Computational Biology and Chemistry 32 (2008) 227–231

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Software Note

TMBETADISC-RBF: Discrimination of β-barrel membrane proteins using RBF networks and PSSM profiles

Yu-Yen Ou [a], M.Michael Gromiha [b,*], Shu-An Chen [a], Makiko Suwa [b]

[a] Department of Computer Science and Engineering, Yuan Ze University, Chung-Li, Taiwan
[b] 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

ARTICLE INFO

Article history:
Received 9 November 2007
Received in revised form 11 March 2008
Accepted 11 March 2008

Keywords:
Outer membrane protein
Discrimination
Radial basis function
PSSM profile

ABSTRACT

Discriminating outer membrane proteins (OMPs) from other folding types of globular and membrane proteins is an important task both for identifying OMPs from genomic sequences and for the successful prediction of their secondary and tertiary structures. We have developed a method based on radial basis function networks and position specific scoring matrix (PSSM) profiles generated by PSI-BLAST and non-redundant protein database. Our approach with PSSM profiles has correctly predicted the OMPs with a cross-validated accuracy of 96.4% in a set of 1251 proteins, which contain 206 OMPs, 667 globular proteins and 378 α-helical inner membrane proteins. Furthermore, we applied our method on a dataset containing 114 OMPs, 187 TMH proteins and 195 globular proteins obtained with less than 20% sequence identity and obtained the cross-validated accuracy of 95%. This accuracy of discriminating OMPs is higher than other methods in the literature and our method could be used as an effective tool for dissecting OMPs from genomic sequences. We have developed a prediction server, TMBETADISC-RBF, which is available at http://rbf.bioinfo.tw/ sachen/OMP.html.

1. Introduction

The successful discrimination of β-barrel or outer membrane proteins (OMPs) from other folding types of globular and membrane proteins would help to dissect them in genomic sequences. In recent years, several methods have been proposed for discriminating OMPs based on amino acid conformational parameters, physical chemical properties and machine learning techniques. However, the accuracy of discriminating OMPs is significantly lower than that of transmembrane helical (TMH) proteins. The occurrence of long stretches of hydrophobic residues in TMH proteins raised the accuracy of discriminating them while the intervention of polar and charged amino acid residues in OMPs makes it difficult to discriminate them.

In our earlier works, we have proposed statistical methods based on amino acid composition, dipeptide preferences and motifs for discriminating OMPs (Gromiha and Suwa, 2005; Gromiha et al., 2005; Gromiha, 2005). On the other hand, machine learning techniques have been used for discriminating OMPs at better accuracy. They include the Hidden Markov models (Martelli et al., 2002; Bagos et al., 2004), *k*-nearest neighbor methods (Garrow et al., 2005), neural networks (Gromiha and Suwa, 2006), support vector machines (Park et al., 2005), the combination of neural networks and support vector machines (Natt et al., 2004) and consensus method (Garrow and Westhead, 2007). These methods attained the discrimination accuracy in the range of 80–94%. All these methods used minimal information for the analysis, and prediction accuracy is rather modest or assessed with limited number of data. The accuracy of statistical methods are influenced with high sensitivity whereas the machine learning techniques show high specificity (Gromiha and Suwa, 2007). Further, no method has been developed with position specific scoring matrice (PSSM) profiles for discriminating OMPs.

Classification based on radial basis function (RBF) network has several applications in bioinformatics. It has been widely used to predict the cleavage sites in proteins (Yang and Thomson, 2005), inter-residue contacts (Zhang and Huang, 2004), protein disorder (Su et al., 2006), selecting siRNA sequences (Takasaki et al., 2006) and so on. In this work, we have developed a method based on RBF network and discriminated the OMPs using amino acid composition and residue pair preference. Further, the position specific scoring matrices have been included for discrimination. The RBF network discriminated the OMPs at an accuracy of 95.1% and the inclusion of alignment profiles remarkably improved the discrimination accuracy up to the level of 96.4%. We have set

* Corresponding author. Tel.: +81 3 3599 8046; fax: +81 3 3599 8081.
E-mail addresses: yien@csie.org (Y.-Y. Ou), michael-gromiha@aist.go.jp (M.Michael Gromiha).

1476-9271/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2008.03.002

up a web server for discriminating OMPs and is available at http://rbf.bioinfo.tw/ sachen/OMP.html.

2. Materials and Methods

2.1. Datasets

In our earlier study (Gromiha and Suwa, 2006), we have used a set of 208 OMPs, 674 globular proteins and 206 TMH proteins for discriminating OMPs. Tusnády et al. (2005) developed a database of transmembrane proteins (PDBTM), which has a non-redundant (NR) set of 32 OMPs and 193 TMH protein structures. We have combined the respective sequences from these datasets and constructed a new set of proteins, which have less than 40% sequence identity using the program CD-HIT (Li et al., 2001). Further, the dataset has been refined with BLAST (Altschul et al., 1997) and verified that no two sequences have the identity of more than 40%. The final dataset contains 206 OMPs, 667 globular and 378 TMH proteins. Further, we have used four other datasets to validate the performance of our method, (i) 1612 globular proteins belonging to 30 major folding types, (ii) 5261 globular proteins obtained from ASTRAL database (Chandonia et al., 2004) with the sequence identity of less than 25%, (iii) a dataset of 114 OMPs, 187 TMH proteins and 195 globular proteins obtained with less than 20% sequence identity and (iv) a 112 OMPs, 181 TMH and 195 globular proteins obtained with e-value $> e^{-50}$. Datasets (iii) and (iv) are subsets of the main dataset, refined with the aid of the program, blastclust (Altschul et al., 1997). In addition, we have applied our method to the whole genomic sequences of *E. coli* for detecting OMPs.

2.2. Design of the Radial Basis Function Networks

An RBFN consists of three layers, namely the input layer, the hidden layer and the output layer. The input layer broadcasts the coordinates of the input vector to each of the nodes in the hidden layer. Upon receiving a stimulus, each node in the hidden layer then generates an activation based on the associated radial basis function and each node in the output layer computes a linear combination of the activations generated by the hidden nodes. The main difference from neural network is that the hidden units perform the computations in RBFN. It has a significant advantage over neural network that the first set of parameters can be determined independently of the second set and still produce accurate classifiers (Witten and Frank, 2005). Comparing RBFN and SVM, it is not necessary to perform the model selection process in RBFN, which makes RBFN more efficient than SVM. The main disadvantage of RBFN is that they give every attribute the same weight and hence it cannot deal effectively with irrelevant attributes.

The parameters that RBFN learn are (i) the centers and widths of the RBFs and (ii) the weights used to form the linear combination of the outputs obtained from the hidden layer. In this work, we have used all training data as hidden neurons for getting the best results. Further, no special parameters have been selected in the network. We have carefully checked the network and avoided the problem of overfitting. The details about network structure and design can be found in Ou et al. (2005).

The general mathematical form of the output nodes in an RBFN is as follows:

$$g_j(\mathbf{x}) = \sum_{i=1}^{k} w_{ji}\phi(\|\mathbf{x} - \mu_i\|; \sigma_i), \tag{1}$$

where $g_j(\mathbf{x})$ is the function corresponding to the j-th output node and is a linear combination of k radial basis functions $\phi(\ )$ with center μ_i and bandwidth σ_i. Also, w_{ji} is the weight associated with the link between the j-th output node and the i-th hidden node.

For data classification applications, it is assumed that each sample is described by a multi-dimensional feature vector. The RBFN will have one output node corresponding to one class of samples and a query sample is predicted to belong to the class of which the corresponding output function gives the maximum value. The tasks that the learning algorithm carries out include: (1) determining where the activation functions of the hidden nodes should be located; (2) figuring out the parameters of the activation functions; (3) optimizing the weights associated with the links between the hidden layer and the output layer.

In our implementation, we put an activation function on each training instance, and the learning algorithm simply sets all the bandwidth parameters to a constant and attempts to minimize

$$J(W) = \sum_{j=1}^{m} P_j E_j\{\|W^{\mathrm{T}}\mathbf{h} - V_j\|^2\} + \lambda \sum_{j=1}^{m} \mathbf{w}_j^{\mathrm{T}}\mathbf{w}_j, \tag{2}$$

where

(1) P_j and $E_j\{\ \}$ are the a priori probability and the expected value of class-j samples, respectively, and m is the number of classes in the training dataset;
(2) $W = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m]$, $\mathbf{w}_j = [w_{j1}, w_{j2}, \ldots, w_{jk}]^{\mathrm{T}}$;
(3) $\mathbf{h} = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_k(\mathbf{x})]^{\mathrm{T}}$;
(4) V_j is the j-th column vector of an $m \times m$ identity matrix;
(5) λ is the regularization parameter.

The typical approach to obtain the optimal W that minimizes J is to solve the following equation:

$$\nabla_W J(W) = 2\sum_{j=1}^{m} P_j E_j\{\mathbf{h}\mathbf{h}^{\mathrm{T}} + \lambda I\}W - 2\sum_{j=1}^{m} P_j E_j\{\mathbf{h}\}V_j^{\mathrm{T}} = [0], \tag{3}$$

where $[0]$ is a $k \times m$ null matrix.

Let K denotes the matrix of the second-order moments under the mixture distribution, then we have

$$K = \sum_{j=1}^{m} P_j E_j\{\mathbf{h}\mathbf{h}^{\mathrm{T}}\}. \tag{4}$$

Furthermore, Eq. (3) becomes

$$(K + \lambda I)W = M, \tag{5}$$

where

$$M = \sum_{j=1}^{m} P_j E_j\{\mathbf{h}\}V_j^{\mathrm{T}}. \tag{6}$$

Since we can set $\lambda > 0$ to guarantee that $(K + \lambda I)$ is a positive definite matrix, we can apply the Cholesky decomposition to solve Eq. (5) efficiently (Press, 1992).

2.3. Compositions of Amino Acids and Amino Acid Pairs

If we have n proteins in the training data, we can use n vectors $\{x_i, i = 1, \ldots, n\}$, to represent all training data. Each vector has a label to show the protein is belonging to which group (e.g. OMPs or non-OMPs).

The vector x_i has 20 elements for the amino acid composition and 400 elements for the amino acid pair composition. The 20 elements are the number of occurrences of 20 amino acids, and the 400 elements are the number of occurrences of 400 different amino acid pairs. In this paper, we also combine amino acid composition and

amino acid pair composition, and then have 420 elements in each vector.

2.4. PSSM Profiles

Recently, several investigators tried to predict the OMPs based on amino acid composition, residue pairs and motifs. In the structural point of view, several amino acid residues can be mutated without altering the structure of a protein and it is possible that two proteins have similar structures with different amino acid compositions. Hence, we have used the position specific scoring matrix profiles for discriminating OMPs, which have been widely used in protein secondary structure prediction, subcellular localization and other bioinformatics problems with significant improvement (Jones, 1999; Xie et al., 2005). The PSSM profiles have been obtained by using PSI-BLAST and non-redundant protein database. In the prediction of OMPs, we use PSSM profiles to generate 400D input vector as input features by summing up the same amino acid rows in PSSM profiles. This will account the probability of changing a specific amino acid into all other types. Every element of 400D input vector was divided by the length of the sequence and then be scaled by $1/1 + e^{-x}$.

2.5. Assessment of Predictive Ability

The prediction performance was examined by 5-fold cross-validation test, in which the three types of proteins were randomly divided into five subsets of approximately equal size. We have trained the data with four subsets and the remaining set was used to test the performance of the method. This process is repeated five times so that every subset is once used as the test data.

We used sensitivity, specificity, accuracy and Matthew's correlation coefficient (MCC) to assess the prediction performance. Here, TP, FP, TN and FN refer to the number of true positives (OMPs identified as OMPs), false positives (non-OMPs identified as OMPs), true negatives (non-OMPs identified as non-OMPs) and false negatives (OMPs identified as non-OMPs), respectively.

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (8)$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (9)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (10)$$

3. Results and Discussion

3.1. Discrimination of OMPs from Globular/TMH Proteins

We have examined the predictive power of the present method for discriminating OMPs from a pool of 206 OMPs and 1045 non-OMPs (globular/TMH proteins) and the results are presented in Table 1. We observed that our method could discriminate the OMPs with the accuracy of 96.4% using PSSM profiles. The sensitivity and specificity are respectively, 89.3% and 97.8%. Our analysis showed that PSSM profiles marginally improved the discrimination accuracy compared with dipeptide composition. We achieved the correlation of 0.87, which is significantly better than that obtained with amino acid composition.

We have analyzed the ability of simple BLAST search to identify the similarities between the OMPs in the dataset. In this procedure, for a given protein in a cross-validation set, we have searched for similar sequences in the other four sets; if the best hit is an OMP, the target protein has been assigned as an OMP or non-OMP, otherwise. In this method we obtained the accuracy of 78.7% for discriminating OMPs and non-OMPs. The sensitivity, specificity and correlation are respectively, 94.7%, 75.5% and 0.54. Hence, we suggest that our method could be an effective tool for detecting OMPs in genomic sequences.

Table 1
Discrimination of OMPs and non-OMPs

	20D	400D	420D	PSSM
Sensitivity	87.4	90.8	90.8	89.3
Specificity	94.9	95.1	96.0	97.8
Accuracy	93.7	94.4	95.1	96.4
MCC	0.78	0.81	0.83	0.87

20D: amino acid composition for the 20 amino acid residues; 400D: composition for the 400 amino acid residue pairs; 420D: combination of amino acids and residue pairs.

Further, we have analyzed the performance of the present method for discriminating globular and OMPs. We observed that the PSSM-based method could discriminate the OMPs with the accuracy of 96.2%. The sensitivity, specificity and MCC are respectively, 91.3%, 97.8% and 0.89. In addition, we have tested the validity of our approach using a set of 206 OMPs and 5639 non-OMPs to represent the real case scenario in genomic sequences. Our method identified the OMPs with the sensitivity of 85.4% and excluded the non-OMPs with the specificity of 99.8%. The accuracy and MCC are, 99.3% and 0.90, respectively.

3.2. Discrimination Results Using the Dataset of Proteins Obtained with Less Than 20% Sequence Identity

We have examined the influence of datasets for discriminating OMPs using a subset of sequences obtained with less than 20% sequence identity and the results are presented in Table 2. We have trained our method with reduced dataset and evaluated the performance with 5-fold cross-validation method. We noticed that the different measures (sensitivity, specificity, accuracy and MCC) showed a similar trend to that obtained with large dataset. The refined dataset yielded the accuracy of 95%, which is better than that obtained with neural network, 89% (Gromiha and Suwa, 2006). On the other hand, we have removed the homologous sequences with the *e*-value cutoff of e^{-50}, which yielded 112 OMPs, 181 TMH and 195 globular proteins. With these datasets, our method showed the 5-fold cross-validation accuracy of 93.7%. The sensitivity, specificity and MCC are, 81.3%, 97.3% and 0.82, respectively.

3.3. Prediction Results for Different Folding Types of Globular Proteins

We have tested the present method in a set of 1612 globular proteins belonging to 30 different folding types. These proteins have been selected from the SCOP database (Murzin et al., 1995) with the

Table 2
Discrimination results obtained with the dataset of less than 20% sequence identity

	20D	400D	420D	PSSM
Sensitivity	82.5 (74.6)	80.8	85.6	84.2
Specificity	95.6 (92.7)	96.9	96.6	98.2
Accuracy	92.5 (88.5)	93.2	93.1	**95.0**
MCC	0.79	0.81	0.81	0.85

The results reported with neural networks are given in parentheses. Highest accuracy is shown in bold. 20D: amino acid composition for the 20 amino acid residues; 400D: composition for the 400 amino acid residue pairs; 420D: combination of amino acids and residue pairs.

Table 3
Exclusion of globular proteins belonging to 30 major folds

Fold	Neural	RBF 20D	RBF 400D	RBF 420D	RBF PSSM
Cytochrome C (a.3)	96	100.0	100.0	100.0	100.0
DNA/RNA binding 3-helical bundle (a.4)	95.1	98.1	100.0	98.1	100.0
Four helical up and down bundle (a.24)	92.3	100.0	100.0	96.2	100.0
EF hand-like fold (a.39)	96	100.0	100.0	100.0	100.0
SAM domain-like (a.60)	100	100.0	100.0	100.0	100.0
α– α superhelix (a.118)	95.7	97.9	100.0	97.9	100.0
Immunoglobulin-like β-sandwich (b.1)	89.6	98.8	99.4	100.0	100.0
Common fold of diphtheria toxin/transcription factors/cytochrome f (b.2)	85.7	100.0	96.4	96.4	100.0
Cupredoxin-like (b.6)	93.3	100.0	100.0	100.0	100.0
Galactose-binding domain-like (b.18)	96	96.0	96.0	96.0	100.0
Concanavalin A-like lectins/glucanases (b.29)	76.9	84.6	96.2	96.2	100.0
SH3-like barrel (b.34)	97.6	100.0	100.0	100.0	100.0
OB-fold (b.40)	96.2	98.7	98.7	100.0	100.0
Double-stranded β-helix (b.82)	97.1	100.0	100.0	100.0	100.0
Nucleoplasmin-like (b.121)	90.5	95.2	97.6	100.0	100.0
TIM β/ α-barrel (c.1)	94.5	98.6	96.6	97.9	100.0
NAD(P)-binding Rossmann-fold domains (c.2)	96.1	97.4	98.7	100.0	100.0
FAD/NAD(P)-binding domain (c.3)	93.5	96.8	96.8	96.8	100.0
Flavodoxin-like (c.23)	98.2	100.0	100.0	100.0	100.0
Adenine nucleotide αhydrolase-like (c.26)	97.1	100.0	100.0	100.0	100.0
P-loop containing nucleoside triphosphate hydrolases (c.37)	98.9	94.7	100.0	98.9	100.0
Thioredoxin fold (c.47)	100	100.0	100.0	100.0	100.0
Ribonuclease H-like motif (c.55)	95.9	100.0	98.0	98.0	100.0
S-Adenosyl-L-methionine-dependent methyltransferases (c.66)	100	100.0	100.0	100.0	100.0
α/ β-Hydrolases (c.69)	89.2	97.3	100.0	100.0	100.0
"beta-Grasp, ubiquitin-like (d.15)"	97.6	97.6	100.0	100.0	100.0
Cystatin-like (d.17)	96	100.0	100.0	100.0	100.0
Ferredoxin-like (d.58)	93.2	98.3	100.0	100.0	100.0
Knottins (g.3)	100	100.0	100.0	100.0	100.0
Rubredoxin-like (g.41)	100	100.0	100.0	100.0	100.0

All records are in accuracy (%).

criteria that there should be at least 25 proteins in each fold and the sequence identity is not more than 25%. The results are presented in Table 3.

We observed that the present method with PSSM profiles could correctly exclude all the proteins in all the considered folding types and the accuracy is 100%. It is interesting that all the proteins belonging to 13 different folding types are correctly excluded by all the four attributes used in this work. Further, the proteins in the family Concanavalin A-like lectins/glucanases, which have strong correlation with the composition of OMPs, are also excluded with the accuracy of 84–100% using different features. These results are better than that reported with neural network (Gromiha and Suwa, 2006). In addition, we have tested a set of 5261 globular proteins obtained from ASTRAL database (Chandonia et al., 2004) and our method excluded them with the accuracy of 96.2%.

3.4. Comparison with Other Methods

We have compared the performance of the present method with that of other methods in the literature and the results are presented in Table 4. We observed that the statistical methods and machine learning techniques correctly discriminated the OMPs with the accuracy in the range of 80–94%. Liu et al. (2003) proposed a method based on the amino acid composition of residues in transmembrane beta-strand segments to discriminate OMPs. They used just 12 proteins for developing the parameters and tested with 241 OMPs, and the accuracy was reported to be 84%. Martelli et al. (2002) devised a method based on HMM using 12 OMPs and tested the method in 145 OMPs, which has yielded the accuracy of 84%. Bagos et al. (2004) used a HMM for discriminating OMPs and obtained the accuracy of 88% for a set of 133 OMPs. Garrow et al. (2005) reported an accuracy of 92.5% for discriminating OMPs. We have used a set of 206 OMPs and 1045 non-OMPs and the present method using RBF networks and PSSM profiles improved the accuracy up to 96.4%, which is better than other methods in the literature. Further, our method could correctly exclude 1612 globular proteins from 30 major folding types with the accuracy of 100%. 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 1251 proteins used in the present work. We observed that this program discriminated the OMPs with an accuracy of 94.3% whereas the cross-validation accuracy obtained by the present work is 96.4%. The MCC obtained with TMB-Hunt is 0.80 and with our method is 0.87. The high performance of the present method might be due to the inclusion of PSSM profiles and large dataset of OMPs.

Table 4
Comparison of eight different methods for discriminating OMPs

Method	Accuracy (%)	Reference
Sequence alignment profile	80	Gnanasekaran et al. (2000)
Statistical	84	Liu et al. (2003)
HMM	84	Martelli et al. (2002)
HMM	88	Bagos et al. (2004)
Statistical	89	Gromiha and Suwa (2005)
k-Nearest neighbor	93	Garrow et al. (2005)
Support vector machines	94	Park et al. (2005)
RBF network and PSSM	96	Present work

3.5. Detecting OMPs in E. coli Genome

We have applied our method for detecting OMPs in the genomic sequenes of *E. coli*. In order to make the test fair, we have trained the data by removing all the *E. coli* OMPs. Our method picked up 220

proteins as OMPs among the 4237 sequences in *E. coli*. This result showed that 5% of proteins are OMPs, which is comparable to other methods in detecting OMPs in *E. coli* (Berven et al., 2004; Bigelow and Rost, 2006). When we tested the predicted OMPs with 11 known structures of OMPs from *E. coli*, our method could detect all the OMPs with the sensitivity of 100%. In addition, our method identified 16 of the 18 OMPs that were removed from the training data.

We have compared the OMPs identified in *E. coli* using different methods with the 36 OMPs deposited in transport classification database, TCDB (Saier et al., 2006). Our method could detect 34 out of 36 OMPs with the accuracy of 94.4%. TMB-Hunt (Garrow et al., 2005) identified 332 OMPs with the cutoff of two ($\log[\text{prob(OMP)}/\text{prob(non-OMP)}] > 2$) and this method could detect 91.7% of the *E. coli* OMPs in TCDB. On the other hand, BOMP (Berven et al., 2004) detected only 80.6% of the *E. coli* OMPs deposited in TCDB and there is a superimposition of 72 sequences between BOMP and our method. This analysis shows that although our method has several false positives it has the ability of picking up the real OMPs in genomes with high sensitivity.

3.6. Prediction on the Web

We have developed a web server, TMBETADISC-RBF, for discriminating OMPs from amino acid sequence. It takes the amino acid sequence in FASTA format as input and displays the type of the protein (OMP or non-OMP) along with the protein name in the output. Further, the users have the feasibility of selecting the method (amino acid composition (20D), residue pair preference (400D), combination of them (420D) and PSSM profiles. It is available at http://rbf.bioinfo.tw/ sachen/OMP.html. Our method can be used to detect new OMPs as well as to annotate OMPs in genomic sequences.

References

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Bagos, P., Liakopoulos, T., Spyropoulos, I., Hamodrakas, S., 2004. A hidden markov model method, capable of predicting and discriminating beta-barrel outer membrane proteins. BMC Bioinformatics 5 (1), 29.

Berven, F., Flikka, K., Jensen, H., Eidhammer, I., 2004. BOMP: a program to predict integral b-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic Acids Res. 32, W394–W399.

Bigelow, H., Rost, B., 2006. PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. Nucleic Acids Res. 34 (web server issue), W186.

Chandonia, J., Hon, G., Walker, N., Conte, L., Koehl, P., Levitt, M., Brenner, S., Journals, O., 2004. The ASTRAL compendium in 2004. Nucleic Acids Res. 32, D189–D192.

Garrow, A., Agnew, A., Westhead, D., 2005. TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. Nucleic Acids Res. 33, W188–W192.

Garrow, A.G., Westhead, D.R., 2007. A consensus algorithm to screen genomes for novel families of transmembrane beta barrel proteins. Proteins 69 (1), 8–18.

Gnanasekaran, T., Peri, S., Arockiasamy, A., Krishnaswamy, S., 2000. Profiles from structure based sequence alignment of porins can identify β stranded integral membrane proteins. Bioinformatics 16, 839–842.

Gromiha, M.M., 2005. Motifs in outer membrane protein sequences. Biophys. Chem. 117, 65–71.

Gromiha, M.M., Ahmad, S., Suwa, M., 2005. Application of residue distribution along the sequence for discriminating outer membrane proteins. Comput. Biol. Chem. 29, 135–142.

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

Gromiha, M.M., Suwa, M., 2006. Discrimination of outer membrane proteins using machine learning algorithms. Proteins: Structure, Function, and Bioinformatics 63, 1031–1037.

Gromiha, M.M., Suwa, M., 2007. Current developments on-barrel membrane proteins: sequence and structure analysis, discrimination and prediction. Curr. Protein Pept. Sci. 8 (6), 580–599.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.

Li, W., Jaroszewski, L., Godzik, A., 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17, 282–283.

Liu, Q., Zhu, Y., Wang, B., Li, Y., 2003. Identification of b-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. Comput. Biol. Chem. 27, 355–361.

Martelli, P.L., Fariselli, P., Krogh, A., Casadio, R., 2002. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics 18, 46–53.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.

Natt, N.K., Kaur, H., Raghava, G.P.S., 2004. Prediction of transmembrane regions of beta-barrel proteins using ann and svm-based methods. Proteins 56, 11–18.

Ou, Y.-Y., Oyang, Y.-J., Chen, C.-Y., 2005. A novel radial basis function network classifier with centers set by hierarchical clustering. In: IJCNN'05. Proceedings, 3, pp. 1383–1388.

Park, K.-J., Gromiha, M.M., Horton, P., Suwa, M., 2005. Discrimination of outer membrane proteins using support vector machines. Bioinformatics 21, 4223–4229.

Press, W.H., 1992. Numerical Recipes in C, 2nd edition. Cambridge University Press, Cambridge.

Saier, M., Tran, C., Barabote Jr., R., 2006. TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Res. 34, D181–D186.

Su, C.-T., Chen, C.-Y., Ou, Y.-Y., 2006. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC Bioinformatics 7 (1), 319.

Takasaki, S., Kawamura, Y., Konagaya, A., 2006. Selecting effective siRNA sequences by using radial basis function network and decision tree learning. BMC Bioinformatics 7 (5), S22.

Tusnády, G., Dosztányi, Z., Simon, I., 2005. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res. 33, D275–D278.

Witten, I., Frank, E., 2005. Data Mining:: Practical Machine Learning Tools and Techniques, 2nd edition. Morgan Kaufmann.

Xie, D., Li, A., Wang, M., Fan, Z., Feng, H., 2005. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res. 33 (1), W105–W110.

Yang, Z., Thomson, R., 2005. Bio-basis function neural network for prediction of protease cleavage sites in proteins. Neural Netw. IEEE Trans. 16 (1), 263–274.

Zhang, G., Huang, D., 2004. Prediction of inter-residue contacts map based on genetic algorithm optimized radial basis function neural network and binary input encoding scheme. J. Comput. Aided Mol. Des. 18 (12), 797–810.