BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 15 2011, pages 2062–2067
doi:10.1093/bioinformatics/btr340

*Sequence analysis*

Advance Access publication June 8, 2011

# Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties

Shu-An Chen[1], Yu-Yen Ou[1,*], Tzong-Yi Lee[1] and M. Michael Gromiha[2,*]

[1]Department of Computer Science and Engineering, Yuan Ze University, Chung-Li, Taiwan, [2]Department of Biotechnology, Indian Institute of Technology (IIT) Madras, Chennai 600 036, India

Associate Editor: Martin Bishop

## ABSTRACT

**Summary:** Transporters are proteins that are involved in the movement of ions or molecules across biological membranes. Currently, our knowledge about the functions of transporters is limited due to the paucity of their 3D structures. Hence, computational techniques are necessary to annotate the functions of transporters. In this work, we focused on an important functional aspect of transporters, namely annotation of targets for transport proteins. We have systematically analyzed four major classes of transporters with different transporter targets: (i) electron, (ii) protein/mRNA, (iii) ion and (iv) others, using amino acid properties. We have developed a radial basis function network-based method for predicting transport targets with amino acid properties and position specific scoring matrix profiles. Our method showed a 10-fold cross-validation accuracy of 90.1, 80.1, 70.3 and 82.3% for electron transporters, protein/mRNA transporters, ion transporters and others, respectively, in a dataset of 543 transporters. We have also evaluated the performance of the method with an independent dataset of 108 proteins and we obtained similar accuracy. We suggest that our method could be an effective tool for functional annotation of transport proteins.

**Availability:** http://rbf.bioinfo.tw/~sachen/ttrbf.html

**Contact:** yien@csie.org; gromiha@iitm.ac.in

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on March 7, 2011; revised on May 23, 2011; accepted on May 25, 2011

## 1 INTRODUCTION

Membrane proteins play a vital role in living organisms, such as transport of ions and molecules across membranes, binding to small molecules within extracellular space, recognition processes in immune system, energy transduction, and so on. Transporters are one of the major classes of membrane proteins, spanning cell membranes and forming an intricate system of pumps and channels. They are different in their membrane topologies, energy coupling mechanisms and the specifics of their substrates (Ren *et al.*, 2007; Saier, 2000). Transporters are generally classified into channels/pores, electrochemical and active transporters, group translocators, electron carriers, as well as factors involved in transport systems (Saier *et al.*, 2006). The transport primarily involves ions, proteins, mRNA and electrons. The classification of transporters based on different families as well as their targets remains an important problem for the advancement of structural and functional genomics.

Recently, few methods have been proposed for discriminating transport proteins based on their functions. Gromiha and Yabuki (2008) analyzed the amino acid composition of transporters and developed a neural network-based method for classifying them into channels/pores, electrochemical and active transporters. Li *et al.* (2008, 2009) utilized nearest neighbor and hidden Markov model methods, which integrate sequence similarity search and topological analysis into a machine-learning framework for categorizing transporters. In our previous work (Ou *et al.*, 2010), we have systematically analyzed the amino acid composition, residue pair preference and amino acid properties in six different families of transporters. Utilizing the information, we have developed a radial basis function (RBF) network method based on profiles obtained with position-specific scoring matrices (PSSMs) for discriminating transporters belonging to three different classes and six families.

In this work, we focused on another important aspect of functional annotation of transporters, i.e. the prediction of transporter targets. First, we constructed a transporter database with target information from the latest version of UniProt database (The UniProt Consortium, 2010). The database was then divided into four major classes of transporters based on their transport targets. The four classes are electron transporters, protein/mRNA transporters, ion transporters and other transporters. We have systematically analyzed the characteristic features of amino acid residues and developed a radial basis network for discriminating transporter targets using amino acid properties and PSSM profiles. Our method showed a 10-fold cross-validation accuracy of 90.1, 80.1, 70.3 and 82.3% for electron transporters, protein/mRNA transporters, ion transporters and others. We evaluated the performance of this method with an independent dataset of 108 proteins and obtained an accuracy of 92.6, 77.8, 69.4 and 80.6% for electron transporters, protein/mRNA transporters, ion transporters and others, respectively. A web server has been developed for discriminating transporters based on different targets and it is available online for the users. Finally, we developed a protocol to analyze newly discovered genomic sequences to annotate putative transporters and their transport targets. We propose that our method could be an effective tool in the annotation of transporters in genomic sequences.

*To whom correspondence should be addressed.

## 2 METHODS

### 2.1 Dataset

We constructed a dataset of 2452 transporters, which are known at protein level and clear target annotation in the UniProt database (The UniProt Consortium, 2010). Using BLAST (Altschul *et al.*, 1997), we removed the sequences that have more than 20% identity. The final dataset contains 651 transporters with 98 electron transporters, 266 protein/mRNA transporters, 200 ion (ammonia, calcium, hydrogen ion, chloride, potassium, sodium, phosphate, sulfate, cobalt, nickel, copper, iron and neurotransmitter) transporters and 87 other transporters.

From the dataset, we have randomly selected 17 electron, 44 protein/mRNA, 33 ion and 14 other transporters for independent tests. The remaining data were used for 10-fold cross-validation. The detailed statistics of the dataset is listed in Supplementary Table S1. The overlap with the data available in Transport Classification Database is given in Supplementary Table S2.

### 2.2 Design of the RBF networks

We have employed the QuickRBF package (Ou, 2005) to construct RBFN classifiers in this work. The architecture of RBF network is shown in Figure 1. As presented in Figure 1, a general 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. Each node in the hidden layer then produces an activation based on the associated radial basis kernel function. Finally, each node in the output layer computes a linear combination of the activations of the hidden nodes. The general mathematical form of the output nodes in RBFN is as follows:

$$c_j(x)=\sum_{i=1}^{k} w_{ji}\phi(\|x-\mu_i\|;\sigma_i); \quad (1)$$

where $c_j(x)$ denotes the function corresponding to the $j$-th output node and it is a linear combination of $k$ RBFs $\phi()$ with center $\mu_i$ and bandwidth $\sigma_i$. Also, $W_{ji}$ denotes the weight associated with the correlation between the $j$-th output node and the $i$-th hidden node.

A fixed bandwidth of 5 for each kernel function is employed in the network. We have carried out the computation with different bandwidths 1, 5, 10 and 20, and we obtained similar results. In addition, we used all training data as centers. Then, the RBFN classifier identifies four types of transporters based on the output function value. More details about network structure and design have been explained in our earlier article (Ou *et al.*, 2005).

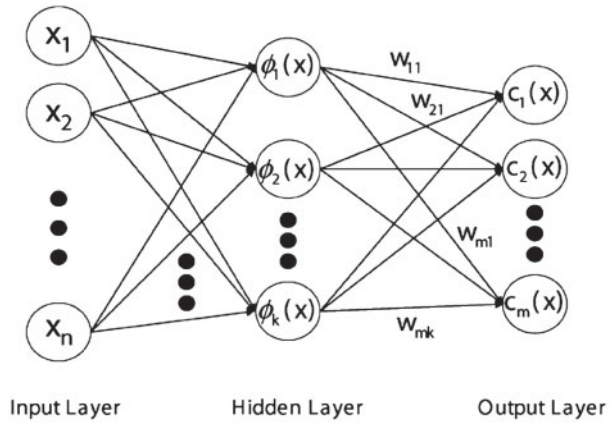


**Fig. 1.** The architecture of RBF network. The input, hidden and output layers are shown with associated weights.

Classification based on RBF networks has several applications in bioinformatics. It has been widely used to predict the cleavage sites in proteins (Yang and Thomson, 2005), interresidue contacts (Zhang and Huang, 2004), protein disorder (Su *et al.*, 2006) and the discrimination of $\beta$-barrel membrane proteins (Ou *et al.*, 2008).

### 2.3 Compositions of amino acids and amino acid pairs

We used $n$ vectors $\{x_i, i=1,\ldots,n\}$, to represent all $n$ proteins in the training data. Each vector was labeled to show the protein groups (e.g. electrochemical transporters or active transporters). The vector $x_i$ has 20 elements for the composition of amino acids, and 400 elements for the composition of amino acid pairs. The 20 elements show the number of occurrences of 20 amino acids normalized with total number of residues in a protein, and the 400 elements show the number of occurrences of those 400 amino acid pairs normalized with the total number of residues in a protein. Further, we have used the combinations of amino acid and residue pair compositions with 420 elements in each vector.

### 2.4 PSSM profiles

From the structural point of view, several amino acid residues could be mutated without altering the structure of the protein, making it possible that two proteins could share similar structures with different amino acid compositions. Hence, we have adopted the PSSM profiles, which have been widely used in protein secondary structure prediction, subcellular localization and other bioinformatics problems with notable improvement (Jones, 1999; Ou *et al.*, 2008; Xie *et al.*, 2005). The PSSM profiles were obtained with PSI-BLAST and the non-redundant (NR) protein database.

In the classification of transport proteins, we used PSSM profiles to generate 400 dimension ($20\times20$ residue pairs) input vectors as input features by summing up each row of same amino acid in the PSSM profiles and the variable is denoted as 'x'. Supplementary Figure S1 shows the details of generating the 400D (dimension) PSSM features from original PSSM profiles. Every element of 400D input vector was divided by the length of the sequence and then be scaled by $\frac{1}{1+e^{-x}}$ for normalizing the values between 0 and 1.

### 2.5 Biochemical properties

To enhance prediction performance, we included the properties of amino acid residues as new features. In this study, we analyzed the composition of 20 amino acids, the composition of 400 amino acid pairs and 544 the biochemical properties in the AAindex database (Kawashima *et al.*, 2008).

We divided the amino acid sequences of each protein into four equal parts, and computed average value of biochemical properties for each part. Then, we calculated the $F$-score for four parts individually using Equation (1) and obtained the average. The properties were ranked with the average of four $F$-score values.

We added these topmost ranking biochemical properties one by one to the PSSM feature sets according to their $F$-score value, and kept the property in the feature set if it improved the performance via 10-fold cross-validation. The $F$-score of the $i$-th feature is defined as:

$$F\text{-score}(i)=\frac{\left(\bar{x}_i^{(+)}-\bar{x}_i\right)^2+\left(\bar{x}_i^{(-)}-\bar{x}_i\right)^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}\left(x_{k,i}^{(+)}-\bar{x}_i^{(+)}\right)^2+\frac{1}{n^--1}\sum_{k=1}^{n^-}\left(x_{k,i}^{(-)}-\bar{x}_i^{(-)}\right)^2} \quad (2)$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$ and $\bar{x}_i^{(-)}$ are the average of the $i$-th feature of the whole, positive and negative datasets, respectively; $n^+$ is the number of positive dataset and $n^-$ is the number of negative dataset; $x_{k,i}^{(+)}$ is the $i$-th feature of the $k$-th positive instance, and $x_{k,i}^{(-)}$ is the $i$-th feature of the $k$-th negative instance (Chen and Lin, 2006). For each classification, specific target (e.g. electron transport) is treated as positive dataset and rest of them (protein/mRNA, ion and others) constitute negative dataset. It may be noted that the properties

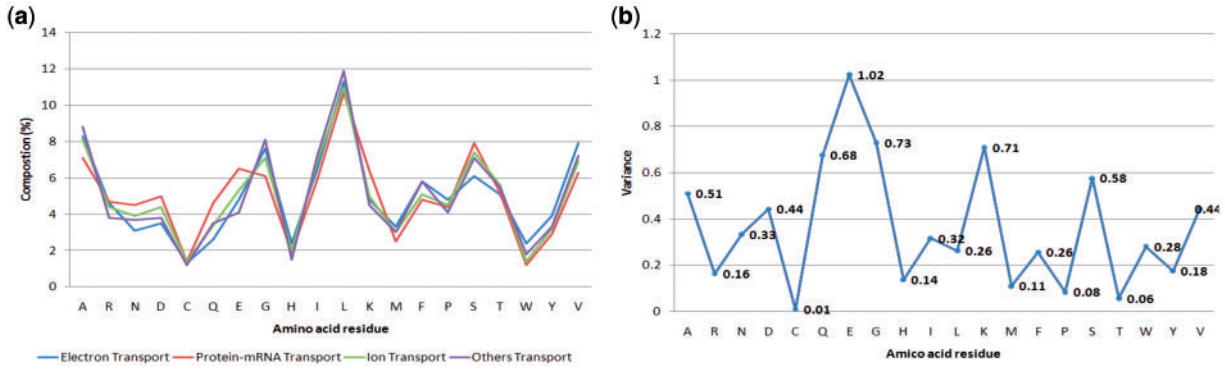


**Fig. 2.** (**a**) Amino acid composition in four classes of transporters; (**b**) variance of 20 amino acid residues among four classes of transporters, electron, protein/mRNA, ion and others.

are selected with 'best fit' for 10-fold cross-validation. However, we have also evaluated the method using an independent dataset of 108 transporters, which verifies the reliability of results.

### 2.6 Assessment of predictive ability

The prediction performance was examined by 10-fold cross-validation test, in which the four types of proteins were randomly divided into 10 subsets of approximately equal size. We trained the data with nine subsets and the remaining set was used to test the performance of the method. This process was repeated 10 times so that every subset had been used as the test data once.

We used sensitivity, specificity, accuracy and Matthew's correlation coefficient (MCC) to measure the prediction performance. TP, FP, TN, FN are true positives, false positives, true negatives and false negatives, respectively.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \quad (4)$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \quad (5)$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (6)$$

## 3 RESULTS

### 3.1 Amino acid composition in various classes of transporters

Figure 2 shows the computed composition of amino acids and their variance in four types of transporters. Figure 2a showed that the composition of Gln, Ser and Trp in electron transporters was notably different from that of other classes. The variance of 20 amino acid residues between four classes can be seen in Figure 2b. The residues Glu, Gly, Lys, Gln, Ser and Ala had a variance higher than 0.5. Interestingly, the variance of these residues was different from other classes and due mainly to the composition of the residues in protein/mRNA transporters.

### 3.2 Dipeptide composition preference in various classes of transporters

We computed the residue pair preference (dipeptide composition) for all 400 possible residue pairs in electron transporters, protein/mRNA transporters, ion transporters and others. Based on the dipeptide composition, we computed the variance, and the residue pairs that have the topmost variances are listed in Supplementary Table S3. The residue pairs, GL, VF, LG and GA showed a variance greater than 0.03; and the residue pair with Gly showed the highest variance among the four classes of transporters.

**Table 1.** Features used for classifying transporters

| | Properties |
|---|---|
| Electron transport | Q, W, S, N, D, Y, V |
| Protein/mRNA transport | Q, E, K |
| Ion transport | B-factors for amino acid residues |
| Others | AG, EE, GA |

### 3.3 Amino acid properties with high F-score for four classes of transporters

We computed the *F*-score for the composition of 20 amino acids, 400 amino acid pair compositions and 544 biochemical properties in the AAindex database (Kawashima *et al.*, 2008). The properties with topmost scores are given in Supplementary Table S4. These properties were added to PSSM profiles and their influence was examined with the ability of improving the accuracy of discrimination. The features selected for the classification of each transporter (electron, protein/mRNA, ion and others) are presented in Table 1. Interestingly, the influence of amino acid properties is vital for all four classes of transporters (Table 2).

### 3.4 Importance of selected features for the structure and function of proteins

We have used amino acid composition, residue pair preference, biochemical properties and evolutionary information in the form of PSSM profiles as main features in the present study. These features play an important role to classify proteins based on their structure and function. It has been shown that the positive charged residues are overrepresented in the binding regions of DNA and RNA binding proteins to interact with the DNA/RNA (Ahmad *et al.*, 2004; Bhardwaj and Lu, 2007; Jeong *et al.*, 2003; Jones *et al.*, 2001; Kumar *et al.*, 2008; Terribilini *et al.*, 2006;

**Table 2.** Discrimination of four classes of transporters with different features

| Method | Sensitivity (%) | | | | Specificity (%) | | | | Accuracy (%) | | | | MCC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T4 | T1 | T2 | T3 | T4 | T1 | T2 | T3 | T4 | T1 | T2 | T3 | T4 |
| Cross-validation dataset (543 proteins) | | | | | | | | | | | | | | | | |
| PSI-BLAST | 42.0 | 66.7 | 65.3 | 45.2 | 95.5 | 78.8 | 74.7 | 92.6 | 87.5 | 73.8 | 71.8 | 86.2 | 0.44 | 0.46 | 0.38 | 0.39 |
| AAC | 56.8 | 72.1 | 59.9 | 57.5 | 91.1 | 74.1 | 66.2 | 81.3 | 86.0 | 73.3 | 64.3 | 78.1 | 0.47 | 0.46 | 0.24 | 0.31 |
| DPC | 63.0 | 70.3 | 67.1 | 56.2 | 85.9 | 69.5 | 61.2 | 80.2 | 82.5 | 69.8 | 63.0 | 77.0 | 0.42 | 0.39 | 0.26 | 0.29 |
| AAC + DPC | 64.2 | 71.2 | 65.3 | 58.9 | 85.3 | 72.0 | 62.0 | 80.6 | 82.1 | 71.6 | 63.0 | 77.7 | 0.42 | 0.43 | 0.25 | 0.31 |
| PSSM | 70.4 | 78.4 | 68.3 | 69.9 | 92.4 | 79.8 | 69.9 | 83.8 | 89.1 | 79.2 | 69.4 | 82.0 | 0.60 | 0.58 | 0.36 | 0.43 |
| PSSM + properties | 71.6 | 80.2 | 70.7 | 72.6 | 93.3 | 80.1 | 70.2 | 83.8 | 90.1 | 80.1 | 70.3 | 82.3 | 0.62 | 0.60 | 0.38 | 0.45 |
| Independent dataset (108 proteins) | | | | | | | | | | | | | | | | |
| PSI-BLAST | 41.2 | 63.6 | 54.5 | 21.4 | 93.4 | 79.7 | 65.3 | 92.6 | 85.2 | 73.1 | 62.0 | 83.3 | 0.39 | 0.44 | 0.19 | 0.16 |
| AAC | 58.8 | 70.5 | 51.5 | 57.1 | 92.3 | 67.2 | 62.7 | 83.0 | 87.0 | 68.5 | 59.3 | 79.6 | 0.51 | 0.37 | 0.13 | 0.32 |
| DPC | 47.1 | 77.3 | 60.6 | 57.1 | 91.2 | 71.9 | 61.3 | 74.5 | 84.3 | 74.1 | 61.1 | 72.2 | 0.39 | 0.48 | 0.20 | 0.23 |
| AAC + DPC | 47.1 | 79.5 | 57.6 | 57.1 | 91.2 | 76.6 | 65.3 | 75.5 | 84.3 | 77.8 | 63.0 | 73.1 | 0.39 | 0.55 | 0.21 | 0.24 |
| PSSM | 76.5 | 72.7 | 63.6 | 50.0 | 94.5 | 79.7 | 69.3 | 84.0 | 91.7 | 76.9 | 67.6 | 79.6 | 0.69 | 0.52 | 0.31 | 0.28 |
| PSSM + properties | 76.5 | 75.0 | 63.6 | 64.3 | 95.6 | 79.7 | 72.0 | 83.0 | 92.6 | 77.8 | 69.4 | 80.6 | 0.72 | 0.54 | 0.34 | 0.38 |

T1, electron transporters; T2, protein/mRNA transporters; T3, ion transporters; T4, other transporters; BLAST : simple sequence similarity search; AAC, amino acid composition; DPC, dipeptide composition (residue pair preference).

Wu *et al.*, 2009). The hydrophobic residues are accumulated in the membrane spanning regions of transmembrane helical proteins. Hence, the concept of conformational parameters and physicochemical properties has been widely used for predicting the membrane spanning segments of α-helical membrane proteins (Gromiha, 1999; Hirokawa *et al.*, 1998; Tusnady and Simon, 1998). The membrane spanning regions of β-barrel membrane proteins showed a periodicity of polar–non-polar residues and the residue pair preference has been used to discriminate such class of proteins (Gromiha *et al.*, 2005). Further, these features have been used in several classifications such as mesophilic and thermophilic proteins, binding regions in protein complexes, etc. (Berezovsky *et al.*, 2007; Gromiha and Suresh, 2008; Zhang and Fang, 2006). The PSSM profiles provide the evolutionary information, which characterize the proteins with similar sequence and function (Jones, 1999; Ou *et al.*, 2008). These analyses showed the importance of the features used in the present study for understanding the structures and functions of proteins.

## 3.5 Discrimination of transporters based on four different classes of transporters

We developed a variety of methods for annotating electron transporters, protein/mRNA transporters, ion transporters and others. The results obtained from the composition of amino acids, residue pair preference, combinations of them, PSSM and the combination of PSSM with the properties of amino acid residues are presented in Table 2.

The results showed that PSSM with amino acid properties was successful in discriminating transporters with an average 10-fold cross-validation accuracy of 90.1, 80.1, 70.3 and 82.3% for electron transporters, protein/mRNA transporters, ion transporters and others, respectively.

We have carried out receiver operator characteristic (ROC) analysis and the results for four transporters are shown in Figure 3. Our results showed the area under the curve (AUC) of 0.90, 0.86, 0.77 and 0.86, respectively, for electron, protein/mRNA, ion and other transporters using PSSM and biochemical properties (Supplementary Table S5).

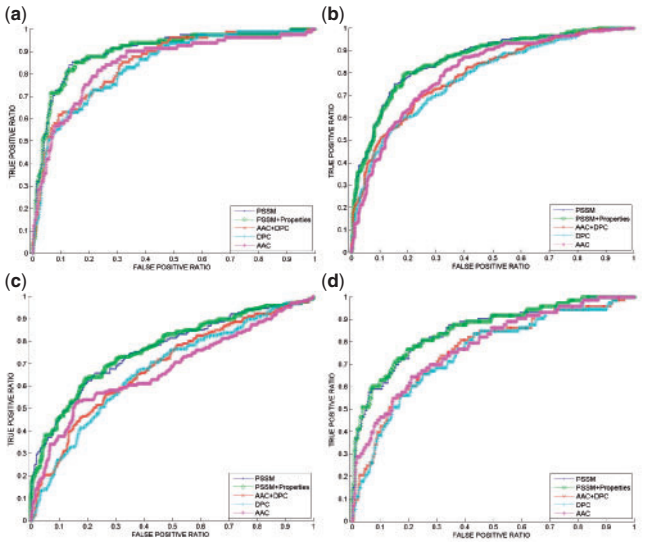


**Fig. 3.** Comparison of ROC curve for different features. The results obtained with amino acid composition (AAC), dipeptide composition (DPC), combination of AAC and DPC (AAC + DPC), PSSM and combination of PSSM and properties (PSSM + properties) are shown. (**A**) Electron transporter; (**B**) protein/mRNA transporter; (**C**) ion transporter; and (**D**) other transporter.

We evaluated the performance of this method with an independent dataset of 108 proteins and obtained an accuracy of 92.6, 77.8, 69.4 and 80.6% for electron transporters, protein/mRNA transporters, ion transporters and others. Our analysis showed that PSSM profiles and the properties of amino acid residues improved the accuracy of discrimination, compared with the composition of amino acids and

**Table 3.** The comparison between different classifiers

| Method: PSSM + properties | Sensitivity (%) | | | | Specificity (%) | | | | Accuracy (%) | | | | MCC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T4 | T1 | T2 | T3 | T4 | T1 | T2 | T3 | T4 | T1 | T2 | T3 | T4 |
| Cross-validation dataset (543 proteins) | | | | | | | | | | | | | | | | |
| Decision Trees (J48) | 53.1 | 59.3 | 38.1 | 26.9 | 89.6 | 66.7 | 74.4 | 88.2 | 84.2 | 63.5 | 62.7 | 80.3 | 0.41 | 0.24 | 0.10 | 0.15 |
| Naïve Bayes | 66.7 | 73.9 | 69.0 | 70.3 | 84.3 | 64.5 | 53.5 | 76.4 | 83.0 | 68.3 | 60.0 | 75.8 | 0.40 | 0.38 | 0.20 | 0.34 |
| KNN | 63.0 | 76.1 | 65.3 | 45.2 | 90.5 | 70.4 | 66.0 | 89.6 | 86.4 | 72.7 | 65.7 | 83.6 | 0.50 | 0.46 | 0.29 | 0.33 |
| Random Forest | 69.1 | 73.9 | 65.9 | 57.5 | 90.0 | 76.0 | 72.6 | 77.9 | 86.9 | 73.3 | 70.5 | 75.1 | 0.54 | 0.46 | 0.36 | 0.27 |
| Lib-SVM | 70.4 | 77/9 | 70.1 | 71.2 | 91.8 | 76.6 | 70.2 | 83.6 | 88.6 | 77.2 | 70.2 | 82.0 | 0.58 | 0.54 | 0.38 | 0.44 |
| QuickRBF | 71.6 | 80.2 | 70.7 | 72.6 | 93.5 | 80.1 | 70.2 | 83.8 | 90.1 | 80.1 | 70.3 | 82.3 | 0.62 | 0.60 | 0.38 | 0.45 |
| Independent dataset (108 proteins) | | | | | | | | | | | | | | | | |
| Decision Trees (J48) | 30.1 | 54.9 | 36.4 | 28.7 | 94.5 | 74.2 | 73.1 | 89.5 | 86.1 | 67.2 | 63.4 | 82.0 | 0.31 | 0.30 | 0.09 | 0.19 |
| Naïve Bayes | 59.4 | 80.1 | 73.8 | 84.3 | 88.1 | 60.6 | 48.0 | 74.6 | 87.1 | 69.2 | 56.2 | 76.9 | 0.44 | 0.41 | 0.21 | 0.43 |
| KNN | 70.6 | 72.7 | 69.7 | 57.1 | 93.4 | 73.4 | 50.7 | 87.2 | 89.8 | 73.1 | 56.5 | 83.3 | 0.63 | 0.46 | 0.19 | 0.38 |
| Random Forest | 69.1 | 81.8 | 57.6 | 71.4 | 90.0 | 78.1 | 62.7 | 79.8 | 86.9 | 79.6 | 61.1 | 78.7 | 0.54 | 0.59 | 0.19 | 0.39 |
| Lib-SVM | 76.5 | 75.0 | 66.7 | 57.1 | 94.5 | 79.7 | 66.7 | 83.0 | 91.7 | 77.8 | 66.7 | 79.6 | 0.69 | 0.54 | 0.31 | 0.32 |
| QuickRBF | 76.5 | 75.0 | 63.6 | 64.3 | 95.6 | 79.7 | 72.0 | 83.0 | 92.6 | 77.8 | 69.4 | 80.6 | 0.72 | 0.54 | 0.34 | 0.38 |

T1, electron transporters; T2, protein/mRNA transporters; T3, ion transporters; T4, other transporters.

dipeptides. We achieved a correlation of 0.72, 0.54, 0.34 and 0.38 in the test set for the four classes, which was 10–49% improvement over the results obtained with other features. The usage of PSSM profiles and biochemical properties might be the reason for this improvement.

### 3.6 Comparison with other methods

We have analyzed the capability of PSI-BLAST to discriminate each of the four types of transporter targets, based on sequence similarity searches. We have examined all transporters in the 543 training and 108 test sets of data and computed the sensitivity, specificity, accuracy and MCC. This method showed accuracy in the range of 64–88% for discriminating the four classes (electron, protein/mRNA, ion and others) of transporters. Our proposed method showed the accuracy of 69–92%, which is superior to PSI-BLAST searches for discrimination. In addition, the simple sequence similarity search method showed a lower sensitivity of approximately 10–40% than our proposed method in the test set of 108 proteins. The detailed results of sequence similarity search are also included in Table 2.

In addition, we have compared the performance of the present method with other algorithms such as decision trees, *k*-nearest neighbors, support vector machines, random forest, etc. and the results are presented in Table 3. The ROC analysis has been done with all classifiers and the results are presented in Supplementary Figure S2 and Table S6. We noticed that our method performs better than other methods in terms of sensitivity, specificity, accuracy, MCC and AUC for discriminating all types of transporter targets.

### 3.7 Importance of the work in biological context and application to new sequences

The annotation of proteins based on their structure and function are important in structural and functional genomics. In our earlier study, we have developed methods to discriminate transporters from other proteins and classify them into three classes and six families (Gromiha and Yabuki, 2008; Ou *et al.*, 2010). It is necessary and important to classify them based on transporting targets as the efficiency, activity, transport and other functions depends on targets. This can be evidenced with the database for functionally important residues in membrane proteins and drug–target interactions (Gromiha *et al.*, 2009). Hence, the classification based on targets used in the present study is biologically relevant to understand the functions.

We developed a protocol for predicting the target of a transporter. In this procedure, we initially examined the query sequence whether it is a transporter or not (Ou *et al.*, 2010). For transporters, we applied PSSM features and other properties to classify into four types of transporter targets. This procedure yields four results (for each transport type) and the transporter target is assigned according to the greatest preference (Fig. 4). This is a sequence-based method, which could be used to annotate genomic sequences. As an example, we utilized this method to annotate the transporters and different classes of transporter targets in *Escherichia coli* genome, with 4237 sequences. Our method detected 67, 225, 262 and 155 proteins as electron transporters, protein/mRNA transporters, ion transporters and others, respectively. Further investigations on these proteins are on progress.

### 3.8 Discrimination on the web

We have developed a web server for discriminating membrane transport proteins based on their targets, (i) electron, (ii) protein/mRNA, (iii) ion and (iv) others. It takes the amino acid sequence in FASTA format as input and predicts the type of the target for membrane transport protein. The server can be freely accessible at http://rbf.bioinfo.tw/~sachen/ttrbf.html.

## 4 CONCLUSIONS

This study focused on methods for the prediction of targets in transport proteins. We analyzed four major classes of transporters with different transporter targets, such as electron, protein/mRNA, ion and others, and revealed the important amino acid residues,

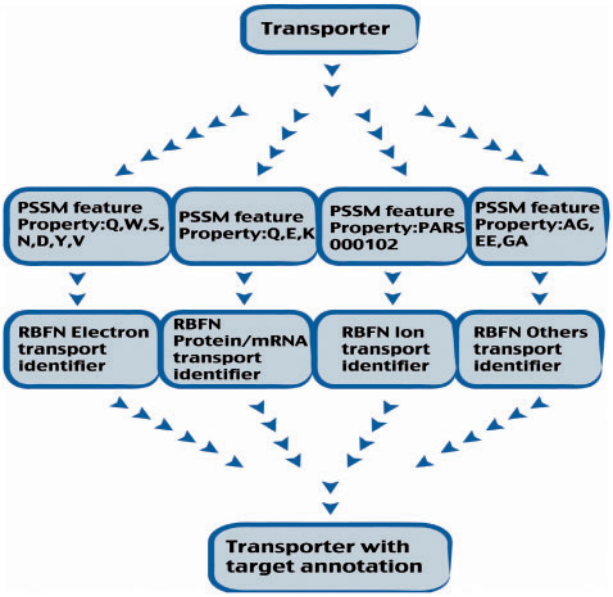


**Fig. 4.** The architecture for annotating transporters targets with three steps: (i) PSSM profiles for specific features; (ii) RBF networks for each target; and (iii) final classification.

residue pairs and amino acid properties. We developed a radial basis network for transporter target annotation using amino acid properties and PSSM profiles. Our method showed a 10-fold cross-validation accuracy of 90.1, 80.1, 70.3 and 82.3% for electron transporters, protein/mRNA transporters, ion transporters and others, respectively. We evaluated the performance of the method with an independent dataset of 108 proteins and we obtained similar results. Based on the results, we have developed a protocol for identifying transporters and predicting their transporting targets. We suggest that our method would serve as an effective tool for the functional annotation of membrane proteins.

## ACKNOWLEDGEMENTS

We thank the reviewers for constructive comments.

*Funding*: Indian Institute of Technology Madras research grant (BIO/10-11/540/NFSC/MICH to M.M.G.). National Science Council (NSC) of Taiwan, NSC99-2221-E155-073 (to Y-Y. O.); and NSC99-2320-B155-001 (to T-Y. L.).

*Conflict of Interest*: none declared.

## REFERENCES

Ahmad,S. *et al.* (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.

Altschul,S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Berezovsky,I.N. *et al.* (2007) Positive and negative design in stability and thermal adaptation of natural proteins. *PLoS Comput. Biol.*, **3**, 498–507.

Bhardwaj,N. and Lu,H. (2007) Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS Lett.*, **581**, 1058–1066.

Chen,Y.W. and Lin,C.J. (2006) Combining SVMs with various feature selection strategies. In Guyon,I. *et al.* (eds) *Feature Extraction: Foundations and Applications*, Springer, Heidelberg, pp. 315–324.

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

Gromiha,M.M. and Suresh,M.X. (2008) Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins*, **70**, 1274–1279.

Gromiha,M.M. and Yabuki,Y. (2008) Functional discrimination of membrane proteins using machine learning techniques. *BMC Bioinformatics*, **9**, 135.

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

Gromiha,M.M. *et al.* (2009) TMFunction: database for functional residues in membrane proteins. *Nucleic Acids Res.*, **37**, D201–D204.

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

Jeong,E. *et al.* (2003) Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes. *Mol. Cells*, **16**, 161–167.

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

Jones,S. *et al.* (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.

Kawashima,S. *et al.* (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202.

Kumar,M. *et al.* (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, **71**, 189–194.

Li,H. *et al.* (2008) A nearest neighbor approach for automated transporter prediction and categorization from protein sequences. *Bioinformatics*, **24**, 1129.

Li,H. *et al.* (2009) TransportTP: A two-phase classification approach for membrane transporter prediction and characterization. *BMC Bioinformatics*, **10**, 418.

Ou,Y.-Y. (2005) QuickRBF: a package for efficient radial basis function networks. Software available at http://csie.org/~yien/quickrbf/.

Ou,Y.-Y. *et al.* (2005) A novel radial basis function network classifier with centers set by hierarchical clustering. *Proc. IJCNN'05* , **3**, 1383–1388.

Ou,Y.-Y. *et al.* (2008) TMBETADISC-RBF: discrimination of $\beta$-barrel membrane proteins using RBF networks and PSSM profiles. *Comput. Biol. Chem.*, **32**, 227–231.

Ou,Y.-Y. *et al.* (2010) Classification of transporters using efficient radial basis function networks with position-specific scoring matrices and biochemical properties. *Proteins*, **78**, 1789–1797.

Ren,Q. *et al.* (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res.*, **35**, D274–D279.

Saier,M.H. Jr (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.*, **64**, 354–411.

Saier,M.H. Jr *et al.* (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.*, **34**, D181–D186.

Su,C.-T. *et al.* (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics*, **7**, 319.

Terribilini,M. *et al.* (2006) Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, **12**, 1450–1462.

The UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.

Tusnady,G.E. and Simon,I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, **283**, 489–506.

Wu, J.S. *et al.* (2009) Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, **25**, 30–35.

Xie,D. *et al.* (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.*, **33**, W105–W110.

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

Zhang,G.Y. and Fang,B.S. (2006) Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. *Process Biochem.,* **41**, 1792–1798.

Zhang,G.Z. and Huang,D.S. (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**, 797–810.