BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 23 2005, pages 4223–4229
doi:10.1093/bioinformatics/bti697

Sequence analysis

Discrimination of outer membrane proteins using support vector machines

Keun-Joon Park[1,2], M. Michael Gromiha[1,*], Paul Horton[1] and Makiko Suwa[1]

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2–42 Aomi, Koto-ku, Tokyo 135–0064, Japan and [2]Laboratory of Functional Analysis *in silico*, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai Minato-ku, Tokyo 108–8639, Japan

Received on June 27, 2005; revised and accepted September 27, 2005
Advance Access publication October 4, 2005

ABSTRACT

Motivation: Discriminating outer membrane proteins from other folding types of globular and membrane proteins is an important task both for dissecting outer membrane proteins (OMPs) from genomic sequences and for the successful prediction of their secondary and tertiary structures.

Results: We have developed a method based on support vector machines using amino acid composition and residue pair information. Our approach with amino acid composition has correctly predicted the OMPs with a cross-validated accuracy of 94% in a set of 208 proteins. Further, this method has successfully excluded 633 of 673 globular proteins and 191 of 206 α-helical membrane proteins. We obtained an overall accuracy of 92% for correctly picking up the OMPs from a dataset of 1087 proteins belonging to all different types of globular and membrane proteins. Furthermore, residue pair information improved the accuracy from 92 to 94%. This accuracy of discriminating OMPs is higher than that of other methods in the literature, which could be used for dissecting OMPs from genomic sequences.

Availability: Discrimination results are available at http://tmbeta-svm.cbrc.jp

Contact: michael-gromiha@aist.go.jp

INTRODUCTION

Outer membrane proteins (OMPs) perform a variety of functions, such as mediating non-specific, passive transport of ions and small molecules, selectively allowing the passage of molecules such as maltose and sucrose (Schirmer *et al*., 1995; Forst *et al*., 1998; Schulz, 2000; Wimley, 2003) and are involved in voltage-dependent anion channels (Mannella, 1998). These proteins contain β-strands as their membrane spanning segments and are found in the outer membranes of bacteria, mitochondria and chloroplasts (Schulz, 2002). A comparative analysis on the distribution of amino acid residues in α-helical and β-barrel membrane proteins shows that the membrane part of OMPs is more complex than that of transmembrane helical proteins due to the intervention of many charged and polar residues in the membrane (Gromiha *et al*., 1997; Gromiha, 1999). Consequently, the success rate of discriminating β-barrel membrane proteins from other proteins is significantly lower than that of α-helical membrane proteins (Hirokawa *et al*., 1998; Chen and Rost, 2002).

Recently, several methods have been proposed for discriminating OMPs from amino acid sequences (Gnanasekaran *et al*., 2000; Wimley, 2002; Martelli *et al*., 2002; Liu *et al*., 2003; Bagos *et al*., 2004; Natt *et al*., 2004; Garrow *et al*., 2005). Gnanasekaran *et al*. (2000) developed a structure-based sequence alignment method for discriminating β-stranded membrane proteins and reported an accuracy of 80%. Wimley (2002) analyzed the architecture of 15 OMPs and proposed a method based on hydrophobicity for identifying β-barrel membrane proteins in genomic sequences. It has been reported that this method correctly identified 75% of the OMPs (Liu *et al*., 2003). Martelli *et al*. (2002) used 12 OMPs and developed a Hidden Markov Model (HMM) for picking up the β-barrel membrane proteins and reported an accuracy of 84% in a set of 145 OMPs. Liu *et al*. (2003) analyzed the amino acid composition in the membrane spanning regions of 12 β-barrel membrane proteins and applied the information for discrimination, which showed 85% accuracy when tested with 241 OMPs. Bagos *et al*. (2004) developed an algorithm based on HMM for discriminating OMPs and reported an accuracy of 89% in a set of 133 OMPs. Natt *et al*. (2004) used a set of 16 OMPs and proposed a machine learning technique for discrimination, which showed an average accuracy of 90% in a set of randomly selected 100 globular and 16 OMPs. Garrow *et al*. (2005) proposed a modified *k*-nearest neighbor algorithm and reported an accuracy of 92.5% using weighted amino acids and evolutionary information. Martelli *et al*. (2003) reviewed the performance of a few methods for the discrimination and prediction of membrane protein structures. All these methods used minimal information for the analysis and the prediction accuracy is rather modest.

Further, few methods have been suggested to screen OMPs from genomic sequences (Zhai and Saier, 2002; Berven *et al*., 2004; Bigelow *et al*., 2004). Zhai and Saier (2002) developed a β-barrel finder program based on secondary structure, hydropathy and amphipathicity parameters and used it for identifying OMPs in *Escherichia coli* genome. This algorithm has recognized 10 families correctly and missed the proteins from 4 OMP families. Berven *et al*. (2004) proposed a program for identifying OMPs using two factors: (1) C-terminal pattern typical of many integral β-barrel proteins and (2) integral β-barrel score based on the extent to which the sequence contains stretches of amino acids typical of

*To whom correspondence should be addressed.

transmembrane β-strands. Bigelow *et al*. (2004) introduced a profile-based HMM for discriminating OMPs and suggested the probable OMPs in genomic sequences of 72 Gram-negative bacteria.

Classification based on support vector machines (SVMs) has several applications in bioinformatics and computational biology. It has been widely used to predict protein secondary structures (Nguyen and Rajapakse, 2005b), solvent accessibility (Yuan *et al*., 2002; Kim and Park, 2004; Nguyen and Rajapakse, 2005a), protein–protein binding sites (Bradford and Westhead, 2004; Res *et al*., 2005), remote protein homology detection (Busuttil *et al*., 2004), protein domains (Vlahovicek *et al*., 2005) protein subcellular localization (Hua and Sun, 2001; Park and Kanehisa, 2003; Nair and Rost, 2005) and gene and tissue classification from microarray expression data (Brown *et al*., 2000). The biological and bioinformatics applications of SVMs have been reviewed in Byvatov and Schneider (2003) and Yang (2004).

In our earlier work, we have developed a statistical method based on amino acid composition and residue pairs for discriminating OMPs (Gromiha and Suwa, 2005; Gromiha *et al*., 2005). It showed an accuracy of 89% in a dataset of 377 OMPs (Gromiha and Suwa, 2005). As SVMs have a wide range of applications and perform well in prediction algorithms, we have developed a method using SVMs for discriminating OMPs. We have examined the performance of SVMs using different kernel functions and parameters, and various sequence features represented by the composition of amino acid residues and residue pairs. We observed that SVMs could discriminate the OMPs at an accuracy of 92% with amino acid composition and the accuracy is improved to 94% using residue pair information. This method has the ability to correctly pick up the OMPs and exclude other folding types of globular proteins at high accuracy levels.

MATERIALS AND METHODS

Datasets

All protein sequences were collected from the dataset of Gromiha and Suwa (2005), extracted from the PSORT-B database (Gardy *et al*., 2003) for membrane proteins and the PDB40D_1.37 database of SCOP (Murzin *et al*., 1995; Berman *et al*., 2000) for globular proteins. The dataset included a total of 377 OMP sequences, 674 globular protein sequences and 268 α-helical membrane proteins. In these datasets, OMPs and α-helical membrane proteins have many redundant sequences and the globular proteins have been filtered with <40% sequence similarity.

Removal of highly similar sequences

Sequences with a high degree of similarity to other sequences were removed by all-to-all sequence similarity check using the program CD-HIT (Li *et al*., 2001), which produces a non-redundant protein database, using a greedy incremental algorithm as implemented by Holm and Sander (1998). We produced non-redundant protein datasets for all types of proteins by CD-HIT with full-length matches of <40% sequence identity. We did not consider the proteins containing B, X or Z in the amino acid sequence.

The total number of proteins in the final dataset are 208 OMPs, 673 globular proteins and 206 α-helical membrane proteins and the sequences are available at http: www.cbrc.jp/~gromiha/omp/dataset2.html.

Support vector machines

SVM is a learning algorithm (Cristianini and Shawe-Taylor, 2000), which from a set of positively and negatively labeled training vectors learns a classifier that can be used to classify new unlabeled test samples. SVM learns the classifier by mapping the input training samples $\{x_1, \ldots, x_n\}$ into a possibly high-dimensional feature space and seeking a hyperplane in this space which separates the two types of examples with the largest possible margin, i.e. distance to the nearest points. If the training set is not linearly separable, SVM finds a hyperplane, which optimizes a trade-off between good classification and large margin.

For actual implementation we used the freely downloadable SVM-light package by Joachims (1999). We tested linear, polynomial and RBF (radial basis function) kernels with various parameters.

Compositions of amino acids and amino acid pairs

Each protein in the training dataset of N proteins is characterized by a vector $\vec{x}_i$ $(i = 1, \ldots, N)$ representing certain sequence features, together with the positive label or the negative labels for discriminating two different groups (e.g. globular and OMPs). In addition to the amino acid composition, we considered the amino acid pair composition. The vector $\vec{x}_i$ has 20 elements for the amino acid composition and 400 elements for the amino acid pair composition. Amino acid composition is defined as the ratio between the number of occurrences of a specific amino acid residue and the total number of residues in a protein.

For each ordered pair of amino acids, the amino acid pair composition C_{ij} is defined as the number of occurrences of amino acid i followed by j divided by the total number of adjacent pairs (i.e. the length of the sequence minus one).

Feature selection

At first, we considered selection with the Fisher discriminant ratio (FDR). It is defined as

$$\mathrm{FDR}_i = \frac{(\mu^i_{\mathrm{OMP}} - \mu^i_{\mathrm{m}})^2}{(\sigma^i_{\mathrm{OMP}})^2 + (\sigma^i_{\mathrm{m}})^2}, \quad (1)$$

where μ^i_{OMP} and σ^i_{OMP} are the mean and variance of the i-th amino acid (20 residues) or amino acid pair (400 residue pairs) composition in OMPs, respectively. μ^i_{m} and σ^i_{m} denote the mean and variance of amino acid i or amino acid pair i composition in globular proteins, α-helical membrane proteins or non-OMPs. In this study, the group of non-OMPs means the sum of globular proteins and α-helical membrane proteins. Selecting features with the highest FDR values is often employed as a simple technique for feature selection (Liu *et al*., 2003).

Another selection procedure is done according to backward and forward selection for handling datasets with amino acid composition and amino acid pair composition. In this work, backward selection (elimination) started with the complete set of 20 amino acid composition features. It evaluates all the subsets by eliminating the composition of one amino acid from the complete set and selects the one with the best performance measure of Matthews correlation coefficient (MCC) [Equation (5)]. It then evaluates all the subsets with one feature less than the best subset from the previous step and selects the second best. The process stops when decreasing the size of current best subset leads to a lower prediction rate. Forward selection for the 400 amino acid pair composition features was started from the result of backward selection subset of amino acid composition features. It evaluates all the one-additional feature subsets and selects the one with the best prediction rate. It then builds all the two-additional feature subsets that include the features already selected from the first step and finds the best one. This process continues until increasing the size of the current subset leads to a lower performance measure. Although FDR has been used for feature selection in some studies, we adopted cross-validated classification accuracy for the selection criteria in this work.

Since we have small number of proteins in our dataset, it is expected that training with the complete set of pair composition features (400D) may cause overfitting. Hence, we performed forward selection with amino acid pair composition features to find a good small feature set.

5-Fold cross-validation test

The prediction performance was examined by the 5-fold cross-validation test, in which the three types of proteins were randomly divided into five subsets of approximately equal size. This means that the data were partitioned into training and test data in five different ways. After training the SVMs with a collection of four subsets, the performance of the SVMs was tested against the fifth subset. This process was repeated five times so that every subset was once used as the test data.

In order to assess the accuracy of prediction methods we use four measures. The sensitivity, specificity and overall accuracy are defined by

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{overall accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{4}$$

where TP, FP, TN and FN refer to the number of true positives, false positives, true negatives and false negatives proteins, respectively.

The MCC (Matthews, 1975) is defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \tag{5}$$

The value of MCC is one for a perfect prediction and zero for a completely random assignment. Sensitivity measures our ability to correctly predict OMPs, while specificity measures our ability for correctly reject non-OMPs.

RESULTS AND DISCUSSION

Amino acid composition in OMPs, globular and α-helical membrane proteins

Table 1 shows the result of amino acid composition for 20 amino acid residues in OMPs, globular and α-helical membrane proteins. The FDR values [Equation (1)] also have been computed and the results are presented in Table 2. The residues Ser, Glu, Cys and His show high FDR values (FDR > 0.2) between OMPs and globular proteins. These residues also show significant differences (>1) in percent composition between globular proteins and OMPs (Table 1). The formation of disulfide bonds between Cys residues requires an oxidative environment and such disulfide bridges are not usually found in intracellular proteins (Branden and Tooze, 1999). The analysis of the three-dimensional (3D) structures of 15 β-barrel OMPs shows the presence of just 8 (0.1%) Cys residues and none of them is in the membrane part (Gromiha and Suwa, 2003). Hence, the occurrence of Cys is significantly higher in globular proteins than in OMPs. Glu is a strong helix former (Chou and Fasman, 1978) and this tendency influences its higher occurrence in globular proteins than OMPs.

The composition of Ser shows the opposite tendency that the composition is higher in OMPs than globular proteins. The structural analysis of several OMPs shows that Ser plays an important role in the stability and function of OMPs (Gromiha and Suwa, 2005). In outer membrane protein A (OmpA, PDB code no. 1QJP), the interior of β-strands contain an extended hydrogen bonding network of charged and polar residues (Pautsch and Schulz, 2000). In outer membrane protease OmpT (PDB code no. 1I78), the side chains of the residues Ser22, Gln228 and Asn258, located above the membrane, form hydrogen bonds to main chain atoms in the β-barrel (Vandeputte-Rutten *et al*., 2001). Interestingly, none of the residues, which have high composition in globular proteins (Glu, His and Cys), is involved in such patterns (Pautsch and Schulz, 2000; Vandeputte-Rutten *et al*., 2001). Further, the importance of Ser to the stability and function of OMPs has been reported for outer

Table 1. Amino acid composition for the 20 amino acid residues in outer membrane, globular and α-helical membrane proteins

Category	Residue	Composition (%) Outer membrane (208)	Globular (673)	α-Helical membrane (206)
Aliphatic	Ala	9.4	8.4	10.3
	Gly	8.7	7.7	8.3
	Ile	4.7	5.8	7.5
	Leu	8.9	8.5	12.7
	Pro	3.7	4.5	4.3
	Val	6.7	7.2	8.2
Aromatic	Phe	3.7	3.8	5.5
	Tyr	4.1	3.4	2.8
	Trp	1.2	1.3	2.0
Negative charged	Asp	5.9	5.8	3.3
	Glu	4.9	6.7	3.7
Positive charged	Arg	5.2	5.1	4.4
	His	1.2	2.2	1.7
	Lys	4.9	6.2	3.4
Polar	Asn	5.4	4.4	3.1
	Gln	4.7	3.9	3.2
	Ser	8.0	5.8	5.9
	Thr	6.3	5.7	5.2
Sulfur containing	Cys	0.4	1.5	0.8
	Met	1.7	2.2	3.6

Table 2. FDR of each amino acid between OMP and globular or α-helical membrane proteins [see Equation (1)]. High FDR values in globular proteins (>0.2) are shown in bold.

Residue	FDR (Fisher discriminant ratio) Globular	α-Helical membrane
Ala	0.033	0.039
Arg	0.003	0.070
Asn	0.142	1.281
Asp	0.000	1.487
Cys	**0.267**	0.198
Gln	0.078	0.341
Glu	**0.312**	0.231
Gly	0.074	0.012
His	**0.247**	0.109
Ile	0.114	0.900
Leu	0.014	1.083
Lys	0.109	0.302
Met	0.096	1.528
Phe	0.000	0.496
Pro	0.082	0.066
Ser	**0.495**	0.745
Thr	0.039	0.254
Trp	0.004	0.295
Tyr	0.070	0.343
Val	0.018	0.293

membrane cobalamin transporter (BtuB, PDB code no. 1NQE), anion-selective porin (Omp32, PDB code no. 1E54), etc. (Zeth *et al*., 2000; Chimento *et al*., 2003a,b).

The amino acid composition of 20 residues in OMPs and α-helical membrane proteins shows that Met, Asp, Asn and Leu have a significant difference (FDR > 1.0) between them (Table 2). The group of OMPs showed higher composition of Asp and Asn than that in α-helical membrane proteins. The occurrence of residues Leu and Met is higher in α-helical membrane proteins than that in OMPs. Further, from Table 1, the residues Ala, Ile, Leu, Val, Phe, Trp and Met have higher composition in α-helical membrane proteins than OMPs. This result reflects the presence of hydrophobic stretches of amino acid residues in the membrane part of α-helical membrane proteins.

Kernel selection

We begin with the selection of a kernel from three possibilities: the simple linear kernel, the polynomial kernel and the RBF kernel. The performance of each classifier was measured by examining how well the classifier identified positive and negative examples in the test sets, according to the 5-fold cross-validation test. In this analysis, the RBF kernel showed the best performance with overall prediction rate or MCC values. Various values of the parameter γ were also tested for the RBF kernel, and our choice was $\gamma = 0.02$ or 0.03 for feature selection. The parameter C, which controls the trade-off between training error and margin, was set to 0.5 or 0.6 in this work. The RBF kernel is defined by

$$K(\vec{x}, \vec{y}) = \exp(-\gamma \|\vec{x} - \vec{y}\|^2), \quad (7)$$

where $\gamma = 1/\sigma^2$ and σ is called the width of the (Gaussian) kernel. Instead of explicitly mapping the objects to the possibly high-dimensional feature space, SVM usually works implicitly in the feature space by only computing the corresponding kernel $K(\vec{x}, \vec{y})$ between any two objects x and y.

Discrimination of OMPs and globular proteins

Table 3 shows the results of the 5-fold cross-validation tests for the RBF kernel SVM classifiers with the parameters $\gamma = 0.03$ and $C = 0.5$ using amino acid composition. The initial overall prediction rate with 20 amino acid composition was 91.1% and the value of MCC was 0.757. Starting from this 20 amino acid composition feature vector, we executed backward feature selection, and the overall prediction rate increased to 92.5%. This reduced set has a 17D feature vector with the exclusion of residues Ala, Glu and Lys. This result reveals that the feature selection with FDR values moderately improved the performance.

We performed a second feature selection (forward) for 400 amino acid pair composition features starting from the 17D feature subset. The second feature selection chose 8 pair compositions, and the final version yields a 25D feature vector. The information of the additional eight pair compositions improved the value of MCC from 0.798 to 0.846. The selected pair compositions were EL, AA, AT, SS, AG, AI, ID and YE.

Discrimination of OMPs and α-helical membrane proteins

Table 4 shows the results of the 5-fold cross-validation tests for the RBF kernel SVM classifiers with the parameters $\gamma = 0.03$ and $C = 0.6$ using amino acid composition. We observed that SVMs could discriminate the OMPs at an overall accuracy of 94.9% with 20 amino acid compositions and the prediction rate improved to 95.9% after backward selection. After backward elimination, the composition information of residues Arg, Asp, Tyr, Cys and His were removed from the training and test datasets (15D feature vector). The performance of this discrimination including overall accuracy and MCC was better than the discrimination of OMPs and globular proteins. In our interpretation, the results shown in Tables 3 and 4 indicate that discrimination of OMPs and α-helical membrane proteins seems to be easier than that of OMPs and globular proteins. It is reasonable that the discrimination of OMPs and α-helical membrane proteins at high accuracy is consistent with higher FDR values for α-helical membrane proteins compared with globular proteins (Table 2). Further, the information about the occurrence of hydrophobic residues in the membrane part of α-helical membrane proteins can discriminate this class of proteins at high accuracy as reported in Mitaku *et al*. (2002).

We have included the information about amino acid pair composition with the result of first feature selection (composition of 15 amino acids) and we observed that there is no improvement in the prediction rate (Table 4). This result revealed that the composition of reduced amino acids could discriminate the groups of OMPs and α-helical membrane proteins successfully.

Table 3. Discrimination of OMPs and globular proteins

Composition	Prediction rate (%)			
	Sensitivity	Specificity	Overall	MCC
Amino acid residues (20D)	82.7	93.8	91.1	0.757
Residue pairs (400D)	83.2	97.3	94.0	0.830
Residues and residue pairs (420D)	63.5	100.0	91.4	0.755
Reduced amino acids[a] (17D)	87.5	94.1	92.5	0.798
Reduced amino acids and residue pairs[b] (25D)	88.0	96.4	**94.4**	**0.846**

The number given in parentheses indicates the number of variables.
[a]After backward feature selection, the composition features of Ala, Glu and Lys were removed from the training and test sets (17D feature vectors).
[b]The second (forward) feature selection included eight kinds of amino acid pair compositions (EL, AA, AT, SS, AG, AI, ID and YE).
Highest prediction rate and MCC are shown in bold.

Table 4. Discrimination of OMPs and α-helical membrane proteins

Composition	Prediction rate (%)			
	Sensitivity	Specificity	Overall	MCC
Amino acid residues (20D)	98.6	91.3	94.9	0.901
Residue pairs (400D)	99.0	90.3	94.7	0.897
Residues and left residue pairs (420D)	95.7	89.8	92.8	0.856
Reduced amino acids[a] (15D)	99.0	92.7	**95.9**	**0.920**
Reduced amino acids left and residue pairs[b] (15D)	99.0	92.7	**95.9**	**0.920**

[a]After backward feature selection, the composition features of Arg, Asp, Tyr, Cys and His were removed from the training and test sets (15D feature vector).
[b]None of the residue pairs was selected by the second feature selection.
Highest prediction rate and MCC are shown in bold.

Table 5. Discrimination of OMPs and non-OMPs

Composition	Prediction rate (%)			
	Sensitivity	Specificity	Overall	MCC
Amino acid residues (20D)	87.5	92.6	91.6	0.752
Residue pairs (400D)	86.5	96.4	94.5	0.823
Residues and residue pairs (420D)	79.3	99.0	95.2	0.840
Reduced amino acids[a] (18D)	89.9	92.5	92.0	0.767
Reduced amino acids and residue pairs[b] (28D)	90.9	94.7	**93.9**	**0.816**

Non-OMPs (879) consist of globular proteins (673) and α-helical membrane proteins (206).

[a]The first feature selection was backward selection for 20 kinds of amino acid and the composition features of Ala and Glu were removed.

[b]In the second feature selection, the composition features QA, DF, DA, KK, EF, NK, DR, YN, FF and LI were added to training and test sets ($20 - 2 + 10 = 28$D feature vector) by forward selection. The overall prediction rate and MCC are shown in bold.

Table 6. Top 10 amino acid pair compositions (by FDR) between OMPs and, globular, α-helical membrane proteins or non-OMPs

Globular Amino acid pair	FDR	α-Helical Amino acid pair	FDR	Non-OMPs Amino acid pair	FDR
LS	0.178	LI	1.000	SS	0.166
LG	0.163	LL	0.894	SY	0.148
SL	0.160	IL	0.694	GY	0.131
SS	0.146	AI	0.660	QS	0.128
SA	0.140	FL	0.625	LS	0.126
EE	0.138	LV	0.604	IL	0.125
SY	0.126	IF	0.603	SN	0.123
AS	0.126	IV	0.572	SA	0.122
GY	0.111	IA	0.526	NS	0.115
QS	0.109	LM	0.490	AQ	0.113

Discrimination of OMPs and non-OMPs

We have examined the discrimination ability of OMPs and non-OMPs by the present method. For this purpose, we defined a non-OMPs dataset by combining the globular and α-helical membrane proteins. Table 5 shows the result of 5-fold cross-validation tests for the RBF kernel SVM classifiers with the parameters $\gamma = 0.02$ and $C = 0.5$ using amino acid and amino acid pair composition. Again, we executed a two-step feature selection, consisting of first the backward selection and then the forward selection. Backward selection reduced the amino acid composition features to 18 by excluding the residues Ala and Glu. This backward selection improved the sensitivity of OMPs from 87.5 to 89.9% and the overall accuracy from 91.6 to 92%.

After backward elimination, we did forward selection for 400 amino acid pair composition to the 18 amino acid composition feature subset. As a result of forward selection, the final feature vector of our method contained 10 additional amino acid pair compositions of QA, DF, DA, KK, EF, NK, DR, YN, FF and LI. Interestingly, most of these amino acid pairs contain either a charged or an aromatic residue. The inclusion of representative amino acid pairs, such as two like/oppositely charged (KK, DR), hydrophobic (LI), aromatic (FF) and the combination of charged and hydrophobic (DA), polar and charged (NK), charged and aromatic (EF), aromatic and polar (YN), polar and hydrophobic (QA) and charged and hydrophobic (DA) improved the prediction accuracy. However, the inclusion of amino acid pairs that show significant difference between OMPs and non-OMPs (Table 6) did not improve the accuracy. This might be due to the fact that such information is already available in the feature selection of 18 amino acid composition. Another reason might be the fact that the hyperplane of an SVM is constructed by combination of several features, whereas the FDR reflects the difference of only one feature. The additional information about the 10 amino acid pairs increased the MCC from 0.767 to 0.816 and the accuracy from 92 to 94%.

The combination of all the amino acid and residue pair compositions raised the correlation up to 0.84 and the overall accuracy is 95.2% (Table 5). Although the accuracy and correlation are high it has 420 variables (20 amino acids and 400 residue pairs). Generally, fitting the data with a minimum number of variables increases the robustness of the results. In the present work we selected 28 feature variables (18 amino acids and 10 residue pairs), giving an accuracy of 93.9%. This is almost as high as the accuracy (95.2%) obtained when using all 420 features. Thus we recommend using the selected features for discrimination.

Implementation

The prediction method presented in this paper is implemented as a computer program named TMBETA-SVM and the web service is made available at http://tmbeta-svm.cbrc.jp. The program predicts OMPs based on the compositions of amino acids and amino acid pairs, using the SVM classifiers with the RBF kernel and the parameters $C = 0.5$ and $\gamma = 0.02$. The datasets used in this work are also available at http://www.cbrc.jp/~gromiha/omp/dataset2.html.

Comparison with other methods

Liu *et al*. (2003) proposed a method based on the amino acid composition of residues in transmembrane β-strand segments of 12 proteins in PDB to discriminate β-barrel membrane proteins and claimed an accuracy of 85.4% on a set of 241 OMPs. As the membrane spanning segments are used to compute the amino acid composition, this method could identify the OMPs, which have a high content of amino acid residues in the membrane but it missed the proteins with fewer membrane spanning β-strand segments. Martelli *et al*. (2002) devised an HMM method using 12 OMPs in PDB and tested the method on 145 OMPs, which yielded an accuracy of 84%. Bagos *et al*. (2004) used an HMM for discriminating β-barrel OMPs and reported an accuracy of 88.8% for a set of 133 OMPs. The method based on amino acid composition showed an accuracy of 89% on a dataset of 377 OMPs (Gromiha and Suwa, 2005). In this work, we have used a set of 208 OMPs, 673 globular proteins, and 206 α-helical membrane proteins. The OMPs were discriminated with an accuracy of 94% from the pool of 1087 sequences, while correctly excluding 96% of the transmembrane α-helical proteins. This compares favorably to an accuracy of 90% reported by an HMM (Martelli *et al*., 2002). Although the direct comparison of accuracies reported by different methods is not appropriate (due to differences in datasets and validation procedures) it may give some information about the performance of different methods. We have examined the discriminative power of the program TMB-Hunt (Garrow *et al*., 2005), which claimed

the highest accuracy of 92.5%, using the publicly available web server and the same dataset of 1087 proteins used in the present work. We observed that this program discriminated the OMPs with an accuracy of 89.2% whereas the cross-validation accuracy obtained by the present work is 93.9%. The MCC obtained with TMB-Hunt is 0.729 and that obtained with our method is 0.816. The high accuracy achieved by the present method is due to the effectiveness of the method as well as the information gained from the large dataset of globular proteins and OMPs.

CONCLUSIONS

We have developed an SVM-based method for discriminating OMPs using amino acid composition and residue pairs. The influence of amino acids and residue pairs for improving the accuracy has been analyzed. We found that the selection of 18 amino acid residues could discriminate the OMPs at an accuracy of 92%. Further, the inclusion of 10 residue pairs raised the accuracy to 94% and the correlation from 0.77 to 0.82. We have developed a web server for discriminating OMPs, which takes the amino acid sequence as input, and the predicted type of the protein is displayed as output. The program is available online at http://tmbeta-svm.cbrc.jp/

Conflict of Interest: none declared.

REFERENCES

Bagos,P.G. *et al*. (2004) A hidden Markov model method, capable of predicting and discriminating β-barrel outer membrane proteins. *BMC Bioinformatics*, **5**, 29.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Berven,F.S. *et al*. (2004) BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res*., **32**, W394–W399.

Bigelow,H.R. *et al*. (2004) Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res*., **32**, 2566–2577.

Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.

Branden,C. and Tooze,C. (1999) *Introduction to Protein Structure*. Garland Publishing Inc., New York.

Brown,M.P.S. *et al*. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Busuttil,S. *et al*. (2004) Support vector machines with profile-based kernels for remote protein homology detection. *Genome Inform. Ser. Workshop Genome Inform*., **15**, 191–200.

Byvatov,E. and Schneider,G. (2003) Support vector machine applications in bioinformatics. *Appl. Bioinformatics*, **2**, 67–77.

Chen,C.P. and Rost,B. (2002) State-of-the-art in membrane protein prediction. *Appl. Bioinformatics*, **1**, 21–35.

Chimento,D.P. *et al*. (2003a) Substrate-induced transmembrane signaling in the cobalamin transporter BtuB. *Nat. Struct. Biol*., **10**, 394–401.

Chimento,D.P. *et al*. (2003b) The *Escherichia coli* outer membrane cobalamin transporter BtuB: structural analysis of calcium and substrate binding, and identification of orthologous transporters by sequence/structure conservation. *J. Mol. Biol*., **332**, 999–1014.

Chou,P.Y. and Fasman,G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol*., **47**, 45–148.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, MA.

Forst,D. *et al*. (1998) Structure of the sucrose-specific porin ScrY from *Salmonella typhimurium* and its complex with sucrose. *Nat. Struct. Biol*., **5**, 37–46.

Gardy,J.L. *et al*. (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res*., **31**, 3613–3617.

Garrow,A.G. *et al*. (2005) TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. *Nucleic Acids Res*., **33**, W188–W192.

Gnanasekaran,T.V. *et al*. (2000) Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins. *Bioinformatics*, **16**, 839–842.

Gromiha,M.M. (1999) A simple method for predicting transmembrane alpha helices with better accuracy. *Protein Eng*., **12**, 557–561.

Gromiha,M.M. and Suwa,M. (2003) Variation of amino acid properties in all-beta globular and outer membrane protein structures. *Int. J. Biol. Macromol*., **32**, *93–98*.

Gromiha,M.M. and Suwa,M. (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics*, **21**, 961–968.

Gromiha,M.M. *et al*. (1997) Identification of membrane spanning beta strands in bacterial porins. *Protein Eng*., **10**, 497–500.

Gromiha,M.M. *et al*. (2005) Application of residue distribution along the sequence for discriminating outer membrane proteins. *Comput. Biol. Chem*., **29**, 135–142.

Hirokawa,T. *et al*. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.

Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Joachims,T. (1999) Making large-scale SVM learning practical. In Schölkopf,B., Burges,C. and Smola,A. (eds) *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA.

Kim,H. and Park,H. (2004) Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins*, **54**, 557–562.

Li,W. *et al*. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Liu,Q. *et al*. (2003) Identification of β-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. *Comput. Biol. Chem*., **27**, 355–361.

Mannella,C.A. (1998) Conformational changes in the mitochondrial channel protein, VDAC and their functional implications. *J. Struct. Biol*., **121**, 207–218.

Martelli,P.L. *et al*. (2002) A sequence-profile-based HMM for predicting and discriminating β-barrel membrane proteins. *Bioinformatics*, **18**, S46–S53.

Martelli,P.L. *et al*. (2003) The prediction of membrane protein structure and genome structural annotation. *Comp. Funct. Genomics*, **4**, 406–409.

Matthews,B.W. (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Mitaku,S. *et al*. (2002) Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics*, **18**, 608–616.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol*., **348**, 85–100.

Natt,N.K. *et al*. (2004) Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins*, **56**, 11–18.

Nguyen,M.N. and Rajapakse,J.C. (2005a) Prediction of protein relative solvent accessibility with a two-stage SVM approach. *Proteins*, **59**, 30–37.

Nguyen,M.N. and Rajapakse,J.C. (2005b) Two-stage multi-class support vector machines to protein secondary structure prediction. *Pac. Symp. Biocomput*., 346–357.

Park,K.-J. and Kanehisa,M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acid pairs. *Bioinformatics*, **19**, 1656–1663.

Pautsch,A. and Schulz,G.E. (2000) High-resolution structure of the OmpA membrane domain. *J. Mol. Biol*., **298**, 273–282.

Res,I. *et al*. (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, **21**, 2496–2501.

Schirmer,T. *et al*. (1995) Structural basis for sugar translocation through maltoporin channels at 3.1 A resolution. *Science*, **267**, 512–514.

Schulz,G.E. (2000) β−Barrel membrane proteins. *Curr. Opin. Struct. Biol*., **10**, 443–447.

Schulz,G.E. (2002) The structure of bacterial outer membrane proteins. *Biochim. Biophys. Acta*, **1565**, 308–317.

Vandeputte-Rutten,L. *et al*. (2001) Crystal structure of the outer membrane protease OmpT from *Escherichia coli* suggests a novel catalytic site. *EMBO J*., **20**, 5033–5039.

Vlahovicek,K. *et al.* (2005) The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines. *Nucleic Acids Res*., **33** (Database Issue), D223–D225.

Wimley,W.C. (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci*., **11**, 301–312.

Wimley,W.C. (2003) The versatile β−barrel membrane protein. *Curr. Opin. Struct. Biol*., **13**, 404–411.

Yang,Z.R. (2004) Biological applications of support vector machines. *Brief Bioinformatics*, **5**, 328–338.

Yuan,Z. *et al.* (2002) Prediction of protein solvent accessibility using support vector machines. *Proteins*, **48**, 566–570.

Zeth,K. *et al.* (2000) Crystal structure of Omp32, the anion-selective porin from *Comamonas acidovorans*, in complex with a periplasmic peptide at 2.1 A resolution. *Structure*, **8**, 981–992.

Zhai,Y. and Saier,M.H.,Jr (2002) The β−barrel finder (BBF) program, allowing identification of outer membrane β−barrel proteins encoded within prokaryotic genomes. *Protein Sci*., **11**, 2196–2207.