Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 18 (2005) 1019–1028
Neural Networks
www.elsevier.com/locate/neunet

2005 Special Issue

A novel approach to extracting features from motif content and protein composition for protein sequence classification ☆

Xing-Ming Zhao[a,b], Yiu-Ming Cheung[c], De-Shuang Huang[a,*]

[a] *Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei, Anhui Province 230031, China*
[b] *Department of Automation, University of Science and Technology of China, Hefei, Anhui Province 230026, China*
[c] *Department of Computer Science, Hong Kong Baptist University, Hong Kong, China*

Abstract

This paper presents a novel approach to extracting features from motif content and protein composition for protein sequence classification. First, we formulate a protein sequence as a fixed-dimensional vector using the motif content and protein composition. Then, we further project the vectors into a low-dimensional space by the Principal Component Analysis (PCA) so that they can be represented by a combination of the eigenvectors of the covariance matrix of these vectors. Subsequently, the Genetic Algorithm (GA) is used to extract a subset of biological and functional sequence features from the eigen-space and to optimize the regularization parameter of the Support Vector Machine (SVM) simultaneously. Finally, we utilize the SVM classifiers to classify protein sequences into corresponding families based on the selected feature subsets. In comparison with the existing PSI-BLAST and SVM-pairwise methods, the experiments show the promising results of our approach.

Keywords: Genetic algorithm; Motif content; Protein composition; Protein sequence classification; Support vector machine

1. Introduction

One core problem in computational biology is the annotation of new protein sequences with the structural and functional features. To deal with this problem, one way is to classify a new protein sequence into a certain known protein family based on sequence similarity so that the structural and functional features of the sequence can be easily identified. In the literature, a variety of approaches, e.g. PSI-BLAST (Altschul et al., 1997), profiles (Gribskov, McLachlan, & Eisenberg, 1987), position-specific weight matrices (Henikoff & Henikoff, 1994), and Hidden Marked Models (HMM) (Krogh, Brown, Mian, Sjolander, & Haussler, 1994) have been developed for protein sequence classification. However, most of these methods belong to generative approaches that build a model for a single protein family and then evaluate each candidate sequence to see how well it fits the model. If the 'fit' is beyond a pre-defined threshold value, the sequence will be classified into the family; otherwise, it is not. The drawback of the generative approaches is that only positive examples are used as training sets. In contrast, discriminative approaches utilize both positive and negative examples as training sets. Consequently, the protein classifier results from the discriminative methods are usually better than those from the generative methods. Recently, a discriminative method, namely, Support Vector Machine (SVM) (Vapnik, 1995), has been successfully applied to protein sequence classification and shown the superiority to the other methods (Ding & Dubchak, 2001; Jaakkola, Diekhans, & Haussler, 2000; Leslie, Eskin, Cohen, Weston, & Noble, 2004; Liao & Noble, 2003; Markowetz, Edler, & Vingron, 2003; Zhao, Huang, Cheung, Wang, & Huang, 2004). Nevertheless, the success of the SVM depends on the selection of features to represent each protein family.

Early work (Brennan & Matthews, 1989) has shown that some short regions of the protein sequences, namely motifs (also called pattern or signature hereinafter), are better conserved than the others during the evolution. These motifs are generally important for the function of a protein. By

☆ This work was partly supported by the NSF of China (Nos. 60472111 and 60405002), Faculty Research Grant of Hong Kong Baptist University with Project Number: FRG/02-03/II-40, and the Research Grant Council of Hong Kong SAR under Project HKBU 2156/04E.
* Corresponding author. Tel.: +86 551 559 1108; fax: +86 551 559 2751.
E-mail address: dshuang@iim.ac.cn (D.-S. Huang).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.07.002

focusing on the limited and highly conserved regions of the proteins, motifs can often reveal important information on functional and structural features of the proteins. For example, the motifs for most catalytic sites and binding sites are conserved over wider taxonomic distance and longer evolutionary time than the sequences of the proteins themselves (Ben-Hur & Brutlag, 2003). That is, motifs can often represent important functional regions of the proteins such as catalytic sites, binding sites, protein–protein interaction sites structural motifs, and so forth.

In fact, some pioneering researchers have investigated the motif content to generate feature vectors for protein classification (Ben-Hur & Brutlag, 2003; Wang, Schroeder, Dobbs, & Honavar, 2003). However, they used the motif content only to characterize the local sequence features. Actually, the classification results of protein sequences are affected by many factors including local features and global features, etc., among which the fundamental one is which features should be extracted from the data. Apparently, if the features are appropriately selected, the protein sequences will be well classified. Currently, there have been a lot of approaches proposed for feature selection, among which the Genetic Algorithm (GA) has shown the superiority to the other methods (Raymer, Punch, Goodman, Kuhn, & Jain, 2000; Yang & Honavar, 1998). Furthermore, some works have used the GA for feature selection in combination with the kernel methods. For example, Fröhlich, Chapelle, and Schölkopf (2003) used the GA technique for feature selection and train SVM. Eads et al. (2002) used GA and SVM for time series classification. Jong, Marchiori, and van der Vaart (2004) used GA and SVM for cancer detection. Friedrichs et al. (Friedrichs & Igel, 2004, 2005; Igel, 2005) used evolutionary algorithms for tuning parameters of SVM. Runarsson and Sigurdsson (2004) used evolutionary strategies (ES) for model selection for SVM. Recently, we have also proposed a hybrid GA/SVM technique for protein classification (Zhao et al., 2004), which performs well for selecting features in high dimensional space.

In this paper, we further improve this hybrid GA/SVM technique to select features from eigen-space and to optimize the regularization parameter of the SVM simultaneously. First, we formulate a protein sequence as a fixed-dimensional vector via the motif content and protein composition. Then, we further project the vectors into a low-dimensional space by the Principal Component Analysis (PCA) (Diamantaras & Kung, 1996) so that they can be represented by a combination of the eigenvectors of the covariance matrix of these vectors. Subsequently, the GA technique is used to extract a subset of biological and functional sequence features from the eigen-space and to optimize the parameters of the SVM simultaneously. Finally, we utilize the SVM classifiers to classify protein sequences into the corresponding families based on the selected feature subsets. In comparison with the existing methods such as PSI-BLAST (Altschul et al., 1997) and SVM-pairwise (Liao & Noble, 2003), the experiments show the outstanding performance of our approach.

The remainder of this paper is organized as follows. Section 2 describes the proposed approach in detail. The experimental results and discussions are reported in Section 3. Finally, a conclusion is drawn in Section 4.

2. A new approach to extracting features from protein sequences

In this section, we present a new method for extracting features from motif content and protein composition for protein sequence classification. Fig. 1 gives an overview of our proposed method: (1) The training protein sequences are first converted into eigenvectors by being projected into the

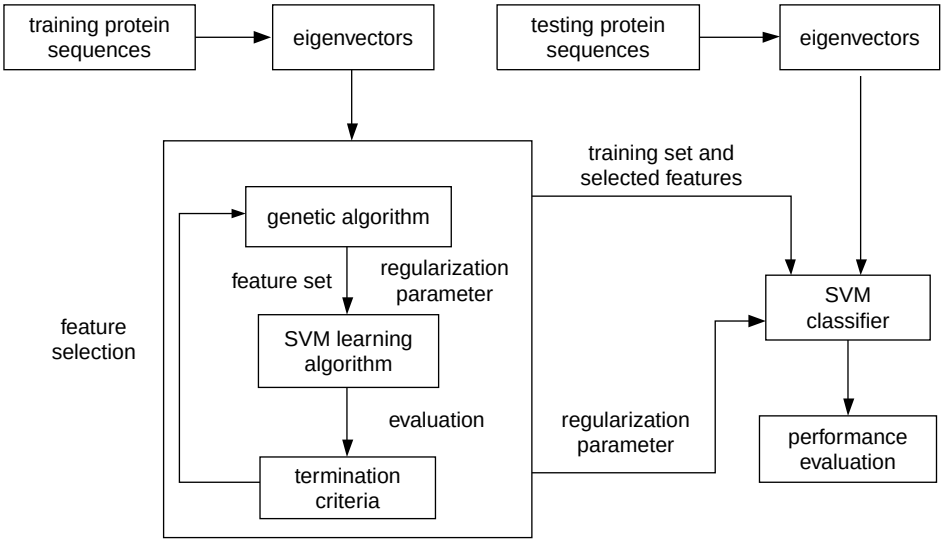

Fig. 1. The procedure of extracting features from motif content and protein composition for protein sequence classification.

eigen-space via PCA; (2) The hybrid GA/SVM technique is then used to select features from the eigen-space and to optimize the regularization parameter of SVM; (3) After acquiring a subset of features, the SVM classifier is utilized to classify an unseen set of protein sequences into corresponding family based on the selected features and optimized regularization parameter. The detailed procedure of extracting features from motif content and protein composition is given in the following sub-sections.

2.1. Vectorization of protein sequences

The first step of our proposed approach is to convert each protein sequence into a vector of fixed dimensionality based on the motif content and protein composition. The set of motifs to be used can be chosen from the existing motif database PROSITE (Falquet et al., 2002), which is a database of protein families and domains. In general, proteins can be grouped into a limited number of families based on the similarities between their sequences. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor. By analyzing the properties of such group of similar sequences, it is possible to derive a signature for a protein family or domain, which distinguishes its members from all other unrelated proteins. Generally, such a protein signature or pattern can be used to assign a new sequenced protein to a specific protein family and thus to induce its function.

PROSITE currently contains patterns and profiles specific for more than a thousand protein families or domains. Each of these patterns comes with documentation providing background information on the structure and function of these proteins. For example, the expression of the PROSITE pattern PS50020 is W−X(9,11)−[VFY]−[FYW]−X(6,7)−[GSTNE]−[GSTQCR]−[FYW]−X(2)−P, where X means any amino acids, [AB] means either A or B, A(6,7) means A appears 6–7 times consecutively. The documentation associated with the pattern PS50020 lists the proteins that contain the pattern and presents the information on the structure and function of the pattern. In this paper, the patterns extracted from the PROSITE database using a perl program, namely ps_scan, are used to describe each protein of interest. Consequently, each protein sequence is converted into an N-dimensional motif-based feature vector, where N is the total number of the motifs in the PROSITE database. Since the PROSITE updated database contains 1744 entries, the number of features will therefore be 1744. Each element of the vectors represents the presence or absence of a motif in the protein sequences. That is, the corresponding feature value will be 1 if a motif is present. Otherwise, it will be 0.

In addition to the motif content, the protein composition is also incorporated into the feature vectors to improve the classification accuracy. A protein sequence S is defined as a linear succession of 20 symbols from a finite alphabet Σ, where Σ consists of {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. Generally, the features describing the protein composition can be defined as

$$v_i = \frac{c_i}{\sum_{j=1}^{20} c_j}, \quad i = 1, \ldots, 20, \tag{1}$$

where v_i is the value for the ith feature, and c_i is the times of the ith amino acid occurring in the given protein sequence. Hence, we can get 20 features for the 20 amino acids in the protein sequences, respectively. Each element in the feature vector denotes the presence frequency of an amino acid. Consequently, the number of features will be 1744 + 20 = 1764 in total.

Although the dimension of the feature vectors is very high, the vectors contain relatively a much smaller number of non-zero features. Edler, Grassmann, and Suhai (2001) have used the PCA technique to determine an appropriate subset of principal components to represent most information of protein sequences for protein fold classification. Here, the PCA technique is also used to reduce the dimensionality and normalize the extracted features. We project a protein into a low-dimensional eigen-space via PCA and then represent it as an eigenvector containing the coefficients of the projection.

2.2. Selecting features from the eigen-space via hybrid GA/SVM

The role of PCA is to reduce the dimensionality and normalize the extracted features. However, it is usually difficult to know how many components should be chosen. Although various rules (Jolliffe, 1986) have been developed to determine the number of principal components, most of them are ad hoc and subjective. Under the circumstances, we first select a large dimensionality of 1000, and then project the protein sequences into the eigen-space via PCA. Subsequently, a subset of features is automatically selected by the hybrid GA/SVM technique.

The basic idea behind the hybrid GA/SVM technique is that GA is used to select features and optimize the hyper-parameter of SVM, while SVM is used to evaluate the selected features. In the hybrid GA/SVM algorithm, a population contains chromosomes that are the potential solutions to feature selection and to optimization of the hyper-parameter of SVM. A fitness function is associated with each chromosome to measure the degree of goodness of the chromosome, where SVM is used to calculate the classification accuracy based on the selected features and optimized hyper-parameter. By applying the genetic operators, i.e. selection, crossover, and mutation, to the chromosomes in the population, a new population is generated for the next generation. The procedure of selection, crossover and mutation is repeated until a termination criteria is satisfied, i.e. one of the generations

contains solutions that are good enough. Suppose the selected feature set is F and the hyper-parameter of SVM is C. The procedure of the hybrid GA/SVM algorithm is summarized in Fig. 2.

It can be seen that the key issues of the hybrid GA/SVM algorithm are the encoding of the chromosomes and the definition of the fitness function. Furthermore, the SVM algorithm is also involved. In the following, we will describe them in more detail.

2.2.1. Support vector machine

In this section, an overview of SVM is presented. Given a set of labelled training pairs (x_i,y_i), $i=1,\ldots,l$, where $x_i \in \mathbb{R}^n$, $y_i \in \{-1,+1\}^l$, SVM maps the input vector x_i into a high dimensional feature space $\mathbb{R}^{nh}$ by a mapping function $\Phi(\cdot)$ and finds a hyperplane, which maximizes the margin, i.e. the distance between the hyperplane and the nearest data points of each class in the space $\mathbb{R}^{nh}$. The decision function implemented by SVM can be written as

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b\right) \tag{2}$$

```
Hybrid GA/SVM()
{
  t=0;
  initialize population p(0);
  evaluate p(0);
  repeat
  {
    t=t+1;
    select p(t) from p(t − 1);
    perform crossover on p(t);
    mutate p(t);
    evaluate p(t);
  }until(termination criteria);
  return feature set F and hyper-parameter C;
}
```

Fig. 2. Hybrid GA/SVM algorithm, where the tournament selection operator and uniform crossover operator are used with the population p evaluated by SVM.

with

$$K(x_i, x) = \langle \Phi(x_i), \Phi(x) \rangle, \tag{3}$$

where α_is are the coefficients obtained by solving the following optimization problem

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{l} \alpha_i - 1/2 \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \\ \text{subject to} \quad & 0 \le \alpha_i \le C, \\ & \sum_{i=1}^{l} \alpha_i y_i = 0, \quad i = 1, 2, \ldots, l, \end{aligned}$$

where C is a regularization parameter, which controls the trade-off between the margin and the misclassification error.

To construct the SVM classifier, we need to determine the following two parameters: the regularization parameter C and the kernel function K. In this paper, the Gaussian kernel is adopted for all the SVM classifiers and C is optimized by GA. The variance of the Gaussian kernel is computed as the median Euclidean distance from any positive training examples to the nearest negative example (Hou, Hsu, Lee, & Bystroff, 2003).

2.2.2. Chromosome representation

In this paper, the chromosome is encoded into a bit string. As shown in Fig. 3, a chromosome is partitioned into two parts: Parts A and B. Part A represents the selected features. Let m ($m=1000$ in our case) be the total number of features, and Part A is represented by a binary vector of dimension m (Fig. 3). If the ith bit of the vector is equal to 1, the corresponding ith feature is selected; Otherwise, the corresponding ith feature will not be selected. In addition, to train an optimized SVM, C should also be selected by GA. Part B represents C using 4 bits to encode the numbers of $-7,\ldots,8$, which are used as the powers of 10. Hence, C can be selected in the discrete value of $10^{-7},\ldots,10^{8}$. For example, the value for C in Fig. 3 is 10^6.

2.2.3. Fitness function for chromosome evaluation

The goal of feature selection is to use fewer features to achieve the same or better performance compared with that obtained using the complete feature set. Hence, chromosome evaluation contains the following two objectives: (1) minimizing the number of features; (2) maximizing the classification accuracy. Obviously, there are some trade-offs between the accuracy and the number of features, among which the accuracy is our major concern. In the literature (Deb & Reddy, 2003a,b, 2004), there have been many methods proposed for combining the above two terms. In this paper, a simple weighting method using linear aggregation of the two objectives is adopted. Given a chromosome g, the fitness function can be defined as

$$f(g) = f_1(g) + w f_2(g), \tag{4}$$

where w is the weighting coefficient, $f_1(g)$ is the recognition

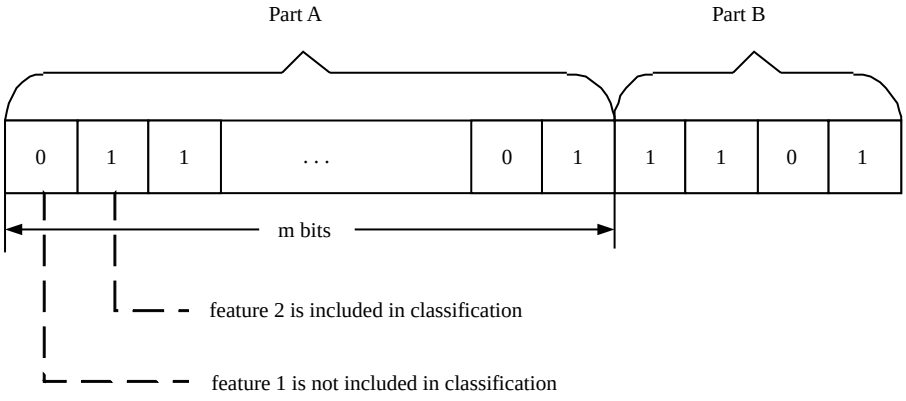

Fig. 3. The binary string that comprises a chromosome of the GA population. Part A represents a potential solution to the feature selection problem, and Part B represents the optimization value of C of SVM.

rate obtained using features presented in g, and $f_2(g)$ is the number of features removed from the original feature set. A small value is given to w because we mainly focus on the recognition rates. In this paper, w is set at 4×10^{-5}. It can be seen that the chromosomes with higher accuracies will outweigh those with lower accuracies, no matter how many features they contain.

In the hybrid GA/SVM algorithm, the recognition rate is obtained by the SVM classifier, where the five-fold cross-validation is used to estimate the classification performance. The training set $\mathbf{T}$ is randomly partitioned into five equal disjoint sets: $\mathbf{T}_1$, $\mathbf{T}_2$, $\mathbf{T}_3$, $\mathbf{T}_4$ and $\mathbf{T}_5$. Consequently, five SVM classifiers are trained on the complement $\bar{\mathbf{T}}_i$ of each partition $\mathbf{T}_i$, and each classifier is then tested on the corresponding unseen test set $\mathbf{T}_i$.

The final cross-validation recognition rate is given by

$$f_1 = \frac{1}{5}\sum_{i=1}^{5} r(\mathbf{T}_i, \bar{\mathbf{T}}_i), \tag{5}$$

where $r(\mathbf{T}_i, \bar{\mathbf{T}}_i)$ is the recognition rate on $\mathbf{T}_i$ using the SVM classifier trained on $\bar{\mathbf{T}}_i$.

2.2.4. Genetic operators

In this work, the tournament selection is adopted to select two parent chromosomes from the current population. Then, the uniform crossover method is applied to the two parent binary string vectors to produce two offsprings, and the mutation operation mutates the offsprings.

If the mutated chromosome is superior to both parent chromosomes, it replaces the similar one; if it is in between the two parents, it replaces the inferior one; otherwise, the most inferior one in the population is replaced. The procedure of selection, crossover and mutation is repeated until a termination criteria is satisfied.

3. Experiments and discussions

In this section, we investigated our proposed method on classifying the protein sequences obtained from the Protein Information Resource (PIR) (Barker et al., 2000) database and Structural Classification of Protein (SCOP) (Murzin, Brenner, Hubbard, & Chothia, 1995) database. The parameters used in GA were: population size 50, generation 1000, crossover rate 0.8, and mutation rate 0.02.

3.1. Experiment I

In this experiment, protein sequences obtained from the PIR database were used to evaluate our proposed technique. Six protein superfamilies data were used as positive datasets and one protein superfamily used as negative dataset. Both the six positive datasets and the one negative dataset were obtained from the PIR Non-Redundant Reference Sequence Database (PIR-NREF), Release 1.35. The six positive protein superfamilies were: Cytochrome C (964), Ferredoxin (312), Plastocyanin (132), Triose (195), Ligase (678) and Lectin (159), respectively. The one negative dataset containing 759 sequences was Cytochrome b. In each of the seven datasets, two-third of the dataset was used as the training set and input for the hybrid GA/SVM algorithm to get a subset of features, and the rest one-third was held as the test set. The hybrid GA/SVM classifier was trained on each positive dataset against the negative dataset to find a subset of features so that we can discriminate the positive dataset from the negative dataset. This procedure was repeated for six times for the six positive families.

We first investigated whether or not using the combination of motif content and protein composition can get better results than using only the motif content. Six kinds of combination were therefore compared: Classification using motif content by SVM; Classification using composition by SVM; Classification using the combination of motif content and protein composition by SVM; Classification using the combination of motif content and protein composition by GA+SVM; Classification using the motif content by PCA+GA+SVM; Classification using the combination of motif content and protein composition by PCA+GA+SVM. Here, PCA+GA+SVM represents our proposed

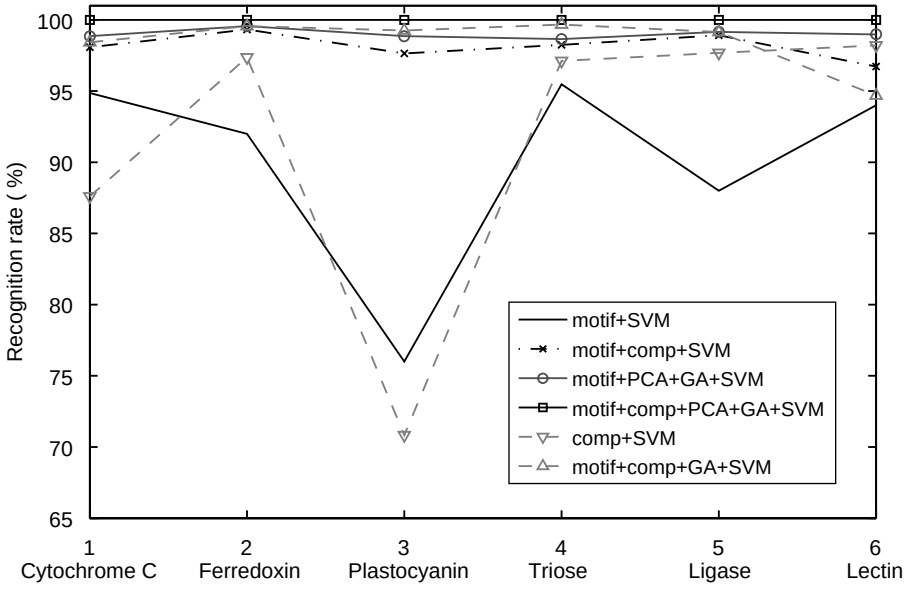

Fig. 4. Comparison of the performance for different kinds of combination of classifiers and feature vectors on the test sets of the six protein superfamilies.

method, and GA/SVM means the hybrid GA/SVM technique without PCA used. Fig. 4 shows the results obtained by the six methods on the test sets of the six protein families. It can be seen that, using both motif content and protein composition, we can get better classification accuracy than using only the motif content. Further, using PCA to reduce the dimensionality and normalize the features, we can get better results than those obtained without PCA. It can also be concluded that using the feature subsets generated from the motif content and protein composition, the classification accuracy can be indeed improved.

Furthermore, we investigated the influence of the weight value w in Eq. (4) on the classification performance. We set w at 4×10^{-5}, 0 and 4×10^{-4}, respectively. Fig. 5 shows the recognition rates and feature numbers obtained on the test sets using different ws. As shown in Fig. 5, when w was set at 0, we can get the same recognition rates as those obtained by setting w at 4×10^{-5}, but the number of features was larger; when w was set at 4×10^{-4}, the number of features was the smallest but the recognition rate was the worst. Therefore, it can be concluded that the larger the value of w is, the smaller the set of features is selected and the worse the recognition rates are obtained. It can be learned from Fig. 5 that the weight coefficient we proposed in this paper is reasonable for protein sequence classification.

Table 1 summarizes the comparison of the recognition rates and feature numbers before and after the dimensionality reduction for the SVM classifier on the test sets of the six superfamilies. It can be seen that using the feature subsets, the feature numbers have been significantly reduced by 52.9–62.3% compared with the complete eigen-feature sets, and the classification accuracies for the given protein sequences can be indeed improved.

In addition, we compared our proposed method with the other three protein sequence classification methods: the C4.5 decision tree algorithm (Quinlan, 1993), the BLAST

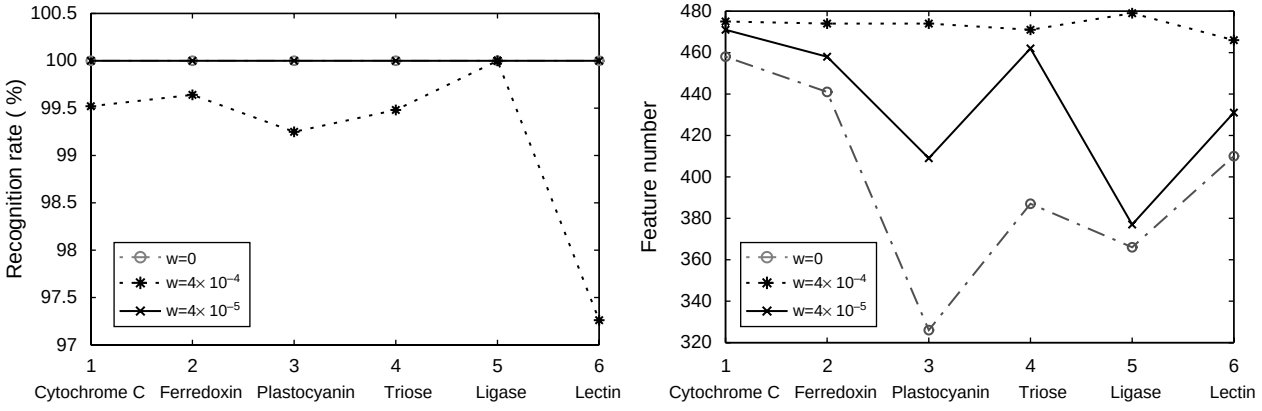

Fig. 5. The results obtained on the test sets of the six protein superfamilies using different w values. The left figure shows the recognition rates obtained using different w values. The right figure shows the number of features obtained using different w values.

Table 1
Comparison of the recognition rates and feature numbers before and after the dimensionality reduction for the SVM classifier on the test sets of the six protein superfamilies

Protein datasets	Original feature set		Selected feature subset	
	Feature number	Classification accuracy (%)	Feature number	Classification accuracy (%)
Cytochrome *c*	1764	98.08	471	100.00
Ferredoxin	1764	99.32	458	100.00
Plastocyanin	1764	97.64	409	100.00
Triose	1764	98.25	462	100.00
Ligase	1764	98.92	377	100.00
Lectin	1764	96.72	431	100.00

(Altschul et al., 1997) method and the HMMer (Durbin, Eddy, Krogh, & Mitchison, 1998) method. Table 2 summarizes the results obtained by the four classifiers on the test sets of the six protein superfamilies, where the C4.5 decision tree algorithm used only the motif content as inputs as described in (Wang et al., 2003). From Table 2, it can be readily found that our proposed hybrid GA/SVM technique based on motif content and protein composition outperforms the decision tree algorithm, the BLAST and the HMMer.

3.2. Experiment II

In this section, the SCOP 1.53 data set presented by Liao and Noble (2003) was used to evaluate the performance of our proposed method in comparison with the other methods. Sequences were selected using the Astral database (Brenner, Koehl, & Levitt, 2000) with an *E*-value threshold of 10^{-25}. This procedure resulted in 4352 distinct sequences that were grouped into 54 families. We compared our proposed technique with four other methods: PSI-BLAST (Altschul et al., 1997), SAM (Krogh et al., 1994), SVM-pairwise (Liao & Noble, 2003) and SVM-Fisher (Jaakkola et al., 2000). The same experiment setup described by Liao (Liao & Noble, 2003) was used so that we can compare our proposed method directly with the other four methods.

For PSI-BLAST, a sequence from the positive training set was randomly selected to serve as the initial query sequence. The resulting profiles were used to search against the test set sequences, and the resulting *E*-values were used to rank the testing sequences. For the SAM method, the hidden Markov models were trained using the Sequence Alignment and Modelling (SAM) toolkit (Krogh et al., 1994). After a hidden Markov model was obtained, the testing sequences can be compared against the model, and the resulting *E*-values were used to rank the testing sequences.

Table 2
Comparison of the recognition rates for the four classifiers on the test sets of the six superfamilies

Protein datasets	SVM (%)	Decision tree (%)	BLAST (%)	HMMer (%)
Cytochrome *c*	100.00	98.43	96.16	82.2
Ferredoxin	100.00	99.44	100.00	96.92
Plastocyanin	100.00	99.33	87.98	97.64
Triose	100.00	99.46	98.89	91.25
Ligase	100.00	99.17	99.25	89.98
Lectin	100.00	99.35	99.35	96.74

For SVM-pairwise and SVM-Fisher, the difference is how to represent the protein sequences as feature vectors. The SVM-Fisher method used the feature vectors generated from the parameters of a profile HMM, while the SVM-pairwise method used the pairwise Smith–Waterman sequence similarity scores to represent the protein sequences. The regularization parameter for both SVM-pairwise and SVM-fisher was set at 10, and the Gaussian kernel was adopted for both methods.

In this paper, the receiver operating characteristic (ROC) scores (Gribskov et al., 1987) and the median rate of false positives (RFP) (Jaakkola et al., 2000) were used to evaluate the performance of all methods. The experimental results were summarized in Figs. 6 and 7, where a higher curve corresponds to a more accurate classification result. As shown in Figs. 6 and 7, using either performance measure, our proposed method outperforms the other four methods. Table 3 summarizes the final feature numbers, ROC scores and median RFP scores on the test sets. As shown in Table 3, after generations of evolution, the number of the features in the final iteration ranges from 431 to 500 for the 54 protein families, reducing the features about 50–56.9% compared with the complete eigen-feature sets. It can be concluded from Figs. 6, 7 and Table 3 that using the information extracted from the motif content and protein composition, the feature subsets extracted from the eigen-space can provide concise representation of corresponding protein families, i.e. our proposed method can find biological significant features for protein classification. This also implies that the feature selection really plays an important role in protein classification.

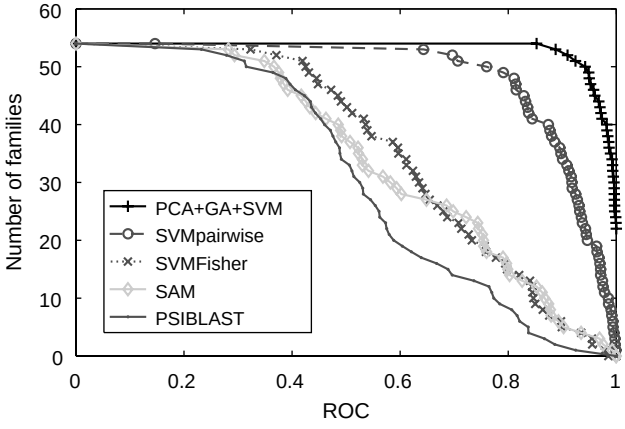

Fig. 6. Comparison of the performance of the five methods on the 54 protein families, where the curves show the number of families versus a ROC score threshold for the five methods.

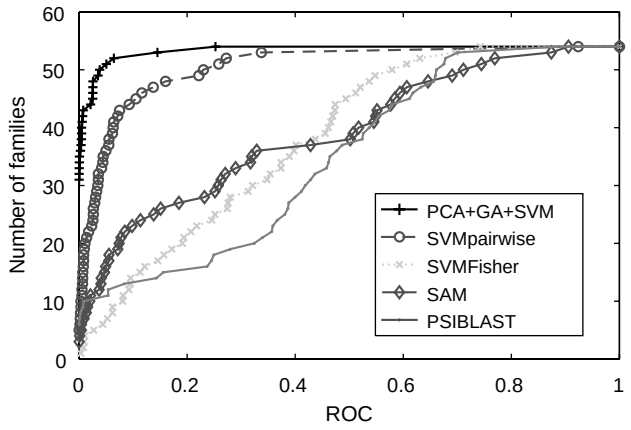

Fig. 7. Comparison of the performance of the five methods on the 54 protein families, where the curves show the number of families versus a median RFP score threshold for the five methods.

In addition, the relative performance of our proposed method was further illustrated in Fig. 8. The left figure in Fig. 8 shows the family-by-family comparison of the 54 ROC scores for our proposed technique versus the SVM-pairwise method. As shown in the left figure of Fig. 8, our proposed technique scores higher than the SVM-pairwise method on every protein family in terms of ROC scores. Further, the right figure of Fig. 8 shows the family-by-family comparison of the 54 median RFP scores for our method versus the SVM-pairwise method. As shown in the right figure of Fig. 8, our proposed method scores higher than the SVM-pairwise method on almost all families in terms of median RFP scores. In other words, using the features extracted from the eigen-space, our proposed technique outperforms other methods in terms of protein classification.

To get an idea about the optimal set of eigenvectors selected by the hybrid GA/SVM technique. We plotted the distribution curves showing the distribution of the selected eigen-features in the training sets and test sets. We took the protein family of 7.3.6.4 as an example. The 7.3.6.4 family contains 3591 negative training sequences, 485 negative test sequences, 37 positive training sequences, and five positive test sequences. Fig. 9 shows the distribution of the final 500 features selected by the hybrid GA/SVM technique from the eigen-space formed by PCA, where x-axis corresponds to the eigen-features and y-axis corresponds to the times of an eigen-feature occurring in the training or test sets. As shown in Fig. 9, the distribution of the eigen-features in the positive test set is consistent with that in the positive training set, while the distribution of the eigen-features in the negative training set is consistent with that in the negative test set. This scenario further implies the reason that our classifiers can achieve better performance.

4. Conclusion

This paper has proposed a new method for classifying

Table 3
The final feature numbers, ROC scores and median RFP scores for the test sets of the 54 protein families

ID	Feature number	ROC score	Median RFP	ID	Feature number	ROC score	Median RFP
1.27.1.1	462	1	0	2.9.1.4	458	1	0
1.27.1.2	461	0.9735	0.0056	3.1.8.1	475	0.9867	0.0042
1.36.1.2	441	1	0	3.1.8.3	469	1	0
1.36.1.5	450	1	0	3.2.1.2	464	0.8878	0.2525
1.4.1.1	457	0.9968	0	3.2.1.3	451	0.9676	0.0065
1.4.1.2	455	0.8525	0.0256	3.2.1.4	474	1	0
1.4.1.3	461	0.9823	0.0004	3.2.1.5	477	0.9525	0.0023
1.41.1.2	462	1	0	3.2.1.6	485	0.9699	0.0252
1.41.1.5	466	0.9970	0.0002	3.2.1.7	456	1	0
1.45.1.2	431	1	0	3.3.1.2	455	0.9893	0.0047
2.1.1.1	463	0.9945	0.0003	3.3.1.5	465	1	0.0036
2.1.1.2	487	0.9958	0	3.32.1.1	467	0.9843	0.0077
2.1.1.3	474	0.9996	0.0213	3.32.1.11	456	0.9245	0.0254
2.1.1.4	468	0.9990	0	3.32.1.13	463	0.9099	0.0045
2.1.1.5	461	0.9603	0.0007	3.32.1.8	451	0.9716	0
2.28.1.1	476	1	0	3.42.1.1	458	0.958	0.0505
2.28.1.3	460	0.9963	0	3.42.1.5	458	0.9934	0
2.38.4.1	455	0.9862	0.038	3.42.1.8	466	0.9425	0.0256
2.38.4.3	483	0.9490	0.145	7.3.10.1	467	0.9965	0
2.38.4.5	438	0.9922	0	7.3.5.2	471	1	0
2.44.1.2	458	0.95	0.0634	7.3.6.1	482	1	0
2.5.1.1	462	1	0	7.3.6.2	487	1	0
2.5.1.3	437	1	0	7.3.6.4	500	1	0
2.52.1.2	475	0.9961	0	7.39.1.2	451	0.95208	0
2.56.1.2	473	1	0	7.39.1.3	462	0.9931	0
2.9.1.2	482	0.9827	0.035	7.41.5.1	458	1	0
2.9.1.3	452	1	0	7.41.5.2	457	1	0

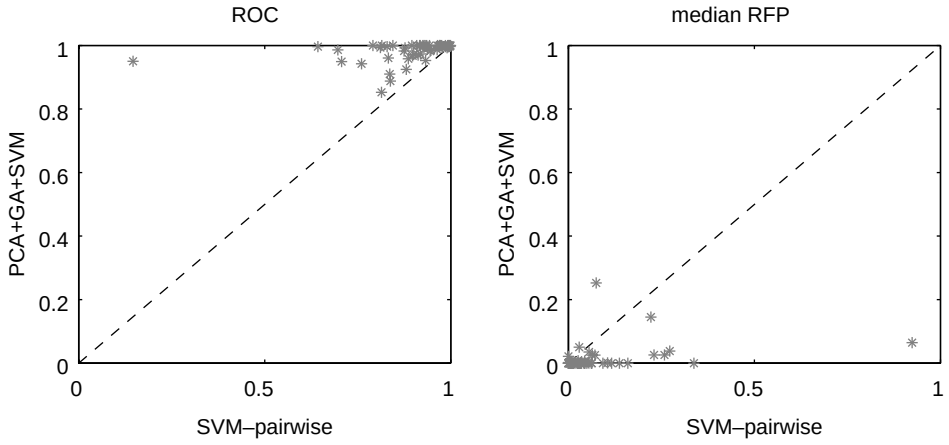

Fig. 8. Comparison of the performance of PCA+GA+SVM versus SVM-pairwise. The left figure shows the family-by-family comparison of the 54 ROC scores. The right figure shows the family-by-family comparison of the 54 median RFP scores.

protein sequences into functional families using features extracted from the motif content and protein composition. Compared with the previous works, the main novelty of our method is that protein sequences are converted into feature vectors using global features and local features represented by the protein composition and motif content, respectively. Furthermore, the PCA technique is also utilized to normalize the feature vectors and project the feature vectors into a low-dimensional eigen-space. Instead of using the complete set of eigenvectors, we select a much smaller feature subset to represent the protein sequences of interest via the hybrid GA/SVM technique, where the hyper-parameter of SVM is also optimized by GA. Having obtained the subset of features, the SVM classifier is utilized to classify protein sequences into corresponding families based on the optimized hyper-parameter. The numerical results on protein sequences obtained from PIR and SCOP databases have shown that features extracted from the motif content and protein composition by our proposed technique are really effective for protein classification.

Acknowledgements

The authors would like to thank Dr Bingyu Sun for helpful discussion. The authors would also like to thank the anonymous reviewers for constructive suggestions.

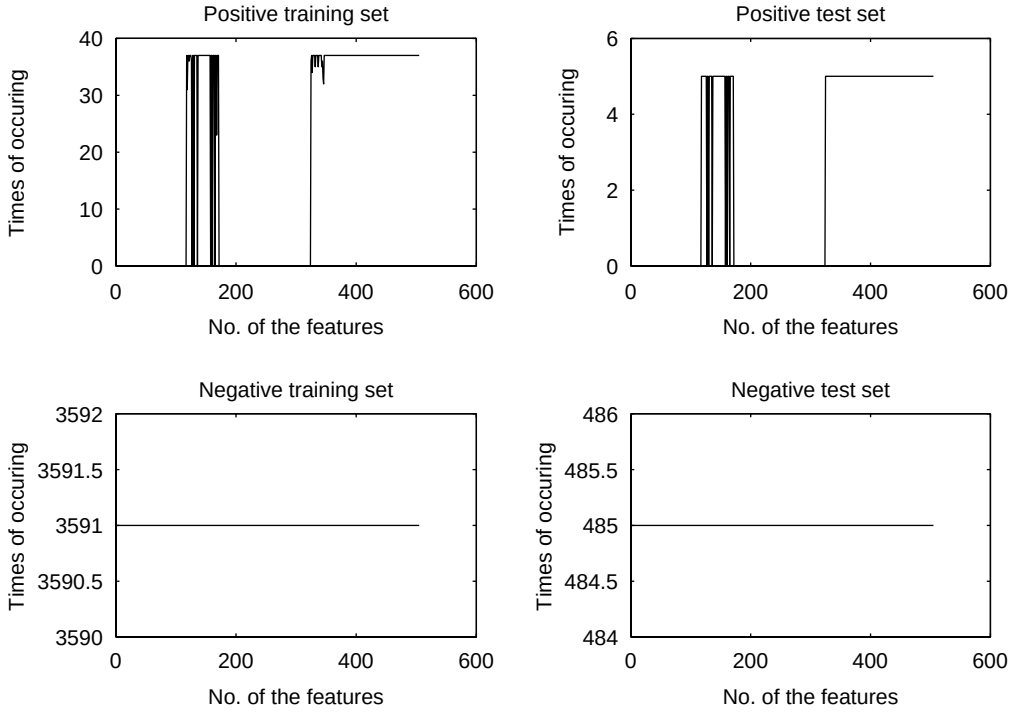

Fig. 9. The distribution of the selected features in the training sets or test sets of 7.3.6.4 protein family.

References

Altschul, S., Madden, T., Schafer, A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped blast and psi-blast: A new generation of protein data. *Nucleic Acids Research*, *25*, 3389–3402.

Barker, W., Garavelli, J., Huang, H., McGarvey, P., Orcutt, B., Srinivasarao, G., et al. (2000). The protein information resource (PIR). *Nucleic Acids Research*, *28*, 41–44.

Ben-Hur, A., & Brutlag, D. (2003). Remote homology detection: A motif based approach. *Bioinformatics*, *19*, 26–33.

Brennan, R., & Matthews, B. (1989). The helix-turn-helix DNA binding motif (minireview). *Journal of Biological Chemistry*, *264*, 1903–1906.

Brenner, S., Koehl, P., & Levitt, M. (2000). The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Research*, *28*, 254–256.

Deb, K., & Reddy, A. (2003). Classification of two and multi-class cancer data reliably using multi-objective evolutionary algorithms. *KanGAL Report, No. 2003006*.

Deb, K., & Reddy, A. (2003). Classification of two-class cancer data reliably using evolutionary algorithms. *KanGAL Report, No. 2003001*.

Deb, K., & Reddy, A. R. (2004). Large-scale scheduling of casting sequences using a customized genetic algorithm. *Lecture Notes in Computer Science*, *2936*, 141–152.

Diamantaras, I., & Kung, S. (1996). *Principal component neural networks, theory and applications*. London: Wiley.

Ding, C., & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, *17*, 349–358.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Eads, D. R., Hill, D., Davis, S., Perkins, S. J., Ma, J., Porter, R. B., et al. (2002). Genetic algorithms and support vector machines for time series classification. *Proceedings of SPIE*, *4787*, 74–85.

Edler, L., Grassmann, J., & Suhai, S. (2001). Role and results of statistical methods in protein fold class prediction. *Mathematical and Computer Modelling*, *33*, 1401–1417.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C., Hofmann, K., et al. (2002). The PROSITE database, its status in 2002. *Nucleic Acids Research*, *30*, 235–238.

Friedrichs, F., & Igel, C. (2004). *Evolutionary tuning of multiple SVM parameters. Proceedings of the 12th European Symposium on Artificial Neural Networks, Bruges, Belgium* pp. 519–524.

Friedrichs, F., & Igel, C. (2005). Evolutionary tuning of multiple SVM parameters. *Neurocomputing*, *64*, 107–117.

Fröhlich, H., Chapelle, O., & Schölkopf, B. (2003). *Feature selection for support vector machines by means of genetic algorithms. Proceedings of the 15th IEEE international conference on tools with artificial intelligence, Sacramento, California, USA* pp. 142–148.

Gribskov, M., McLachlan, A., & Eisenberg, D. (1987). *Profile analysis: Detection of distantly related proteins. Proceedings of the national academy of sciences of the USA*, Vol. 84 pp. 4355–4358.

Henikoff, S., & Henikoff, J. (1994). Position-based sequence weights. *Journal of Molecular Biology*, *243*, 574–578.

Hou, Y., Hsu, W., Lee, M. L., & Bystroff, C. (2003). Efficient remote homology detection using local structure. *Bioinformatics*, *19*, 2294–2301.

Igel, C. (2005). *Multi-objective model selection for support vector machines Proceedings of the third international conference on evolutionary multi-criterion optimization, Mexico* pp. 534–546.

Jaakkola, T., Diekhans, M., & Haussler, D. (2000). A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, *7*, 95–114.

Jolliffe, I. (1986). *Principal component analysis*. New York: Springer-Verlag.

Jong, K., Marchiori, E., & van der Vaart, A. (2004). Analysis of proteomic pattern data for cancer detection. *Lecture Notes in Computer Science*, *3005*, 41–51.

Krogh, A., Brown, M., Mian, I., Sjolander, K., & Haussler, D. (1994). Hidden markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, *235*, 1501–1531.

Leslie, C., Eskin, E., Cohen, A., Weston, J., & Noble, W. S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics*, *20*, 467–476.

Liao, L., & Noble, W. S. (2003). Combining pairwise sequence similarity and support vector machines for detecting remoteprotein evolutionary and structural relationships. *Journal of Computational Biology*, *10*, 857–868.

Markowetz, F., Edler, L., & Vingron, M. (2003). Support vector machines for protein fold class prediction. *Biometrical Journal*, *45*(3), 377–389.

Murzin, A., Brenner, S., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of protein databases for investigation of sequences and structures. *Journal of Molecular Biology*, *247*, 536–540.

Quinlan, J. R. (1993). *C4.5: Programs for empirical learning*. San Francisco, CA: Morgan Kaufmann.

Raymer, M., Punch, W., Goodman, E., Kuhn, L., & Jain, A. (2000). Dimensionality reduction using genetic algorithms. *IEEE Transaction on Evolutionary Computation*, *4*, 164–171.

Runarsson, T. P., & Sigurdsson, S. (2004). Asynchronous parallel evolutionary model selection for support vector machines. *Neural Information Processing—Letters and Reviews*, *3*(3), 59–68.

Vapnik, V. (1995). *The nature of statistical learning theory*. Berlin: Springer-Verlag.

Wang, X., Schroeder, D., Dobbs, D., & Honavar, V. (2003). Automated data-driven discovery of motif-based protein function classifiers. *Information Sciences*, *155*, 1–18.

Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, *13*(1), 44–49.

Zhao, X. M., Huang, D. S., Cheung, Y. M., Wang, H. Q., & Huang, X. (2004). A novel hybrid GA/SVM system for protein sequences classification. *Lecture Notes in Computer Science*, *3177*, 11–16.