Vapnik-Chervonenkis) theory, and they can achieve high performance in practical applications [13]. SVM classifiers are basically kernel-based learning algorithms and find the optimal hyperplane decision boundary in the feature space. In kernel-based algorithms, a kernel trick leads us to process the data in a higher-dimensional feature space constructed by a nonlinear mapping, without the explicit knowledge of the nonlinear mapping. In a view of statistics, the high dimensionality of the feature space can cause the curse of dimensionality. However, the optimal separating hyperplane with a maximal margin in the feature space, can relieve this problem. In statistical learning theory, we an minimize the complexity term of the upper bound of the expected risk by maximizing the margin of the separating hyperplane. The minimization of the upper bound can be viewed as relieving the over-fitting problem [14]. The maximization of the margin can be formulated as a quadratic optimization program so that a global solution can be easily obtained.

In the present study, we used OSU SVM Matlab toolbox 3.00 for the SVM classifier that is freely available at here[2]. The prediction of subcellular localization is a multi-class classification problem but the SVM classifier can only deal with the binary classification problem. Therefore, we need to construct a set of binary classifier for multi-class classification. We constructed $(M-1)M/2$ binary classifiers for M classes. In this pairwise classification, each possible pair of classes is considered and a test pattern is classified by the majority voting. The kernel function used in this study is the *radial basis function* (RBF) kernel with one parameter γ:

$$k(\mathbf{x}, \mathbf{y}) = \exp\left\{-\gamma \|\mathbf{x} - \mathbf{y}\|^2\right\}. \tag{6}$$

During the training and testing, only the RBF kernel parameter γ and the regularization parameter C were considered and the remaining parameters were kept constant.

3.2 Weighted majority voting

An SVM ensemble is a collection of several SVM classifiers whose individual decisions are combined in some aggregation methods. It is known that the performance of the SVM ensemble is often much better than that of individual SVM classifiers because of the independently trained SVM classifiers and their uncorrelated errors [15]. Since we trained several independent SVM classifiers for each feature, we need to aggregate them in an appropriate manner. The majority voting is the simplest and widely used aggregation method. Let $\hat{C}_k$, $k = 1, ..., K$, be the predicted class label of the kth SVM classifier in the SVM ensemble and C_j, $j = 1, ..., M$, denote the jth class label. The final predicted class label of the SVM ensemble $\hat{C}(\mathbf{x})$ for a given input $\mathbf{x}$ is determined by

$$\hat{C}(\mathbf{x}) = \arg\max_j \sum_{k=1}^{K} I_{kj} \tag{7}$$

in which

$$I_{kj} = \begin{cases} 1 & \text{if } \hat{C}_k(\mathbf{x}) = C_j \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

This voting scheme treats all SVM classifiers with equal weights. Since the prediction errors of the classifiers are often different, it is more realistic to give different weights in proportional to their prediction performance. In the weighted majority voting, the predicted class label of the SVM ensemble is given by

$$\hat{C}(\mathbf{x}) = \arg\max_j \sum_{k=1}^{K} W(k, j) I_{kj} \tag{9}$$

where $W(k, j)$ is the weight when the predicted class label of the kth classifier is C_j.

[1] home.postech.ac.kr/~blkimjk/aaindex1m.txt

[2] http://www.ece.osu.edu/~maj/osu_svm

3.3 The proposed prediction system

The overall schematic diagram of our prediction system is illustrated in Fig. 1. In our system, four major steps are needed to get the final decision. In the first preprocessing step, a test protein sequence is truncated after first 40 or 80 residues to get the N-terminal regions. The truncated N-terminal sequences, in the next feature extraction step, are converted into the two feature vectors by computing the scores of pairwise sequence alignments based on the profile HMM and the Needleman-Wunsch algorithm. The full sequence is also converted into the three feature vectors by computing the scores of the Smith-Waterman algorithm, or by calculating the compositional fractions of all dipeptides and the average values of the 121 physico-chemical properties. The representative sequences and profile HMM models can be divided into two parts which are positive and negative vectorization set. The positive vectorization set means all sequences or models of this set belong to the same class with the target sequence. The negative vectorization set denotes the opposite case. Therefore, the discriminative power of the feature vector is expected to increase since it contains the information of positive and negative examples. After these feature extraction steps, we obtain the fixed-length feature vectors. At the classification step, each of the five feature vectors is used as the input to $(M-1)M/2$ binary SVM classifiers for M classes. In this pairwise classification, the feature vector is assigned to the class associated with the highest value in the majority voting. After that, in the weighted majority voting step, the final predicted class label is decided based on the five predicted class labels.

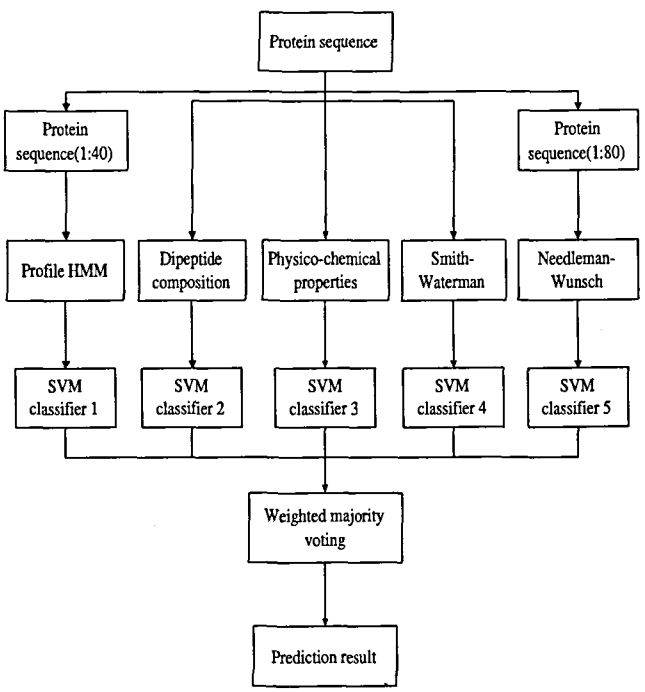

Figure 1: The schematic diagram of the proposed prediction system is illustrated.

4 NUMERICAL EXPERIMENTS AND RESULTS

4.1 Data sets

We used the animal data set generated by [4] for training and evaluating our prediction system. All sequences in the data set were extracted from SWISS-PROT release 42.7, and their cellular locations were chosen by referring the SUBCELL field. More information on the data creation steps is available at here[3]. We excluded protein sequences containing ambiguous amino acids such as B, Z, or X. As shown in Table 1, the data set consists of 11688 eukaryotic animal proteins with 9 cellular locations: cytoplasm, endoplasmic reticulum (ER), extracellular, golgi, lysosome, mitochondrion, nucleus, plasm membrane, and peroxisome.

Cellular location	Number of sequences
Cytoplasm	1945
ER	607
Extracellular	4410
Golgi	184
Lysosome	163
Mitochondrion	1220
Nucleus	2940
Plasma membrane	111
Peroxisome	108
Total	11688

Table 1: The number of proteins of each cellular locations in the data set.

4.2 Evaluation

The performance of our prediction system was evaluated using the 5-fold cross-validation. To measure the performance, sensitivity, specificity and Matthew's correlation coefficient (MCC) and overall accuracy were calculated using the following equations:

$$\text{Sensitivity}(i) = \frac{\text{tp}(i)}{\text{tp}(i)+\text{fn}(i)}, \tag{10}$$

$$\text{Specificity}(i) = \frac{\text{tp}(i)}{\text{tp}(i)+\text{fp}(i)}, \tag{11}$$

$$\text{MCC}(i) = \frac{\text{tp}(i)\text{tn}(i)-\text{fp}(i)\text{fn}(i)}{\sqrt{\text{de}(i)}}, \tag{12}$$

$$\text{Overall accuracy} = \frac{\sum_{i=1}^{k}\text{tp}(i)}{N} \tag{13}$$

where

$$\text{de}(i) = (\text{tp}(i)+\text{fn}(i))(\text{tp}(i)+\text{fp}(i))(\text{tn}(i)+\text{fp}(i))(\text{tn}(i)+\text{fn}(i)), \tag{14}$$

and N is the total number of sequences, k is the number of class, $\text{tp}(i)$ (true positive) is the number of correctly predicted

[3] http://www.cs.ualberta.ca/~bioinfo/PA/Subcellular/experiments/Extract_Data_42_7.html

sequences of class i, tn(i) (true negative) is the number of correctly predicted sequences which is not in class i, fp(i) (false positive) is the number of over predicted sequences of class i, and fn(i) (false negative) is the number of under predicted sequences of class i.

4.3 Results

The performance of all the feature extraction methods is shown in Table 3. To select the appropriate parameter values, we tested various values of the RBF kernel parameter γ and the regularization parameter C via the 5-fold cross-validation. The N-terminal profile HMM based prediction method ($\gamma = 0.003$ and $C = 10$) showed the overall accuracy of 83.62%. In the case of the N-terminal Needleman-Wunsch algorithm based method ($\gamma = 2$ and $C = 100$), the prediction accuracy reached 83.25% which is slightly lower that that of the profile HMM based method. The Smith-Waterman algorithm based method ($\gamma = 80$ and $C = 10$) showed the prediction accuracy of 83.23%. The overall accuracy of dipeptide composition ($\gamma = 170$ and $C = 10$) and physico-chemical properties ($\gamma = 4$ and $C = 10$) based methods reached 85.82% and 82.29%, respectively. To construct an SVM ensemble from the collection of the separately trained SVM classifiers, we tested four different aggregation methods based on the majority voting. In the case of the weighted majority voting, we used three different criteria (sensitivity, specificity, and MCC), for the weight $W(k, j)$ in Eqs. 9. Table 2 shows that the overall accuracy of the specificity based majority voting is the highest. Using the specificity based majority voting, we developed various SVM ensembles by changing the combination of the five SVM classifiers. The first SVM ensemble (ensemble 1) was constructed by combining the three SVM classifiers based on the pairwise sequence alignment (N-terminal profile HMM model, N-terminal Needleman-Wunsch algorithm, and full sequence Smith-Waterman algorithm). The overall accuracy of the ensemble 1 reached 86.17%, which was higher than that of any individual SVM classifier. The second SVM ensemble (ensemble 2) was developed based on the dipeptide composition and physico-chemical properties. The prediction accuracy of the ensemble 2 is slightly lower than the ensemble 1, and even worse than the dipeptide composition based method. To compare the effect of the two composition based features on the ensemble 1, we built two different ensembles from the ensemble 1 (ensemble 3 and 4). As shown in Table 3, the dipeptide composition based feature gives much higher effect on the performance of the constructed ensemble than the physico-chemical properties based feature. Finally, the SVM ensemble combining all the five SVM classifiers (ensemble 5) showed the overall accuracy of 88.53%. The prediction accuracy of our developed SVM ensemble is nearly 10% higher than the previous methods relying solely on amino acid sequence properties [16]. However, it is less meaningful to compare the prediction accuracy directly because the data sets are different. Comparing with the ontology-based approach, whose accuracy is about 4% higher than our method, is also unfair since the ontology-based methods use various extra information extracted from ontological labels [4].

Table 4 shows that the location of targeting information has a strong influence on the average sensitivity of the N-terminal and full sequence based feature extraction methods. The proteins targeted to extracellular, mitochondrion, and nucleus were predicted by the N-terminal based methods with higher accuracy. This result, in the case of nuclear proteins, are not well matched with the fact that the nuclear localization signals can be located anywhere. But, from this result, we could logically assume that most nuclear localization signals are located at the N-terminus. For the proteins whose targeting information are not restricted to the N-terminal region, the full sequence based methods showed better sensitivity.

Voting scheme	Overall accuracy
Unweighted majority voting	0.8842
Weighted majority voting (sensitivity)	0.8775
Weighted majority voting (MCC)	0.8813
Weighted majority voting (specificity)	0.8853

Table 2: A comparison of different majority voting schemes

Method	Location	Specificity	Sensitivity	MCC	Accuracy
Ensemble 1 (SW+NW+ HMM)	Cyt	0.8905	0.7064	0.7548	0.8617
	ER	0.9485	0.6985	0.8046	
	Ext	0.8717	0.9780	0.8668	
	Gol	0.9701	0.3533	0.5817	
	Lys	0.9667	0.3558	0.5832	
	Mit	0.8885	0.8492	0.8516	
	Nuc	0.8062	0.9184	0.8054	
	Pla	0.9130	0.3784	0.5853	
	Per	0.9516	0.5463	0.7190	
Ensemble 2 (DC+PC)	Cyt	0.8822	0.6776	0.7318	0.8504
	ER	0.8270	0.8666	0.8366	
	Ext	0.8482	0.9692	0.8374	
	Gol	0.8830	0.4511	0.6267	
	Lys	0.8872	0.7239	0.7985	
	Mit	0.8701	0.7631	0.7222	
	Nuc	0.8311	0.8755	0.7964	
	Pla	0.9123	0.4685	0.6513	
	Per	0.9412	0.5926	0.7448	
Ensemble 3 (SW+NW+ HMM+PC)	Cyt	0.8963	0.7064	0.7584	0.8691
	ER	0.9554	0.7414	0.8333	
	Ext	0.8649	0.9902	0.8711	
	Gol	1.0000	0.3587	0.5954	
	Lys	0.9848	0.3988	0.6235	
	Mit	0.9181	0.8549	0.8715	
	Nuc	0.8253	0.9126	0.8166	
	Pla	0.9423	0.4414	0.6427	
	Per	0.9531	0.5648	0.7317	
Ensemble 4 (SW+NW+ HMM+DC)	Cyt	0.8822	0.7548	0.7845	0.8815
	ER	0.9500	0.7512	0.8367	
	Ext	0.8756	0.9907	0.8822	
	Gol	1.0000	0.3750	0.6090	
	Lys	0.9789	0.5706	0.7446	
	Mit	0.9373	0.8582	0.8842	
	Nuc	0.8483	0.9129	0.8345	
	Pla	0.9455	0.4685	0.6633	
	Per	0.9701	0.6019	0.7624	
Ensemble 5 (SW+NW+ HMM+PC+ DC)	Cyt	0.8727	0.7578	0.7770	0.8853
	ER	0.9498	0.8105	0.8707	
	Ext	0.8844	0.9871	0.8876	
	Gol	0.9877	0.4348	0.6519	
	Lys	0.9798	0.5951	0.7610	
	Mit	0.9342	0.8615	0.8844	
	Nuc	0.8565	0.9116	0.8398	
	Pla	0.9474	0.4865	0.6767	
	Per	0.9706	0.6111	0.7684	

Cyt: Cytoplasm, Ext: Extracellular, Gol: Golgi, Lys: Lysosome, Mit: Mitochondrion, Nuc: Nucleus, Pla: Plasma membrane, Per: Peroxisome, HMM: N-terminal profile HMM model, NW: N-terminal Needleman-Wunsch algorithm, SW: Full sequence Smith-Waterman algorithm, PC: Full sequence physico-chemical properties, DC: Full sequence dipeptide composition

Table 3: A comparison of the five different SVM ensembles

Location	Targeting info.	N-terminal	Full sequence
Cytoplasm	No	0.6949	0.7554
ER	ERS+α	0.6607	0.8094
Extracellular	ERS	0.9530	0.9382
Golgi	ERS+α	0.3233	0.4330
Lysosome	ERS+α	0.3006	0.6728
Mitochondrion	MS	0.7869	0.7681
Nucleus	NLS	0.8977	0.8365
Plasma membrane	ERS+α	0.3108	0.4535
Peroxisome	PS	0.5093	0.5371

No: No targeting information, ERS: ER signal sequence (N-terminal), MS: Mitochondrion signal sequence (N-terminal), NLS: Nuclear localization signals (anywhere), PS: Peroxisome signal sequence (N-terminal or C-terminal), α: additional targeting information, N-terminal: average sensitivity of profile HMM and Needleman-Wunsch algorithm, Full sequence: average sensitivity of dipeptide composition, physico-chemical properties, and Smith-Waterman algorithm

Table 4: A comparison of the average sensitiviy of the N-terminal and full sequence based feature extraction methods with the targeting information.

5 CONCLUDING REMARKS

In this paper we have proposed various feature extraction methods based solely on amino acid sequence properties for prediction of subcellular localization. Taking into account the improved prediction performance, we conclude that our feature extraction methods using pairwise sequence alignment are well fitted to this classification problem. From the comparative study of the performance of the several SVM ensembles, we have shown that the performance of the feature extraction methods based on pairwise sequence alignment is significantly improved by combining the composition based methods. By comparing the average sensitivity of the N-terminal and full sequence based methods, we could get the biological insight on the location of targeting information. There are, however, a main problem that remain to be explored. In the case of proteins whose targeting information is not restricted to the N-terminal region, the sensitivity is considerably low. Therefore, more research is needed to resolve this low sensitivity problem. We hope this study will serve as a platform from which studies on developing feature extraction methods based on amino acid sequence property may be undertaken with greater depth and specificity.

Acknowledgment

This work was supported by National Core Research Center for Systems Bio-Dynamics.

REFERENCES

[1] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky, and J. Darnell. Molecular Cell Biology. W. H. Freeman, New York, 2003.

[2] 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., 300:1005–1016, 2000.

[3] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. Bioinformatics, 17:721–728, 2001.

[4] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics, 20:547–556, 2004.

[5] J. K. Kim, G. P. S. Raghava, K. S. Kim, S. Y. Bang, and S. Choi. Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine. In Proc. 3rd annual conference for the Korean society for bioinformatics, pages 158–166, Seoul, Korea, 2004.

[6] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., 48:443–453, 1970.

[7] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, 1998.

[8] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucelic Acids Res., 25:4876–4882, 1997.

[9] S. R. Eddy. Profile hidden Markov models. Bioinformatics, 14:755–763, 1998.

[10] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. J. Mol. Biol., 147:195–197, 1981.

[11] J. Cedano, P. Aloy, J. A. Perezpons, and E. Querol. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol., 266:594–600, 1997.

[12] S. Kawashima and M. Kanehisa. AAindex: amino acid index database. Nucelic Acids Res., 28:374–374, 2000.

[13] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge, 2000.

[14] K. R. Müller, S. Mika, G. Ratsch, K. Tsuda, B. Schölkopf. An introduction to kernel-based learning algorithms. IEEE Trans. on Neural Networks, 12:181–202, 2001.

[15] H. Kim, S. Pang, H. Je, D. Kim, and S. Y. Bang. Constructing support vector machine ensemble. Pattern Recognition, 36:2757–2767, 2003.

[16] K. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 19:1656–1663, 2003.