A Solution to the Curse of Dimensionality Problem in Pairwise Scoring Techniques

Man-Wai Mak[1] and Sun-Yuan Kung[2]

[1] Center for Multimedia Signal Processing,
Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University, China
enmwmak@polyu.edu.hk*
[2] Dept. of Electrical Engineering, Princeton University, USA
kung@princeton.edu

Abstract. This paper provides a solution to the curse of dimensionality problem in the pairwise scoring techniques that are commonly used in bioinformatics and biometrics applications. It has been recently discovered that stacking the pairwise comparison scores between an unknown patterns and a set of known patterns can result in feature vectors with nice discriminative properties for classification. However, such technique can lead to curse of dimensionality because the vectors size is equal to the training set size. To overcome this problem, this paper shows that the pairwise score matrices possess a symmetric and diagonally dominant property that allows us to select the most relevant features independently by an FDA-like technique. Then, the paper demonstrates the capability of the technique via a protein sequence classification problem. It was found that 10-fold reduction in the number of feature dimensions and recognition time can be achieved with just 4% reduction in recognition accuracy.
Keywords: Feature selection; Fisher discriminant analysis; curse of dimensionality; protein sequence analysis; subcellular localization; support vector machines.

1 Introduction

In computational biology, the subcellular location and structural family of a protein provide important information about its biochemical functions. However, experimental analysis of proteins is time-consuming and cannot be performed on genome-wide scales. Therefore, a reliable and efficient method is essential for automating the prediction of proteins' subcellular locations and the classification of protein sequences into functional and structural families.

It has been found that converting variable-length protein sequences to fixed-length feature vectors via preprocessing techniques can improve the accuracy

* This work was supported by The Hong Kong Polytechnic University, Grant No. APH18 and the RGC of Hong Kong, Grant No. PolyU5230/05E. The authors thank Jian Guo for providing the pairwise alignment score matrices.

of subcellular localization and protein classification. In most cases, the preprocessing is embedded in a kernel function to facilitate subsequence classification by support vector machines (SVMs). For example, in SVM-Fisher [7], 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 [11], 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. In the mismatch kernel [10], a set of subsequences of length k, namely k-grams, is defined. A query sequence is compared with the k-grams to count the number of times the k-grams appear in the sequence. The concatenation of the counts corresponding to all k-grams forms a feature vector.

While the 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 [3, 9, 13]. 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 profile of the known sequence with that of the unknown sequence [13]. 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. Research has also found that profile alignment can achieve better performance than sequence alignment in predicting subcellular locations [4].

The comparison of two temporal sequences are often hampered by the fact that the two sequences often have different lengths whether or not they belong to the same family. To overcome this problem, pairwise comparison between a sequence with a set of known sequences has been a popular scheme for creating fixed-size feature vectors from variable length sequences. Many of the methods mentioned earlier (e.g., [4, 8, 11]) fall into this scheme. Although this pairwise approach can usually create feature vectors with better discriminative properties, it also has its own limitation. The main problem is that the feature dimension is the same as the number of training patterns, c.f., Figure 1 where we have T training patterns with vector dimension equal to T too. This creates a curse of dimensionality, because for most biometric and bioinformatic applications, the training size could be very large. In fact, for the applications addressed in this paper, they are in the range of several thousands. The downside of such a curse of dimensionality is that it could hurt both training and recognition speed.

An obvious solution to the curse of dimensionality problem is to reduce the feature size and yet retaining the most important information critical for the classification of the training patterns. Research has found that just over 10% of the profile contributes 90% of the total score for positive training sequences [9],

suggesting that some features are more relevant to the classification task than the others. The feature sized reduction can be accomplished via either finding principle subspace or via weeding out those less significant features. This paper takes the latter approach. Moreover, in this paper, the importance of each feature is independently computed, unlike the subspace reduction schemes. More specifically, for each component in the feature vectors, the symmetric divergence between the densities of the feature values from the positive class and the negative class is computed. Then, the features with symmetric divergences significantly greater than the mean divergence are selected. New feature vectors with reduced dimension are then used to train SVM classifiers.

The paper is organized as follows. Section 2 outlines the sequence and profile alignment algorithms. Section 3 explains how the non-discriminative features can be weeded out to reduce the dimensionality of the feature vectors, which are then classified by multi-class SVMs in Section 4. The feature selection technique is evaluated in Section 5 where significant reduction in recognition time is demonstrated.

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 attempts to find 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 algorithm is the Smith-Waterman algorithm [16]. 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

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

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 [5], 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, a sequence alignment score $\rho'(S^{(i)}, S^{(j)})$ can be obtained, which easily leads to a normalized alignment score:

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

To facilitate SVM classification, we can convert a variable-length sequence $S^{(i)}$ into a fixed-length feature vector

$$\boldsymbol{\zeta}'^{(i)} = [\zeta'(S^{(i)}, S^{(1)}) \quad \ldots \quad \zeta'(S^{(i)}, S^{(T)})]^{\mathsf{T}} \tag{2}$$

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 naturally be obtained as $< \boldsymbol{\zeta}'^{(i)}, \boldsymbol{\zeta}'^{(j)} >$. This leads to a class of algorithms referred to as the SVM-pairwise method adopted by [8, 11].

The sensitivity of detecting subtle local homogenous segments can be improved by replacing pairwise sequence alignment with pairwise profile alignment. In the next subsection, we will use the similarity scores of local pairwise profile alignment to generate fixed-length feature vectors for SVM classification.

2.2 Local Pairwise Profile Alignment

Following [15], here we use a protein sequence (called query sequence) as a seed to search and align homogenous sequences from the SWISSPROT 46.0 [2] protein database using the PSI-BLAST program [1] 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 [1]. 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. A PSFM contains the weighted observation frequencies of each position of the aligned sequences. Specifically, the (i, j)-th entry of a PSFM represents the possibility of having amino acid type i in position j of the query sequence.

Let us denote the PSSM of $S^{(i)}$ and the PSFM of $S^{(j)}$ as

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

respectively, where

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

We adopt the scoring function introduced by [13] to compute the similarity score between $\mathbf{p}_u^{(i)}$, $\mathbf{q}_v^{(j)}$, $\mathbf{p}_v^{(j)}$, and $\mathbf{q}_u^{(i)}$:

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

which is then applied to the the Smith-Waterman algorithm to obtain the profile alignment score $\rho(S^{(i)}, S^{(j)})$ (see [6] for details). The local pairwise profile alignment score is then normalized as follows:

$$\zeta(S^{(i)}, S^{(j)}) = \frac{\rho(S^{(i)}, S^{(j)})}{\sqrt{\rho(S^{(i)}, S^{(i)}) \rho(S^{(j)}, S^{(j)})}}. \tag{3}$$

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

Similar to the sequence alignment introduced in Section 2.1, we convert a variable-length sequence $S^{(i)}$ into a fixed-length feature vector

$$\boldsymbol{\zeta}^{(i)} = [\zeta(S^{(i)}, S^{(1)}) \quad \ldots \quad \zeta(S^{(i)}, S^{(T)})]^{\mathsf{T}} \tag{4}$$

for SVM classification.

3 Dimensionality Reduction by Fisher Feature Selection

A problem of the pairwise comparison scheme is that the feature vector dimension is equal to the number of training sequences. For most datasets, this leads to feature vectors with thousands of dimensions. Figure 1 shows all feature vectors ($\boldsymbol{\zeta}'^{(i)}$ and $\boldsymbol{\zeta}^{(i)}$ $\forall i$) formed by the pairwise alignment scores obtained from a dataset used in this paper. A collection of all feature vectors forms a score matrix in which each column represents a feature vectors with number of dimensions equal to the number of training samples and each row represents a feature. As can be observed from Figure 1, both matrices are symmetric and diagonally dominant, i.e., the closer an entry is to the diagonal the more significant the correlation is. Due to this diagonally dominant property, for each individual feature there is a natural home class. This property allows us to apply a Fisher-type criterion to determine how discriminative is the home class from the rest of the classes in a feature-by-feature basis.

Consider a training set containing T training sequences. Without loss of generality, let us assume that the first T_p training sequences are obtained from the positive class and the rest T_n $(= T - T_p)$ sequences are obtained from the negative class. For the m-th feature dimension (m-th row in the score matrix), we have two sets of scores:[3]

$$\gamma_p^{(m)} = \{\zeta(S^{(m)}, S^{(1)}), \ldots, \zeta(S^{(m)}, S^{(T_p)})\} \tag{5}$$

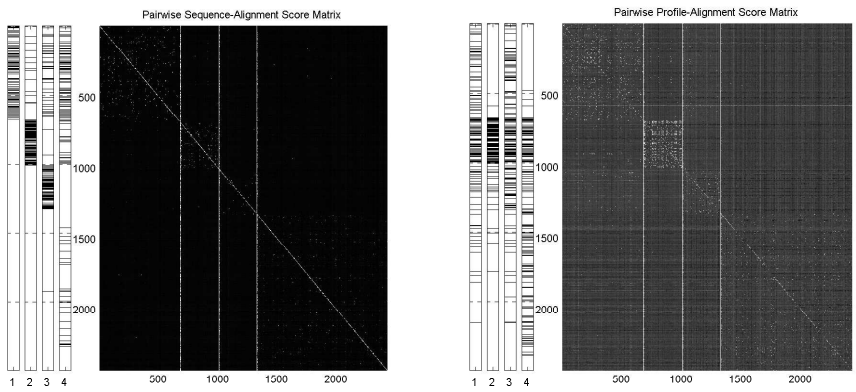
$$\gamma_n^{(m)} = \{\zeta(S^{(m)}, S^{(T_p+1)}), \ldots, \zeta(S^{(m)}, S^{(T)})\}. \tag{6}$$

The separation between these two sets of scores represents the discriminative power of the m-th feature. The separation can be quantified by the symmetric divergence between the score distributions in the two sets as follows:

$$D(\gamma_p^{(m)}, \gamma_n^{(m)}) = \frac{1}{2}\left(\frac{(\sigma_n^{(m)})^2}{(\sigma_p^{(m)})^2} + \frac{(\sigma_p^{(m)})^2}{(\sigma_n^{(m)})^2}\right) - 1 + \frac{1}{2}\left(\mu_p^{(m)} - \mu_n^{(m)}\right)^2 \left(\frac{1}{(\sigma_p^{(m)})^2} + \frac{1}{(\sigma_n^{(m)})^2}\right). \tag{7}$$

where $\mu_p^{(m)}$ and $\sigma_p^{(m)}$ are the mean and standard derivation of the m-th feature in the positive class, and $\mu_n^{(m)}$ and $\sigma_n^{(m)}$ are the corresponding parameters in the negative class. Figure 2 shows the histograms of the symmetric divergences

[3] For notational simplicity, we use profile alignment score matrix. The method can also be applied to sequence alignment score matrix.



(a) (b)

Fig. 1. (a) Pairwise sequence-alignment score matrix and (b) pairwise profile-alignment score matrix. The horizontal lines in the 4 columns on the left of the score matrices denote the features selected by the Fisher feature selection method for the four 1-vs-rest RBF-SVM classifiers. The numbers under the 4 columns represent the class labels: (1) Cytoplasm, (2) Extracellular, (3) Mitochondrial, and (4) Nuclear. The vertical lines on the score matrices partition the dataset into these four classes.

corresponding to the profile alignment scores. Apparently, only a small fraction of the features have large symmetric divergences, which means that only a few feature dimensions are relevant for classification. Based on this observation, a relevant feature set is obtained as follows:

$$\mathcal{M} = \left\{ m : D(\gamma_p^{(m)}, \gamma_n^{(m)}) > \mu_D + k\sigma_D \quad \forall\ 1 \leq m \leq T \right\} \tag{8}$$

where μ_D and σ_D are the mean and standard derivation of the symmetric divergence, respectively, and k is a user-defined parameters. We then select the elements in $\boldsymbol{\zeta}$ according to the set $\mathcal{M}$ to construct feature vectors $\boldsymbol{\phi}$ of $|\mathcal{M}|$ dimensions for SVM training and classification. The same process is also applied to sequence alignment score vectors $\boldsymbol{\zeta}'$ to obtain $|\mathcal{M}|$-dimensional vectors $\boldsymbol{\phi}'$. The four columns on the left of the score matrices in Figure 1 depict the feature dimensions selected by this scheme. Evidently, for each class, the scheme tends to select the features that have higher pairwise scores in that class.

4 Multi-Classification using SVM

The multi-class problem can be solved by a 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

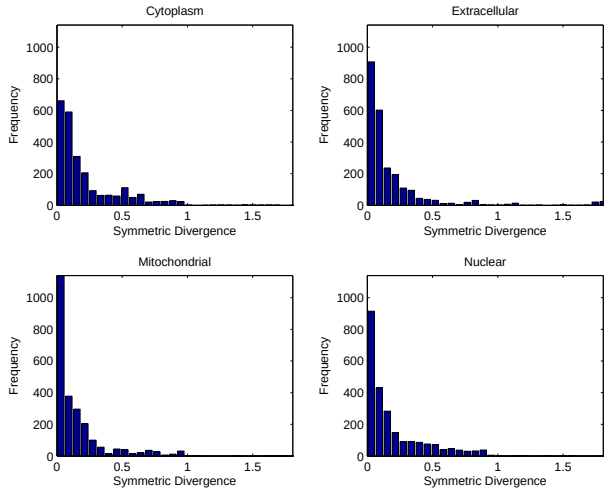


Fig. 2. Histograms of the symmetric divergences between the profile alignment scores of each positive class and its corresponding negative classes.

unknown protein sequence S, the output of the c-th SVM is computed as:

$$f_c(S) = \sum\nolimits_{i \in \mathcal{S}_c} y_{c,i} \alpha_{c,i} K(S, S^{(i)}) + b_c, \tag{9}$$

where $\mathcal{S}_c$ is a set composed of the indexes of the support vectors, $y_{c,i}$ is the label of the i-th sample, $\alpha_{c,i}$ is the i-th Lagrange multiplier of the c-th SVM, and

$$K(S, S^{(i)}) = \begin{cases} < \boldsymbol{\zeta}', \boldsymbol{\zeta}'^{(i)} > & \text{Sequence alignment using full features} \\ < \boldsymbol{\zeta}, \boldsymbol{\zeta}^{(i)} > & \text{Profile alignment using full features,} \\ \exp\left\{-\|\boldsymbol{\phi}' - \boldsymbol{\phi}'^{(i)}\|^2/\sigma^2\right\} & \text{Sequence alignment using selected features} \\ \exp\left\{-\|\boldsymbol{\phi} - \boldsymbol{\phi}^{(i)}\|^2/\sigma^2\right\} & \text{Profile alignment using selected features} \end{cases}$$

is the kernel function. Note that because the dimensionality of $\boldsymbol{\phi}'$ and $\boldsymbol{\phi}$ is considerably smaller than that of $\boldsymbol{\zeta}'$ and $\boldsymbol{\zeta}$, it is possible to use RBF-SVMs to classify $\boldsymbol{\phi}'$ and $\boldsymbol{\phi}$, whereas linear SVMs are chosen for classifying $\boldsymbol{\zeta}'$ and $\boldsymbol{\zeta}$ to avoid curse of dimensionality. Finally, the class of S is determined by a MAXNET:

$$y(S) = \arg\max_c f_c(S),$$

where $y(S)$ is the predicted class of S. In the following, we refer $y(S)$ with kernel $K(\cdot, \cdot)$ obtained from profile alignment to as pairwise profile alignment SVM (or simply PairProSVM), and $y(S)$ with kernel obtained from sequence alignment to as pairwise sequence alignment SVM (PairSeqSVM).

5 Experiments and Results

Reinhardt and Hubbard's eukaryotic protein dataset [14], which contains 2427 amino acid sequences, was employed to test the performance of our method. The sequences in this dataset were extracted from SWISSPORT 33.0 and their subcellular locations (684 cytoplasm, 325 extracellular, 321 mitochondrial, and 1097 nuclear proteins) have been annotated.

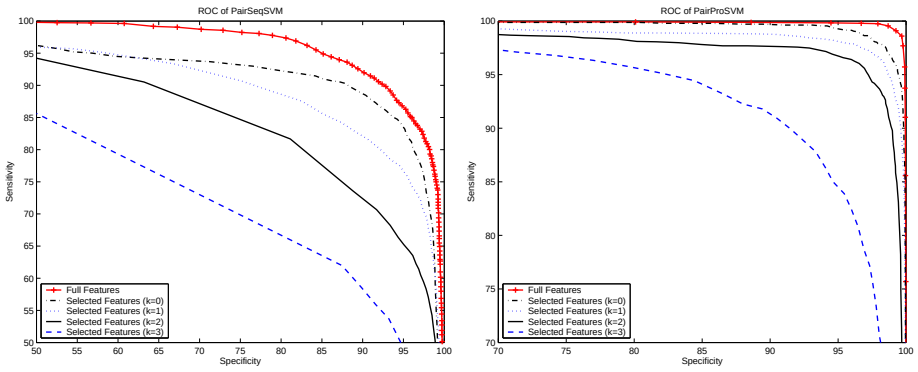
We used 5-fold cross validation for performance evaluation, i.e., the original dataset was randomly divided into 5 subsets. Each subset was singled out in turn as a testing set, and the remaining ones were merged as the training set. The process was iterated 5 times until every subset has been used for testing. The prediction results from all iterations were averaged. The overall prediction accuracy (OA), and the Matthew's correlation coefficient (MCC) [12] were used to assess the prediction result. MCC can overcome the shortcoming of using accuracy as a performance measure on unbalanced data. For example, in an unbalanced dataset where the majority of test samples belong to the positive class, a classifier predicting all samples as positive cannot be regarded as a good classifier if it fails to detect any negative samples. In such case, the MCC will be zero although the overall accuracy is high. Therefore, MCC is a better measure for unbalanced classification.

The prediction results of PairProSVM and PairSeqSVM are listed in Table 1. The overall accuracy of PairProSVM achieves 99.4%, which compares favorably with PairSeqSVM (87.9%). This suggests that profile alignment provides more information on subcellular location than sequence alignment. Also shown in Table 1 are the number of feature dimensions selected by the proposed feature selection scheme and the corresponding recognition time. Evidently, about 10-fold reduction in the number of feature dimensions can be achieved without jeopardizing the overall accuracy and MCC significantly. This dimensionality reduction not only solves the curse of dimensionality problem but also leads to about 10-fold reduction in recognition time. Figure 3 shows the ROC performance of PairSeqSVM and PairProSVM with and without feature selection. The results show that the performance progressively degrades when the value of k in Eq. 8 increases. However, the amount of performance degradation (especially in PairProSVM) is not very significant when k is small.

Although fast recognition may not be critical for bioinformatics because sequence classification can be done off-line, it is critically important for biometrics where real-time recognition is required. Therefore, the feature reduction techniques proposed in this paper is potentially useful for reducing the cost of biometric systems.

6 Conclusions

A novel method to speed up the recognition of pairwise scoring features has been presented. The method can also alleviate the curse of dimensionality problem commonly encountered in the pairwise scoring techniques. It was found that the



(a) (b)

Fig. 3. ROC curves showing the prediction performance of (a) PairSeqSVM and (b) PairProSVM using full features or selected features with different values of k in Eq. 8.

Table 1. The average number of feature dimensions, overall accuracy, MCC, and recognition time for pairwise scoring features with or without feature selection. $\boldsymbol{\phi}'$ and $\boldsymbol{\phi}$ ($\boldsymbol{\zeta}'$ and $\boldsymbol{\zeta}$) denote feature vectors derived from sequence- and profile-alignment scores with (without) feature selection, respectively. k is the user-defined parameter in Eq. 8, which controls the number of features to be selected. *Feature Dimension* denotes the number of selected features, average over the 4 classes. The penalty factor C and scaling factor σ^2 in the RBF-SVM were set to 100 and 0.5, respectively. For the linear SVM, the value of C was set to 1. All results are based on 5-fold cross validation runs.

Classifier	Feature	k	Feature Dimension	Accuracy (%)	MCC	Rec. Time (sec.)
Linear SVM	$\boldsymbol{\zeta}'$	N/A	1942	87.9	0.79	89.9
RBF-SVM	$\boldsymbol{\phi}'$	0	730	85.0	0.77	30.0
RBF-SVM	$\boldsymbol{\phi}'$	1	245	80.8	0.71	6.1
RBF-SVM	$\boldsymbol{\phi}'$	2	94	75.2	0.62	3.1
RBF-SVM	$\boldsymbol{\phi}'$	3	39	64.9	0.48	2.1
Linear SVM	$\boldsymbol{\zeta}$	N/A	1942	99.4	0.99	18.3
RBF-SVM	$\boldsymbol{\phi}$	0	595	97.4	0.96	21.3
RBF-SVM	$\boldsymbol{\phi}$	1	244	96.4	0.94	4.2
RBF-SVM	$\boldsymbol{\phi}$	2	110	95.5	0.93	1.4
RBF-SVM	$\boldsymbol{\phi}$	3	37	87.1	0.81	1.1

symmetric and diagonally dominant property of pairwise scoring matrices allows the most important features in the pairwise scoring vectors to be selected independently. Experimental evaluation on a benchmark protein sequence dataset shows that the proposed feature selection scheme can reduce the feature dimension from thousands to hundreds, making subsequent classification much easier

and robust. With just a small reduction in recognition accuracy, a substantial speed up in recognition time can be achieved.

References

1. 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.*, 25:3389–3402, 1997.
2. B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res.*, 31:365–37, 2003.
3. S. Busuttil, J. Abela, and G.J. Pace. Support vector machines with profile-based kernels for remote protein homology detection. *Genome Informatics*, 15(2):191–200, 2004.
4. J. Guo, M.W. Mak, and S.Y. Kung. Eukaryotic protein subcellular localization based on local pairwise profile alignment SVM. In *IEEE Workshop on Machine Learning for Signal Processing*, 2006.
5. S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.*, pages 10915–10919, 1992.
6. http://www.eie.polyu.edu.hk/∼mwmak/BSIG/PairProSVM.htm.
7. T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, 7:95–114, 2000.
8. J.K. Kim, G.P.S Raghava, S.Y. Bang, and S. Choi. Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine. *Pattern Recog. Lett.*, 2006.
9. R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. Profile-based string kernels for remote homology detection and motif extraction. *J. Bioinfom. Comput. Biol.*, 3:527–550, 2005.
10. C.S. Leslie, E. Eskin, A. Cohen, J. Weston, and W.S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 2004.
11. L. Liao and W.S. Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.*, 10(6):857–868, 2003.
12. B.W. Matthews. Comparison of predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451, 1975.
13. H. Rangwala and G. Karypis. Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, 21(23):4239–4247, 2005.
14. A. Reinhardt and T. Hubbard. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, 26:2230–2236, 1998.
15. L Rychlewski, B Zhang, and A. Godzik. Fold and function predictions for *mycoplasma genitalium* proteins. *Fold Des.*, 3(4):229–238, 1998.
16. T.F. Smith and M.S. Waterman. Comparison of biosequences. *Adv. Appl. Math.*, 2:482–489, 1981.