EUKARYOTIC PROTEIN SUBCELLULAR LOCALIZATION BASED ON LOCAL PAIRWISE PROFILE ALIGNMENT SVM

Jian Guo and Man-Wai Mak

Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University
USA

ABSTRACT

This paper studies the use of profile alignment and support vector machines for subcellular localization. In the training phase, the profiles of all protein sequences in the training set are constructed by PSI-BLAST and the pairwise profile-alignment scores are used to form feature vectors for training a support vector machine (SVM) classifier. During testing, the profile of a query protein sequence is computed and aligned with all the profiles constructed during training to obtain a feature vector for classification by the SVM classifier. Tests on Reinhardt and Hubbard's eukaryotic protein dataset show that the total accuracy can reach 99.4%, which is significantly higher than those obtained by methods based on sequence alignments and amino acid composition. It was also found that the proposed method can still achieves a prediction accuracy of 96% even if none of the sequence pairs in the dataset contains more than 5% identity. This paper also demonstrates that the performance of the SVM is proportional to the degree of its kernel matrix meeting the Mercer's condition.

1. INTRODUCTION

In protein classification, it has been found that substantial improvement in prediction accuracy can be achieved by converting variable-length sequences into to fixed-length feature vectors via preprocessing techniques. In most cases, the preprocessing is embedded in a kernel function to facilitate subsequent classification by support vector machines (SVMs). For example, in SVM-Fisher [1], a hidden Markov model (HMM) is trained from examples of a protein family. Then, given an unknown protein sequence, the derivative of the log-likelihood score for the protein sequence with respect to each of the HMM parameters is computed. The composition of these derivatives (Fisher scores) form a fixed-length vector, which is to be classified by an RBF-SVM. In SVM-Pairwise [2], each training sequence is compared with all other training sequences to form a list of pairwise alignment scores. These scores are then packed to form a feature vector. This process is repeated for every training sequence in the training set. In the mismatch kernel proposed by Leslie et al. [3], a set of subsequences of length k, namely k-mers, is defined. A query sequence is compared with the k-mers to count the number of times the k-mers appear in the sequence. The concatenation of the counts corresponding to all k-mers forms a feature vector in the k-mer feature space. Subsequent to the pioneering work of [1–3], the advantage of constructing kernel functions from sequences have been further demonstrated and improved by a number of investigators [4, 5].

While the preceding sequence-based methods perform reasonably well in protein homology detection, they may not be able to capture sufficient information from the sequences to detect remote homology. To overcome this difficulty, profile-based methods have been actively investigated in recent years [6–8]. A profile is a matrix in which elements in a column specify the frequency of each amino acid appears in that sequence position. Given a sequence, a profile can be derived by aligning it with a set of similar sequences. The similarity score between a known and an unknown sequence can be computed by aligning the unknown sequence with the profile of the known sequence [6] or by aligning the profile of the known sequence with that of the unknown sequence [7]. In the latter case, because the comparison involves not only two sequences but also their closely related sequences, the score is more sensitive to detecting weak similarity between protein families.

The focus of this paper is placed upon subcellular localization. The subcellular location is a key functional characteristic of potential gene products such as proteins. Because experimental subcellular localization is time-consuming and cannot be performed on a genome-wide scale, an accurate, reliable and efficient system is necessary to automate the prediction process. Kernel techniques based on sequence alignment mentioned earlier have been proven to be powerful for this task, as demonstrated by for example Kim [5].

This paper applies the pairwise profile alignment SVM,

This work was supported by the Research Grant Council of Hong Kong SAR (Project Nos. PolyU 5214/04E and 5230/05E).

which has been used successfully in detecting remote homologous proteins, to predict eukaryotic protein subcellular locations. Instead of extracting feature vectors directly from sequences, this method trains an SVM classifier by using scores of local pairwise profile alignment. Specifically, the profiles of all protein sequences are generated by PSI-BLAST [9] and the profiles of the testing proteins are aligned with the profiles of the proteins in the training set. Different SVM kernels are then created from these alignment scores for classification. Our experimental results demonstrate the advantage of embedding pairwise profile alignment scores into SVM kernels for eukaryotic protein subcellular localization. We also did experiments to demonstrate that meeting the Mercer's condition is an important step in designing the SVM kernels.

2. SEQUENCE VERSUS PROFILE ALIGNMENT

2.1. Local Pairwise Sequence Alignment

Pairwise sequence alignment has been widely used to compute the similarity between two DNA or two protein sequences. It finds the best match between two sequences by inserting some gaps into proper positions of the two sequences. One of the most successful local pairwise sequence alignment algorithms is the Smith-Waterman algorithm [15]. Denote

$$\mathcal{D} = \{S^{(1)}, \ldots, S^{(i)}, \ldots, S^{(j)}, \ldots, S^{(T)}\}$$

as a training set containing T sequences. Here, the i-th protein sequence is denoted as

$$S^{(i)} = S_1^{(i)}, S_2^{(i)}, \ldots, S_{n_i}^{(i)}, \quad 1 \le i \le T$$

where $S_k^{(i)} \in \mathcal{A}$, which is the set of 20 amino acid symbols, and n_i is the length of $S^{(i)}$. Using the BLOSUM62 substitution matrix [10], a set of similarity scores $\varepsilon'(S_u^{(i)}, S_v^{(j)})$ between position u of $S^{(i)}$ and position v of $S^{(j)}$ can be obtained. Then, based on these scores and the Smith-Waterman alignment algorithm [15], a sequence alignment score $\rho'(S^{(i)}, S^{(j)})$ can be obtained, which easily leads to a normalized alignment score:

$$\zeta'(S^{(i)}, S^{(j)}) = \frac{\rho'(S^{(i)}, S^{(j)})}{\sqrt{\rho'(S^{(i)}, S^{(i)})\rho'(S^{(j)}, S^{(j)})}}. \quad (1)$$

To facilitate SVM classification, we define four SVM kernels based on Eq. 1:

$$K_1'(S^{(i)}, S^{(j)}) = \zeta'(S^{(i)}, S^{(j)}) \quad (2)$$

$$K_2'(S^{(i)}, S^{(j)}) = \left(\zeta'(S^{(i)}, S^{(j)}) + 1\right)^d \quad (3)$$

$$K_3'(S^{(i)}, S^{(j)}) = \zeta'(S^{(i)}, S^*)\zeta'(S^{(j)}, S^*) \quad (4)$$

$$K_4'(S^{(i)}, S^{(j)}) = \left(\zeta'(S^{(i)}, S^*)\zeta'(S^{(j)}, S^*) + 1\right)^d \quad (5)$$

where d governs the degree of nonlinearity and S^* is a pseudo-sequence that can be approximated by a real sequence $S^{(l)}$ $(1 \le l \le T)$ as follows:

$$l = \arg \max_{1 \le t \le T} \zeta'(S^{(i)}, S^{(t)})\zeta'(S^{(j)}, S^{(t)}).$$

Note that these kernels only map the variable-length sequences to scores. To produce a better kernel for SVM classification, we can convert a variable-length sequence $S^{(i)}$ into a fixed-length feature vector

$$\boldsymbol{\zeta}'^{(i)} = [\zeta'(S^{(i)}, S^{(1)}) \ \ldots \ \zeta'(S^{(i)}, S^{(T)})]^{\mathsf{T}}$$

by aligning $S^{(i)}$ with each of the sequences in the training set. A kernel inner product between $S^{(i)}$ and $S^{(j)}$ can then be naturally obtained as $\langle \boldsymbol{\zeta}'^{(i)}, \boldsymbol{\zeta}'^{(j)} \rangle$. This leads to a class of algorithms referred to as SVM-pairwise adopted by [2,5]. Mathematically, the corresponding kernel is defined as

$$K_5'(S^{(i)}, S^{(j)}) = \sum_{t=1}^{T} \zeta'(S^{(i)}, S^{(t)})\zeta'(S^{(j)}, S^{(t)}), \quad (6)$$

for $1 \le i, j \le T$. Kernel K_5' has two advantages over K_1' to K_4'. First, every element in K_5' is the combined result of all pairwise comparisons, whereas in K_1' to K_4', every entry represents the pairwise comparison of two sequences only. Second, K_5' can be written as the dot product of two functions, i.e., $K_5'(S^{(i)}, S^{(j)}) =< \boldsymbol{\zeta}'^{(i)}, \boldsymbol{\zeta}'^{(j)} >$, whereas such dot product may not exist in K_1' and K_2'. Therefore, K_5' is guaranteed to be a valid kernel.

The sensitivity of detecting subtle local homogenous segments can be improved by replacing pairwise sequence alignment with pairwise profile alignment. In the next section, we will use the similarity scores of local pairwise profile alignment to generate kernel matrices for SVM classification.

2.2. Local Pairwise Profile Alignment

Following [16], here we use a protein sequence (called query sequence) as a seed to search and align homogenous sequences from the SWISSPROT 46.0 [17] protein database using the PSI-BLAST program [9] with parameters h and j set to 0.001 and 3, respectively. These aligned sequences share some homogenous segments and belong to the same protein family. The aligned sequences are further converted into two profiles to express their homogenous information: position-specific scoring matrix (PSSM) and position-specific frequency matrix (PSFM). Both PSSM and PSFM are matrices with 20 rows and L columns, where L is the total number of amino acids in the query sequence. Each column of a PSSM represents the log-likelihood of the residue substitutions at the corresponding positions in the query sequence [9]. The (i, j)-th entry of the matrix represents the

chance of the amino acid in the j-th position of the query sequence being mutated to amino acid type i during the evolution process. The PSFM contains the weighted observation frequencies of each position of the aligned sequences. Specifically, the (i, j)-th entry of PSFM represents the possibility of having amino acid type i in position j of the query sequence.

Denote the PSSM of $S^{(i)}$ and the PSFM of $S^{(j)}$ as

$$\begin{aligned} \mathbf{P}^{(i)} &= [\mathbf{p}_1^{(i)}, \mathbf{p}_2^{(i)}, \ldots, \mathbf{p}_{n_i}^{(i)}] \\ \mathbf{Q}^{(j)} &= [\mathbf{q}_1^{(j)}, \mathbf{q}_2^{(j)}, \ldots, \mathbf{q}_{n_j}^{(j)}] \end{aligned}$$

respectively, where

$$\begin{aligned} \mathbf{p}_r^{(i)} &= [p_{r,1}^{(i)}, p_{r,2}^{(i)}, \ldots, p_{r,20}^{(i)}]^{\mathsf{T}}, \quad 1 \le r \le n_i, \\ \mathbf{q}_s^{(j)} &= [q_{s,1}^{(j)}, q_{s,2}^{(j)}, \ldots, q_{s,20}^{(j)}]^{\mathsf{T}}, \quad 1 \le s \le n_j. \end{aligned}$$

We adopt the scoring function introduced by [7] to compute the similarity score between $\mathbf{p}_u^{(i)}$, $\mathbf{q}_v^{(j)}$, $\mathbf{p}_v^{(j)}$, and $\mathbf{q}_u^{(i)}$:

$$\varepsilon(S_u^{(i)}, S_v^{(j)}) = \sum_{h=1}^{20} \left(p_{u,h}^{(i)} q_{v,h}^{(j)} + p_{v,h}^{(j)} q_{u,h}^{(i)} \right). \quad (7)$$

In recent years, the Smith-Waterman algorithm has been extended to compute the similarity between two profiles [7]. In this paper, we further apply the Smith-Waterman algorithm [15] and its affine gap extension [18] to obtain the profile alignment score $\rho(S^{(i)}, S^{(j)})$ (see the supplementary webpage [20] for details). The local pairwise profile alignment score is then normalized as follows:

$$\zeta(S^{(i)}, S^{(j)}) = \frac{\rho(S^{(i)}, S^{(j)})}{\sqrt{\rho(S^{(i)}, S^{(i)}) \rho(S^{(j)}, S^{(j)})}}. \quad (8)$$

The normalization allows us to compare the alignment scores arising from profile matrices with different numbers of columns.

Let us further define five kernels based on the normalized scores (Eq. 8) for training SVM classifiers:

$$K_1(S^{(i)}, S^{(j)}) = \zeta(S^{(i)}, S^{(j)}) \quad (9)$$

$$K_2(S^{(i)}, S^{(j)}) = \left(\zeta(S^{(i)}, S^{(j)}) + 1 \right)^d \quad (10)$$

$$K_3(S^{(i)}, S^{(j)}) = \zeta(S^{(i)}, S^*) \zeta(S^{(j)}, S^*) \quad (11)$$

$$K_4(S^{(i)}, S^{(j)}) = \left(\zeta(S^{(i)}, S^*) \zeta(S^{(j)}, S^*) + 1 \right)^d \quad (12)$$

$$K_5(S^{(i)}, S^{(j)}) = \sum_{t=1}^{T} \zeta(S^{(i)}, S^{(t)}) \zeta(S^{(j)}, S^{(t)}) \quad (13)$$

where S^* is a pseudo-sequence that can be approximated by the profile of a real sequence $S^{(l)}$ $(1 \le l \le T)$ as follows:

$$l = \arg \max_{1 \le t \le T} \zeta(S^{(i)}, S^{(t)}) \zeta(S^{(j)}, S^{(t)}).$$

In this work, the degree d in Eqs. 3, 5, 10, and 12 were optimized empirically. Specifically, for K_2' and K_4', d was set to 1, and for K_2 and K_4, d was set to 20 and 10, respectively.

3. MULTI-CLASSIFICATION USING SVM

The multi-class problem can be solved by the one-vs-rest approach. Specifically, for a C-class problem (here $C = 4$) C independent SVM classifiers are constructed. The c-th SVM is trained from positively labelled samples of the c-th class and negatively labelled samples of all other classes. During classification, given an unknown protein sequence S, the output of the c-th SVM is computed as:

$$f_c(S) = \sum\nolimits_{i \in \mathcal{S}_c} y_{c,i} \alpha_{c,i} K(S, S^{(i)}) + b_c, \quad (14)$$

where $K(S, S^{(i)})$ is one of the kernels defined by Eqs. 2 to 6 or Eqs. 9 to 13, $\mathcal{S}_c$ is a set composed of the indexes of the support vectors, $y_{c,i} \in \{-1, +1\}$ is the label of the i-th sample, and $\alpha_{c,i}$ is the i-th Lagrange multiplier of the c-th SVM. Finally, the class of S is determined by a MAXNET:

$$y(S) = \arg \max_c f_c(S), \quad c = 1, \ldots, C$$

where $y(S)$ is the predicted class of S. In the following, we refer $y(S)$ with kernel K_5 to as pairwise profile alignment SVM (or simply PairProSVM), and $y(S)$ with kernel K_5' to as pairwise sequence alignment SVM (PairSeqSVM). See the supplementary web page [20] for the structure of the prediction system. The Spider SVM toolbox [21] was used to implement the SVM classifiers.

4. EXPERIMENTS AND RESULTS

4.1. Dataset

Reinhardt and Hubbard's eukaryotic protein dataset [11] was employed to test the performance of our method. This dataset has been used extensively for evaluating subcellular localization methods in the literature [11–14]. The sequences in this database were extracted from SWISSPORT 33.0 and the subcellular location of every protein has been annotated. The sequences were filtered, i.e., only those appeared to be complete and having reliable annotations were kept. Transmembrane proteins were excluded because reliable methods for predicting these proteins have been well developed. Plant sequences were also removed to ensure sufficient difference in composition. The resulting dataset comprises 2427 eukaryotic proteins (684 cytoplasm, 325 extracellular, 321 mitochondrial, and 1097 nuclear proteins).

4.2. Assessment of the Prediction Results

We used 5-fold cross validation to evaluate the performance, i.e., the original dataset was randomly divided into 5 subsets. Each subset was singled out in turn as a testing set, and the remaining ones were merged as the training set. The process was iterated 5 times until every subset has been used

for testing. The prediction results from all iterations were averaged. The overall prediction accuracy (OA), the accuracy for each subcellular location (Acc), and the Matthew's correlation coefficient (MCC) [19] were used to assess the prediction result. MCC allows us to overcome the shortcoming of accuracy (Acc) on unbalanced data.

Denote $M \in \Re^{C\times C}$ as the confusion matrix of the prediction result, where C is the number of classes. Then $M_{i,j}$ $(1 \le i,j \le C)$ represents the number of proteins that actually belong to class i but are predicted as class j. Denote

$$p_c = M_{c,c}, \quad q_c = \sum_{i=1,i\neq c}^{C}\sum_{j=1,j\neq c}^{C} M_{i,j},$$
$$r_c = \sum_{i=1,i\neq c}^{C} M_{i,c}, \quad s_c = \sum_{j=1,j\neq c}^{C} M_{c,j}, \tag{15}$$

where c $(1 \le c \le C)$ is the index of a particular class. For class c, p_c is the number of true positives, q_c is the number of true negatives, r_c is the number of false positives, and s_c is the number of false negatives. Based on the notations above, the overall accuracy (OA), the accuracy of class c (Acc_c), the Matthew's Correlation Coefficient of class c (MCC_c), the overall MCC (OMCC) and the weighted average MCC (WAMCC) are defined respectively as:

$$\text{OA} = \frac{\sum_{c=1}^{C} M_{c,c}}{\sum_{i=1}^{C}\sum_{j=1}^{C} M_{i,j}} \qquad \text{Acc}_c = \frac{M_{c,c}}{\sum_{j=1}^{C} M_{c,j}} \tag{16}$$

$$\text{MCC}_c = \frac{p_c q_c - r_c s_c}{\sqrt{(p_c+s_c)(p_c+r_c)(q_c+s_c)(q_c+r_c)}} \tag{17}$$

$$\text{OMCC} = \frac{\hat{p}\,\hat{q} - \hat{r}\hat{s}}{\sqrt{(\hat{p}+\hat{s})(\hat{p}+\hat{r})(\hat{q}+\hat{s})(\hat{q}+\hat{r})}} \tag{18}$$

$$\text{WAMCC} = \sum_{c=1}^{C} \frac{p_c + s_c}{N}\text{MCC}_c \tag{19}$$

where $N = \sum_{c=1}^{C}(p_c + s_c)$, $\hat{p} = \sum_{c=1}^{C} p_c$, $\hat{q} = \sum_{c=1}^{C} q_c$, $\hat{r} = \sum_{c=1}^{C} r_c$ and $\hat{s} = \sum_{c=1}^{C} s_c$.

4.3. Results on Reinhardt & Hubbard's dataset

The prediction results of PairProSVM and PairSeqSVM are listed in Table 1. Also listed are the results of other three existing methods (NNPSL [11], SubLoc [12], and ESLpred [14]) for comparison. The overall accuracy of PairProSVM achieves 99.4%, which compares favorably with NNPSL (66%), SubLoc (79.4%), ESLpred (88%), and PairSeqSVM (87.9%). This suggests that profile alignment extracts more information on subcellular location than sequence alignment and amino acid composition. The prediction performance on mitochondria is particularly interesting because they are usually difficult to be accurately predicted by existing methods. The prediction accuracy of PairProSVM on mitochondria reaches 98.4%, which compares favorably with NNPSL (61%), SubLoc (56.7%), and ESLpred (68.2%). The MCC of PairProSVM on mitochondria reaches 0.98, which represents 69% and 42% relative improvement with respect to SubLoc and ESLpred, respectively.

Table 1 shows that PairSeqSVM performs better than NNPSL and SubLoc but poorer than ESLpred and PairProSVM. Note that ESLpred combines different protein features—including amino acid composition, physico-chemical properties, dipeptide compositions, and PSI-BLAST decisions—for subcellular localization. The inferiority of PairSeqSVM as compared to ESLpred suggests that the fusion of different protein features has advantages. Therefore, using a diversity of protein features in PairSeqSVM and PairProSVM may further improve the prediction performance.

4.4. Results on Redundance-Removed Datasets

Note that Reinhardt and Hubbard's dataset includes some protein sequences with high homologous (identity $> 80\%$). Therefore, it is worthwhile to investigate whether the good performance of PairProSVM is due to the similarity in the sequences. To answer this question, we constructed a series of redundance-removed datasets by eliminating the most similar sequences. Specifically, any pairs of sequences in a redundance-removed dataset should not have an identity higher than λ, where λ is a predefined threshold. The blastclust program in the NCBI BLAST software was employed to implement the filtering process.[1] Different λ, from 5% to 100% with intervals of 5%, were tested. Note that when $\lambda = 100\%$, no proteins were removed, which means that Reinhardt and Hubbard's dataset was used.

Figure 2(a) shows the number of samples in the redundance-removed datasets with different λ. We did a 5-fold cross validation on the reduced datasets and the prediction results are shown in Figure 2(b), which clearly shows that PairProSVM is less sensitive to the similarity among the training sequences than the PairSeqSVM.

4.5. Comparison on Alignment Kernels

The performance of the profile alignment kernels (K_1 to K_5) were are shown in Table 2. Evidently, the performance of the five profile alignment kernels vary significantly. To reveal the cause of this performance variation, we calculated the percentage of negative eigenvalues (the ratio of the number of negative eigenvalues to the number of all eigenvalues) for each kernel matrix and listed the results in the bottom of Table 2. The results show that the lower the per-

[1] We used the command "blastclust -L 0 -S lambda", where lambda $= 5, 10, \ldots, 100$.

Table 1. Comparison of different prediction methods for Reinhardt and Hubbard's eukaryotic protein dataset. NNPSL [11] and SubLoc [12] use amino acid composition as features; ESLpred [14] is a mixture method combining amino acid composition, dipeptide composition, physico-chemical properties, and BLAST decisions; PairSeqSVM uses pairwise sequence alignment to create SVM kernels [2]; PairProSVM is the method introduced in this paper. The results of SubLoc were obtained by leave-one-out cross validation. The results of NNPSL were obtained by 10-fold cross validation, and those of ESLpred, PairSeqSVM, and PairProSVM were obtained by 5-fold cross validation. Acc: accuracy; MCC: Matthew's correlation coefficient.

Subcellular	NNPSL	SubLoc		ESLpred		PairSeqSVM (K_5')		PairProSVM (K_5)	
Location	Acc(%)	Acc(%)	MCC	Acc(%)	MCC	Acc(%)	MCC	Acc(%)	MCC
Cytoplasm	55	76.9	0.64	85.2	0.79	85.5	0.79	99.9	1.00
Extracellular	75	80.0	0.78	88.9	0.91	84.6	0.89	98.5	0.98
Mitochrondria	61	56.7	0.58	68.2	0.69	66.7	0.71	98.4	0.98
Nuclear	72	87.4	0.75	95.3	0.87	96.8	0.87	99.7	1.00
Overall	66	79.4	–	88.0	–	88.0	0.84	99.4	0.99
Weighted Average	–	–	0.70	–	0.83	–	0.83	–	0.99

centage of negative eigenvalues the higher the overall accuracy. For example, K_1 and K_3 have as many as 8.5% and 6.2% negative eigenvalues, which result in very poor performance. Among the five pairwise profile alignment kernels (K_1 to K_5), only K_5 satisfies the Mercer's condition and consequently it achieves the best performance.

We also found that none the eigenvalues of the five sequence alignment kernel matrices is less than 0, which means that all these kernels satisfy the Mercer's condition and consequently their performance are almost identical (ranging from 87% to 88%).

The ROC curves in Figure 1 show the prediction performance of these kernels under various decision thresholds. Because the five sequence alignment kernels have similar performance, only the ROC curve of K_5' was plotted. All profile alignment kernels (except K_5) do not satisfy the Mercer's condition and consequently perform worse than the sequence alignment kernels. Among the kernels that satisfy the Mercer's condition, K_5 performs the best. This demonstrates that profile alignment kernels can extract more useful information than sequence alignment kernels provided that they meet the Mercer's condition.

In sum, the comparison of accuracy, MCC and ROC curves demonstrates the importance of Mercer's condition, and a lower percentage of negative eigenvalues of a kernel matrix generally results in better prediction performance.

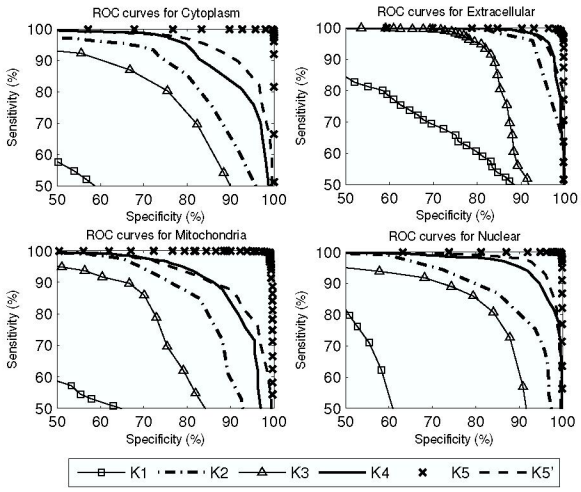

Fig. 1. ROC curves showing the prediction performance of K_1 to K_5 and K_5'.

5. CONCLUSIONS

This paper applies profile-to-profile alignment to eukaryotic protein subcellular localization. Protein alignment profiles are calculated by searching the SWISSPROT 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. We have tested this method on Reinhardt and Hubbard's eukaryotic protein dataset and found that the overall accuracy can reach 99.4%, which is substantially higher than those obtained by existing methods in the literature. It was also found that the overall accuracy still remains at 95% even if the majority of redundant proteins have been removed from the dataset. We hope this *in-silico* method can be complementary to experimental subcellular localization techniques.

6. REFERENCES

[1] Jaakkola,T., Diekhans,M., and Haussler,D. "A discriminative framework for detecting remote protein homologies," *J. Comput. Biol.*, vol.

Table 2. Comparison of the five profile alignment kernels on Reinhardt and Hubbard's eukaryotic protein dataset. Acc: accuracy; MCC: Matthew's correlation coefficient; PNE: percentage of negative eigenvalues.

Location	K_1		K_2		K_3		K_4		K_5	
	Acc (%)	MCC	Acc (%)	MCC	Acc (%)	MCC	Acc (%)	MCC	Acc (%)	MCC
Cytoplasm	36.4	0.03	69.3	0.68	74.0	0.55	78.8	0.76	99.9	1.00
Extracellular	41.9	0.58	67.1	0.79	75.1	0.83	82.5	0.89	98.5	0.98
Mitochondria	25.6	0.18	41.1	0.59	48.9	0.57	57.0	0.70	98.4	0.98
Nuclear	59.0	0.20	95.5	0.63	84.8	0.67	97.9	0.79	99.7	1.00
Overall	45.9	0.28	77.1	0.70	75.7	0.68	85.0	0.80	99.4	0.99
Weighted Average	–	0.20	–	0.66	–	0.64	–	0.78	–	0.99
PNE (%)	8.5		0.3		6.2		0.2		0	

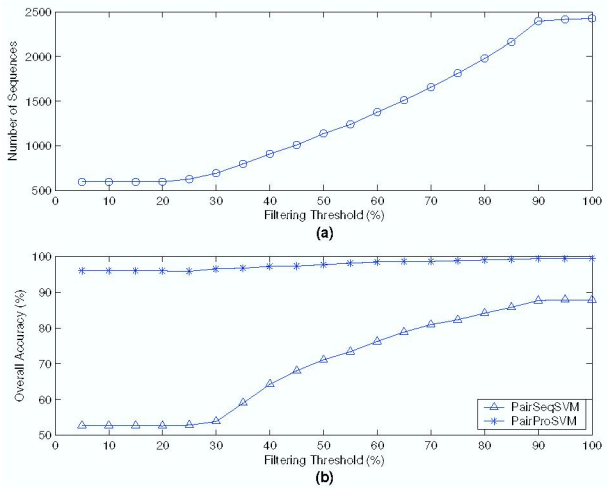

Fig. 2. The relationship between the filtering threshold λ and (a) the number of samples in the redundance-removed dataset and (b) the overall accuracy of PairProSVM on the redundance-removed dataset.

7, pp. 95–114, 2000.

[2] Liao, L., and Noble, W.S. "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *J. Comput. Biol.*, vol. 10(6), pp. 857–868, 2003.

[3] Leslie, C. S., Eskin, Cohen, A., Eeston, J., and Noble, W.S. "Mismatch string kernels for discriminative protein classification," *Bioinformatics*, vol. 20(4), pp. 467-476, 2004.

[4] Saigo, H., Vert, J.P., Ueda, N., and Akutsu, T., "Protein homology detection using string alignment kernels," *Bioinformatics*, vol. 20, pp. 1682-1689, 2004.

[5] Kim, J.K., Raghava, G.P.S, Bang, S.Y., and Choi, S. "Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine," *Pattern Recog. Lett.*, In Press.

[6] Busuttil, S., Abela, J., and Pace, G.J. "Support vector machines with profile-based kernels for remote protein homology detection," *Genome Informatics*, vol. 15(2), pp. 191–200, 2004.

[7] Rangwala, H. and Karypis, G. "Profile-based direct kernels for remote homology detection and fold recognition," *Bioinformatics*, vol. 21, no. 23, pp. 4239–4247, 2005.

[8] Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. "Profile-based string kernels for remote homology detection and motif extraction," *J. Bioinfom. Comput. Biol.*, vol. 3, pp. 527–550, 2005.

[9] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, 1997.

[10] Henikoff, S. and Henikoff, J. G. "Amino acid substitution matrices from protein blocks," *Proc. Natl. Acad. Sci.*, USA 89, pp. 10915-10919, 1992.

[11] Reinhardt, A. and Hubbard, T. "Using neural networks for prediction of the subcellular location of proteins," *Nucleic Acids Res.*, vol. 26, pp. 2230–2236, 1998.

[12] Hua, S.J. and Sun, Z.R. "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, pp. 721–728, 2001.

[13] Huang,Y. and Li Y.D. "Prediction of protein subcellular locations using fuzzy k-NN method," *Bioinformatics*, vol. 20, no. 1, pp. 21–28, 1999.

[14] Bhasin, M. and Raghava, G.P.S. "ESLpred: SVM based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST," *Nucleic Acids Res., Webserver Issue*, vol. 32, pp. 414–419, 2004.

[15] Smith, T.F. and Waterman, M.S. "Comparison of biosequences," *Adv. Appl. Math.*, vol. 2, pp. 482–489, 1981.

[16] Rychlewski L, Zhang B, Godzik A. "Fold and function predictions for *Mycoplasma genitalium* proteins," *Fold Des.*, vol. 3(4), pp. 229–238, 1998.

[17] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K.O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Res.*, vol. 31, pp. 365–370, 2003.

[18] Gotoh, O. "An improved algorithm for matching biological sequences," *J. Mol. Biol.*, vol. 162, pp. 705–708, 1982.

[19] Matthews, B.W. "Comparison of predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta*, vol. 405, pp. 442–451, 1975.

[20] http://www.eie.polyu.edu.hk/∼mwmak/BSIG/PairProSVM.htm.

[21] http://www.kyb.tuebingen.mpg.de/bs/people/spider/