RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

A consensus subunit-specific model for annotation of substrate specificity for ABC transporters

Yayun Hu, Yanzhi Guo*, Yinan Shi, Menglong Li, Xuemei Pu*

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

Members of ATP-binding cassette (ABC) transporter family are present in three kingdoms of life and play a vital role in most cellular functions. ABC transporters function as either importers that bring nutrients and other molecules into cells, or as exporters that pump toxins, drugs and lipids across membranes. Currently, the limitation of 3D structures highlights the importance of the functional annotation for transporters using bioinformatics-based methods. In this work, we focused on annotation of substrate specificity for ABC transporters. Three types of the subunit proteins of ABC transporters, namely permease protein, ATP-binding protein and substrate binding protein all contribute much to transport process, but have unique structures and properties. However previous computational methods only consider the three subunit proteins as the same and cannot individually characterize each type of subunit proteins. Here, through individual feature evaluation and selection, specific representation for each type of subunit proteins was implemented. Then three subunit-specific models were built to consistently analyse four major classes of ABC transporters with different transport targets. Our method achieved a 5-fold cross validation accuracy of 93.35%, 84.34%, 87.24% and 81.96% for sugar transporter, ion transporter, amino acid/protein transporter and others, respectively. Our method also showed an overall prediction accuracy of 88.02% with a Mathew's correlation coefficient of 0.6736 on an independent dataset. The results suggest that considering three subunit proteins separately and developing individual models for three substrate protein groups are recommendable. This method would be an effective tool for computational annotation of substrate specificity for ABC transporters.

1. Introduction

Transporter is a necessary medium for transport process that is an essential biological action for all life and endows many processes, including metabolism, communication, biosynthesis, and reproduction. They allow the entry of all essential nutrients into cell and efflux toxic substance to provide cell a benign growing environment [1]. The importance of transporters to cells can been illustrated by the fact that transporters typically make up 5–15% of the total gene content of sequenced organisms [2] to accommodate the diversity of molecules that a cell might need to acquire from the environment. ATP-binding cassette (ABC) transporter is one of the largest membrane transporter super-families that plays a vital role in the cellular functions and universally distributes in three kingdoms of the life. In *Escherichia coli*, the genes encoding ABC transporters occupy almost 5% of the genome with about 80 distinct transporters [3] and 4% of the genome are encoded as ABC transporters in the *Xanthomonas citri* [4]. And in human about 50 ABC transporters are present [5]. ABC transporter proteins carry various substrates in and out of membrane to provide nutrients and efflux toxics for organism. Depending on the direction of transport, ABC transporter can be classified as importers or exporters. For importer, each ABC transport has three types of subunit proteins [6]. For example, the *E. coli* maltose importer system [7] (shown in Figure 1) contains permease proteins MalF and MalG, ATP-binding protein MalK and maltose binding protein MBP. These three types of subunits form a complex, *e.g.* MBP-MalFGK2, to conduct the transport process. And for ABC exporter, substrate binding protein is absent [8]. Three type subunit proteins have their particular structures and functions. Permease protein and ATP-binding protein are dominated by subunit of transmembrane domain (TMD) and nucleotide binding domain (NBD), respectively [9]. The former offers a membrane-spanning channel for transmitting the substrates and the latter provides the energy during the transport process. For importer, additional substrate binding protein is specifically associates with the ligand in the periplasm for delivery to the appropriate permease protein [10]. Mutations of these proteins have been implicated in various diseases such as immune deficiency, cystic fibrosis and progressive familial intrahepatic cholestasis-2 [11]. Currently, the limitation of 3D structures highlights the importance of the functional annotation for ABC transporters with bioinformatics-based methods and the classification of ABC transporters based on their specific substrates remains an important challenge in structural and functional biology.

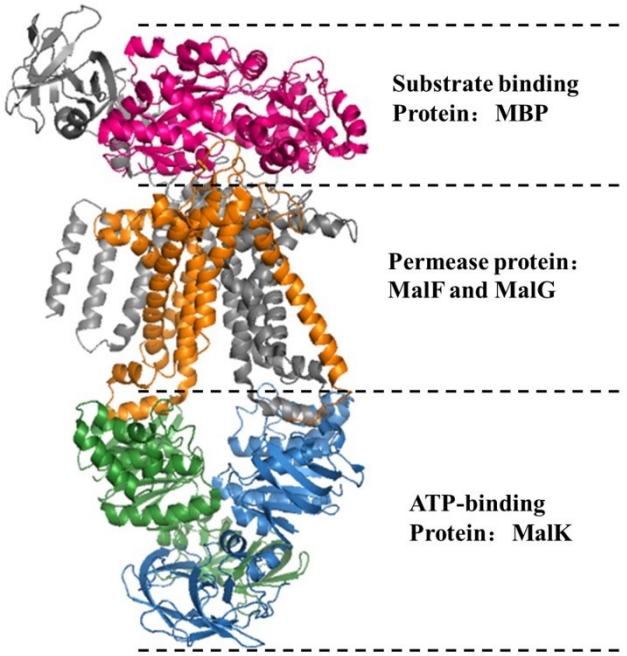

Fig. 1 Structure of E. coli maltose importer system (PDB ID: 2R6G). Stereo view of three subunits is shown in a ribbon diagram. Color code: hot pink, MBP; grey, MalF; orange, MalG; blue and green, MalK dimer..

Early studies have employed homology search and motif search to classify and predict transporters. For example, transportDB [12] is a transporter database annotated through BLAST [13] search. Hidden Markov Model [14] and PST [15] are commonly applied in motif-based methods. However these traditional similarity search-based approaches have their shortcomings, because the level of conversation within many transporter families may be very low [16]. Machine learning methods sidestep lack of conversation by using various features that can truly reflect intrinsic correlation between biological samples and the target to be predicted. Several computational methods have been proposed to identify and predict membrane transport proteins, also known as transporters. A SVM machine method has been proposed by Lin et al. [17] to discriminate five super-families and three families of transporters. Li et al. [18] integrated traditional analysis methods into a machine learning framework and proposed an automatic transporter prediction system, TransportTP with a two-phase classification approach [19]. Barghash et al. [20] have functionally annotated transporters at family level and substrate level from three organisms based on sequence similarity and sequence motifs. Different features or descriptors have been used to characterize the transporters in different machine learning models such as amino acid occurrence [21], amino acid composition [22], pair amino acid composition and pseudo amino acid composition [23]. Position-specific scoring matrices (PSSM) and biochemical properties have been used by Ou et al. [24] and Chen et al. [25] to develop network-based models for functional classification and prediction of transporters. More recently, Nitish K et al. [26] have proposed a method to achieve the goal to classify transporter into the maximum possible number of transported substrates classes by building a model based on biochemical composition and PSSM. These methods may get a good performance in overall transporters, but they have not considered family characteristic, such as ABC transporter super-family. Previous bioinformatics studies on substrate specificity of ABC transporter are focus on the prediction of substrate properties for multidrug transporters. P-Glycoprotein [27-28] and breast cancer resistance protein (BCRP) [29-31] have been widely studied on their specific substrate properties. These studies on transport candidates by multidrug transporters might helpful to drug screening and cancer therapy.

Our work aims to predict the substrate-specific classes for ABC transporters. As we know, ABC transporters have two or three subunit proteins, namely permease protein, ATP-binding protein and substrate binding protein. They have unique structures and properties, but all have significant contributions to transport process. Ashok et al. [32] focused on functional genomic re-annotation of ABC transporter in G. sulfurreducens PCA using five one-dimensional tools including BLAST, family, domain, orthologous groups and signature recognition search. To best of our knowledge, there is no functional annotation of ABC transporters from protein sequence information. Moreover, ABC transporters have their own characteristic and previous researches on functional prediction of transporters only consider any one of the three types of subunits as the same one transporter protein. In this work, through comparative analysis on sequence features including amino acid composition and biochemical properties, great differences are found between three subunit protein sequences in ABC transporter system. Moreover, we have made difference analysis among four substrate-specific classes in different datasets. We found that differences are more obvious among these four classes in three subunit protein sets than that in the dataset of all proteins, which proposed us to build subunit-specific models. First step is to subunit-specific representation of ABC transporters. Through individual feature evaluation and selection for each type of subunit, the different optimal features were obtained for the three types of subunit. Then three consensus support vector machine models respond to three types of subunit were constructed to computationally predict the substrate specificity for four major classes of ABC transporters with different transport targets. Each one is a multi-classification model that can simultaneously classify sugar, ion, amino acid/protein and other transporters respectively. At last, we combined the prediction results of the three subunit-specific models to justify the substrate specificity of ABC transporters.

Our method yields a promising result with the prediction accuracy of higher than 80% for all the four substrate classes either on the cross validation test or on the independent dataset. Moreover, it performs better than uniform model which uses same descriptors to characterize all ABC transporters, indicating that it is a more reasonable way to consider three subunit proteins separately and develop individual models for them. Despite each type of subunit is individually represented, the method is still concise because the three models can consensually predict four classes of substrate at one time.

2. Materials and methods

2.1 Dataset

We originally downloaded dataset from the IUBMB-endorsed

transporter classification database by Saier et al. [33] and collected 1507 transporter sequence IDs of ABC transporter super-family. Then we mapped these IDs to UniProt database (release 2014_01) and deleted sequences recorded in TrEMBL database or labelled as "putative", "probable" or "uncharacterized". The remaining 518 ABC transporters were divided into three sets according to three types of subunit proteins including permease proteins, ATP-binding proteins and substrate binding proteins. We also removed the sequences with sequence identity higher than 70% in each subunit set using CD-hit software [34]. So 259 permease proteins, 127 ATP-binding proteins and 87 substrate binding proteins were remained. At last, each subunit protein dataset was further divided into four classes based on the substrate specificity. In this work, four major classes of substrates were considered, including sugar, ion, amino acid/protein and others. According to the ratio of 4:1, each dataset was divided into training set and independent dataset and the detailed information about the datasets is listed in Table 1. The training sets were used to construct and valid the SVM models and the independent sets were for testing the practical performance of the model. The data can be freely accessible at http://cic.scu.edu.cn/bioinformatics/ABCtrans_pred.zip.

Table 1 The statistics of cross validation dataset and independent dataset for four transport target classes on three subunit protein sets.

Subunit protein Set / Target class	Permease Protein set	ATP-binding protein set	Substrate binding protein set
Sugar	22	14	14
Ion	43	27	24
Amino acid/protein	46	26	23
other	96	34	8
Total	207	101	69
Sugar	5	4	4
Ion	11	7	6
Amino acid/protein	12	6	6
other	24	9	2
Total	52	26	18

2.2 Feature extraction

Here, we tried to represent ABC transporters from protein primary sequences. Four sequence features, including AAC, CTD, PSSM and biochemical properties have been commonly considered to functionally characterize different transporters. So the four features were also extracted and fused as one feature vector to represent the substrate specificity of ABC transporters.

2.2.1 Amino acid composition (AAC).

The composition of amino acid has been computed to evaluate the number of occurrences of 20 amino acids normalized with total number of residues in a protein. Detailed information and calculation can be accessible at PROFEAT [38]
It is defined as:

$$composition(\mathrm{i}) = \frac{n_i}{N} \qquad (1)$$

Where i stands for one of the 20 amino acids, n_i represents the number of each type of amino acid and N is the total number of residues in the protein sequence.

2.2.2 Composition, Transition and Distribution (CTD).

Three descriptions, composition, transition and distribution have been widely used in protein functional classification [35-37]. Our implementation is based on PROFEAT [38]. The 20 amino acids are divided into three groups (labelled as 1, 2 and 3 respectively) based on seven given properties. For each property, every residue is replaced by the index "1", "2" or "3" according to one of three property groups to which the residue belongs. Then one amino acid sequence is transformed into a numerical vector to calculate CTD properties.
Composition describes the global composition of each class in the sequence. It can be defined as:

$$C(\mathrm{r}) = \frac{n_r}{N} \qquad \mathrm{r}=1, 2, 3 \qquad (2)$$

Where n_r is the number of r in the encoded sequence and N is the length of the sequence. The encoded classes "1", "2", and"3" represent three states of a special property. So composition feature contains 3 attributes for each property.
Transition represents the frequencies with which the encoded class changes along the entire length of the protein. It is defined as:

$$\mathrm{T(rs)} = \frac{n_{rs} + n_{sr}}{N-1} \qquad \mathrm{rs}=\text{"12", "13" or "23"} \qquad (3)$$

Where n_{rs} and n_{sr} is the number of dipeptides encoded as "rs" and "sr" respectively in the sequence and N is the length of the sequence. Three attributes are included for each property in transition feature.
Distribution (D) is calculated to measures the chain length where the first, 25%, 50%, 75% and 100% of the amino acids of a particular property are located. It is calculated as following:

$$\mathrm{D(ri)} = \frac{p_{ri}}{N} \qquad \mathrm{i}=0, 25\%, 50\%, 75\%, 100\%$$

$$\mathrm{r}=\text{"1", "2", "3"} \qquad (4)$$

Where p_{ri} is the positon of the first occurrence, the 1st, 2nd, and 3rd quantile and the last occurrence of r in the encoded sequence and N is the length of sequence. For each property, the distribution feature has $5\times3=15$ attributes.
The seven properties associated with CTD features are hydrophobicity, normalized Van der Waals volume, polarity, polarizibility, charge, secondary structures and solvent accessibility. Thus, CDT feature set gives a fixed length of $(3+3+15)\times7=147$ attributes.

2.2.3 Position-specific scoring matrices (PSSM) profile.

PSSM profile is an attribute that is popularly used to reflect the evolutionary information in protein sequence functional classification [24-26] and other bioinformatics studies [39-41]. In this work, we generated the PSSM profile based on running the position-specific iterative basic local alignment search tool (PSI-

BLAST) against Swiss-Prot database. The original PSSM profile from PSI-BLAST was further used to calculate the PSSM feature by summing rows in PSSM profile that correspond to the same amino acid residues and then is scaled to [-1,1]. Finally, we got 400 feature elements for each protein sequence.

2.2.4 Biochemical properties.

Biochemical properties of amino acids are also valuable attributes and useful to analyse. For example, biochemical properties (AAindex) have been successfully used for earlier studies on protein folding and stability [42-44] and functional classification [24-25]. In our work, a set of 49 physical, chemical, energetic and conformational properties of amino acids was collected. The normalized values of these 49 properties were downloaded from http://www.cbrc.jp/~gromiha/fold_rate/property.html.

We used the following formula to compute each property value:

$$AAindex_i = \frac{\sum_{j=1}^{n} AAindex_{ij}}{n} \quad (5)$$

Where $AAindex_i$ is the value for ith biochemical property in a given sequence; $\sum_{j=1}^{n} AAindex_{ij}$ is the arithmetic sum of the ith biochemical property, and n is the length of the protein sequence.

2.3 Support vector machine (SVM)

SVM provided by Vapnik [45] et al. is one of the most popular learning-machine algorithms to solve classification problems. Here, we used the commonly used software LIBSVM to construct the prediction model by downloading this tool from http://www.csie.ntu.edu.tw/~cjlin/libsvm. LIBSVM is proposed by Lin et al. and is integrated software for support vector classification, regression and distribution estimation [46]. Besides, it can solve multi-class classification problems efficiently based on one-against-one strategy. For an n classification problem, it constructs $n \times (n-1)/2$ binary classifiers and fuses them into one multi-class classifier model using a voting strategy. Each binary classification is considered to be a voting where votes can be cast for all data points and the end point is designated to be in a class with maximum number of votes. In case that two classes have identical votes, it simply selects the one with the smallest index. In this work, we selected radial basis function as kernel function to develop multi-class classification model and found best corresponding parameters C and γ by grid search.

In our work, each subunit-specific set contains four classes of substrates. We used LIBSVM to construct one multi-class classifier model to simultaneously discriminate the four classes, so only three subunit-specific SVM models are needed to be constructed.

2.4 Model validation and evaluation

In statistical prediction, sub-sampling test is a good validation method to verify the stability of the model and estimate accuracy. Here, 5-fold cross-validation was used to investigate the training set. In the 5-fold cross validation, the training set is randomly partitioned into 5 equal size subsets. In every case, each of the 5 subsets is used as testing set and the remaining are used for training. The process of model training and testing is repeated 5 times and the performance of each model is computed as the average of the five runs.

Four parameters including sensitivity (SE), specificity (SP), accuracy (ACC) and Mathew's correlation coefficient (MCC) were used to evaluate the prediction ability of our model. They are defined as following:

$$SE = \frac{TP}{TP + FN} \quad (6)$$

$$SP = \frac{TN}{TN + FP} \quad (7)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$MCC = \frac{(TP + TN) - (FP + FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (9)$$

Where TP, FP, TN and FN are true positive, false positive, true negative and false negative, respectively.

3. Results and Discussion

3.1 Compositional biases among three subunit proteins

We firstly analysed the compositions of 20 standard amino acids in three subunit proteins of ABC transporters. Figure 2(a) indicates that several amino acids, such as Ile, His, Asp and Asn express significant difference among the three subunit proteins. The variances of 20 AACs among three subunit proteins are shown in Figure 2(b) and we can see that the mean variance has reached to 1.2. The hydrophobic amino acid, Ile presents the largest variance in permease protein, ATP-binding protein and substrate binding protein.

Moreover, Figure 2(b) shows that eight AACs give big differences with the variances lager than the mean value of 1.2. Except Ile, the other seven are His, Asp, Asn, Gly, Pro, Cys and Ala respectively. Most of these residues are hydrophobic or neutral, which may be relevant to the different biological environment and functions of the three subunit proteins.

For the biochemical compositions, we also calculated the distribution of amino acid properties in the three subunit proteins and further compared these properties among them. Here six physicochemical and structural properties were selected, including positive charge (Lys, Arg), hydrophobicity (Cys, Leu, Val, Ile, Met, Phe, Trp), van der Waals volume (Gly, Ala, Ser, Thr, Pro, Asp), polarity (Pro, Ala, Thr, Gly, Ser), polarizability (Gly, Ala, Ser, Asp, Thr) and secondary structure-heilx (Glu, Ala, Leu, Met, Gln, Lys, Arg, His). We divided each property composition into four equal parts in protein sequence and computed the share of sequence length for each part. As shown in Figure 3, obvious distribution differences among three subunit proteins are found, especially for properties of positive charge (Figure 3(a)), hydrophobicity (Figure 3(b)) and secondary structure (Figure 3(c)). Moreover, for each property, the composition in both ends of protein sequence show more diversity in three subunit proteins. It suggests that two ends of the subunit protein sequences are less conservative than the middle segments.

From the above comparative analysis, we can further confirm that the three subunit proteins of ABC transporters have unique

sequential properties. So it is a more reasonable way to individually represent them during the model construction, rather than describe them with the same feature vector.

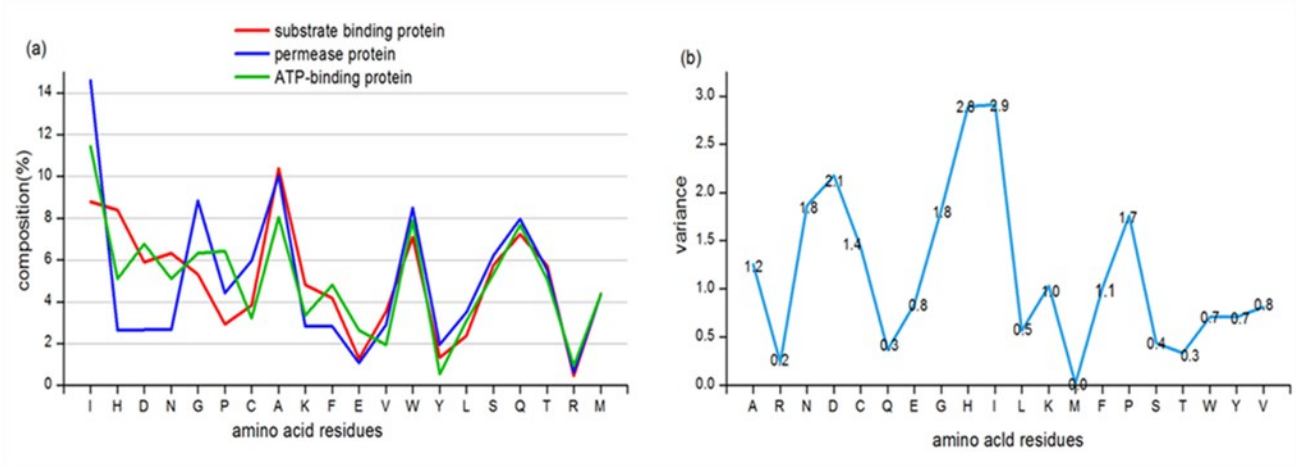

Fig. 2 (a) Amino acid compositions of three subunit proteins. The compositions of 20 amino acids of premease protein (blue), ATP-binding protein (green) and substrate binding protein (red) are plotted. (b) Variance of amino acid composition among three subunit proteins. The line chart shows variances of 20 amino acid compositions among permease protein, ATP-binding protein and substrate binding protein and corresponding values are marked on the curve.

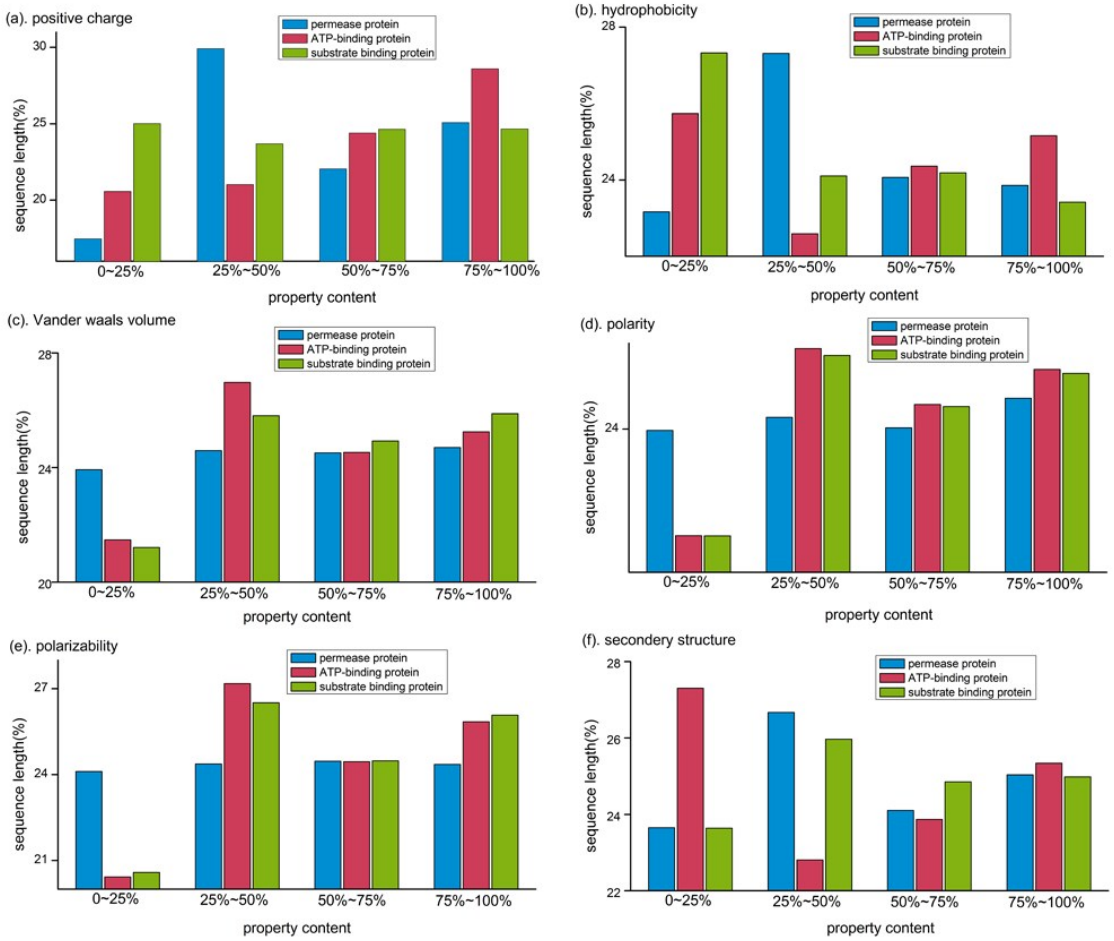

Fig. 3 Distribution of amino acid properties among three subunit proteins. The shared sequence lengths of four equal property content parts for different properties in permease protein, ATP-binding protein and substrate binding protein are shown with column of different colours. Blue, red and green represents for permease protein, ATP-binding protein and substrate binding protein, respectively. Here we analysed six properties: (a) positive charge, (b) hydrophobicity, (c) van der Waals volume, (d) polarity, (e) polarizability and (f) secondary structure.

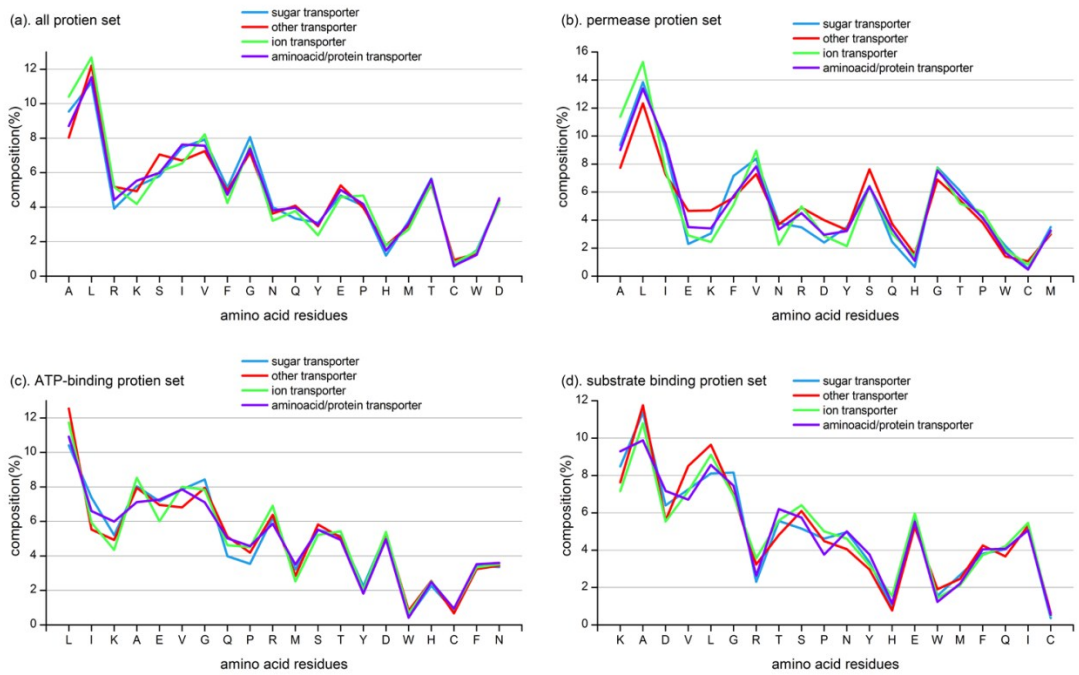

Fig. 4 Amino acid compositions of four transport target classes. Blue, green, purple and red curve shows 20 amino acid compositions of sugar transporter, ion transporter, amino acid/protein transporter and other transporter, respectively. Each sub-graph reflects differences among four target classes on each protein set: (a). all protein set, (b). permease protein set, (c). ATP-binding protein set and (d). substrate binding protein set

3.2 Property comparisons among different transport targets

Based on the comparative analysis among three subunit proteins, we further computed the amino acid composition of ABC transporters with different transport targets on the three subunit protein sets and all protein set (put all ABC transporters together regardless of subunit proteins). The amino acid comparisons among the four classes of transport targets (substrates) in every one of the four sets are shown in Figure 4. Almost no difference can be seen among sugar transporter, ion transporter, amino acid/protein transporter and other transporter on all protein set, but more obvious differences are found among the three subunit protein sets, especially for permease protein set and substrate binding protein set. Since ATP-binding proteins have strict conservative nucleotide binding domain [47] which makes them less diverse than other two subunit proteins, and it is also the reason why we called this super-family as ABC (ATP-binding cassette) transporter. In addition, amino acid residues that have much variation across four classes on each set are different (shown in Figure 4). For the all protein set, six most diverse amino acids among sugar transporter, ion transporter, amino acid/protein transporter and other transporter are Ala, Leu, Arg, lys, Ser and Ile. There are six most diverse amino acids of Ala, Leu, Ile, Glu, Lys and Phe in permease protein set, of Leu, Ile, Lys, Ala, Glu and Val in ATP-binding protein set and of Lys, Ala, Asp, Val, Leu and Gly in substrate binding protein set, respectively. These diverse amino acids for each set have different characteristic and functions, and it suggests that individual consideration for each subunit protein is needed when we construct the prediction model for ABC transporters.

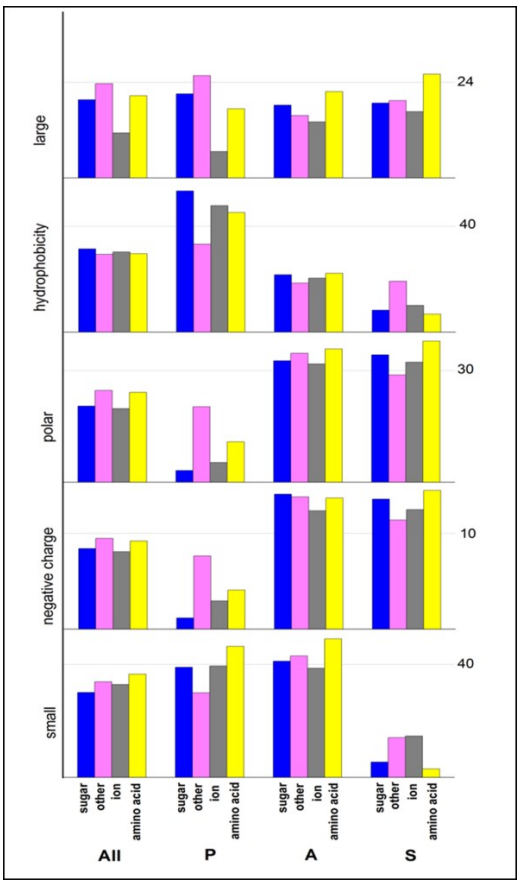

Fig. 5 Biochemical properties of four transport target classes. This histogram elaborates biochemical property biases of sugar transporter

(blue), other transporter (pink), ion transporter (grey) and amino acid/protein transporter (yellow) on four protein sets. Properties analyzed on the figure are large van der Waals volume, hydrophobicity, polar, negative charge and small van der Waals volume. Here All, P, A and S stands for all protein, permease protein, ATP-binding protein and substrate binding protein, respectively; small and large are short names of large van der Waals volume and small van der Waals volume.

In addition, biochemical properties of large van der Waals volume (Met, His, Lys, Phe, Arg, Tyr, Trp), hydrophobicity (Cys, Leu, Val, Ile, Met, Phe, Trp), polar (Arg, Lys, Glu, Asp, Gln, Asn), negative charge (Asp, Glu) and small van der Waals volume (Asn, Val, Glu, Gln, Ile ,Leu) are also popularly used for functional annotations of proteins.

So we used them to analyse the property biases among different transport target of ABC transporters. As Figure 5 shows, sugar transporter, ion transporter, amino acid/protein transporter and other transporter have different characteristic properties. On all protein set, composition of large van der Waals volume shows obvious difference across four transporter classes, but for other properties, variances is not significant. While, property composition among four transporter classes on the three subunit protein sets show more differences and variances are also larger than that on all protein set. This disparity can be easily seen on permease protein set and substrate binding protein set.

So the comparative analysis among different transporter targets further suggests that dividing ABC transporters into three groups according to the subunit type is good for functional classification about transport target because it gives prominence to the differences between sugar transporter, ion transporter, amino acid/protein transporter and other transporter and to their own characteristic on each subunit protein set.

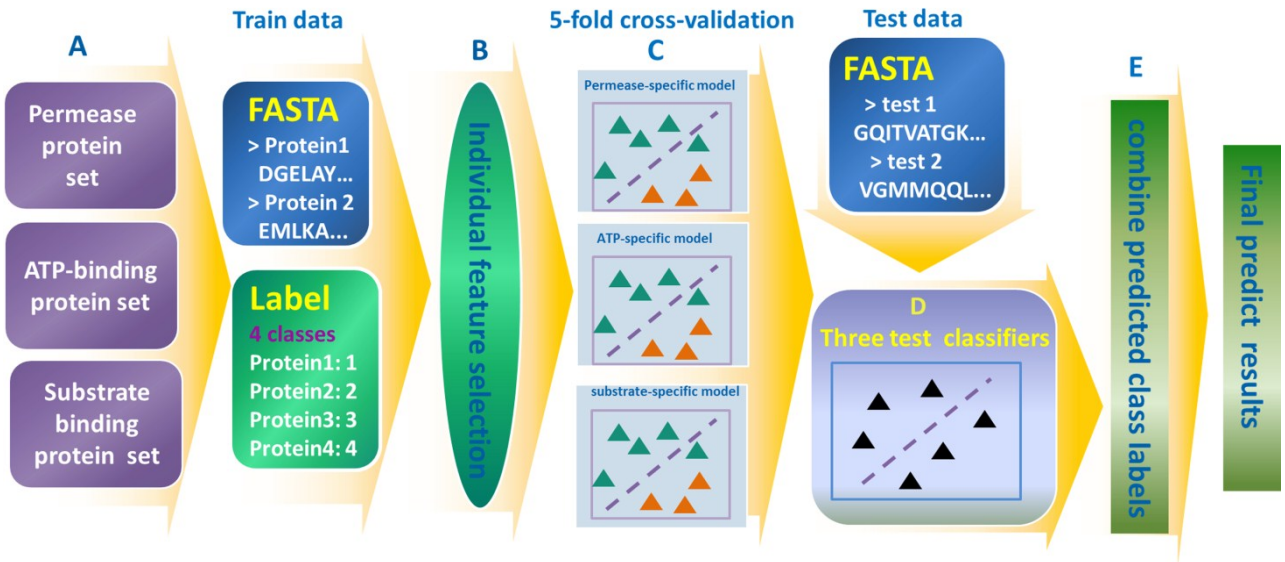

Fig. 6 Schematic overview of the main steps in our model design.

3.3 The optimization of models on different feature sets

We used AAC, CTD, PSSM, AAindex and their different combinations to build a series of models to discriminate sugar, ion, amino acid/protein and other ABC transporters for each subunit protein set. The entire flowchart of our work can be seen in Figure 6. Performances of different models on 15 feature sets using the training set and 5-cross validation test for three subunit protein sets can be seen in Supplementary Table S1, S2 and S3.

For the permease-specific model construction, we can see that PSSM performs better than AAC, CTD and AAindex, and the model based on PSSM feature set is also the best one in all 15 models with an average accuracy of 89.59% and a MCC of 0.7078. Since the model based on the feature set of AAindex and PSSM gives a nearly result to the best model with an accuracy of 89.40%, but sensitivity of it (74.95%) is lower than that of the best model (76.38%). So for optimal permease-specific model was selected based on PSSM profile.

For ATP-binding protein dataset, the performance of the models on different feature sets is relatively poor, which may be caused by the conservation of NBD in ATP-binding protein, as mentioned above. However, we still easily find the optimal model that is the one based on the feature set of AAindex combined with PSSM and it achieves an average accuracy of 80.36% with a MCC of 0.5050.

Similarly, PSSM feature set also performs best for substrate binding protein set with an accuracy of 86.12% and a MCC of 0.6784. So we can conclude that PSSM is a good feature in functional classification and bioinformatics annotation for ABC transporter as the previous reports have confirmed [25-26].

Finally, we summed up the results of the three optimal subunit-specific models to finally annotate substrate-specific function of ABC transporters. Our method that integrates the best models of three subunit protein sets achieved an accuracy of 93.35%, 84.34%, 87.24% and 81.96%, and a MCC value of 0.7411, 0.5709, 0.6658 and 0.6324 for sugar transporter, ion transporter, amino acid/protein transporter and others respectively, using the training set and 5-fold cross validation test (shown in Table 2).

Table 2. The prediction results of the four substrate classes based on the combination of the three optimal subunit-specific models on cross validation set and independent test set.

Transporter class	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
Cross validation dataset				
Sugar	75.20	96.56	93.35	0.7411
Ion	65.21	91.23	84.34	0.5709
Amino acid/protein	76.90	90.90	87.24	0.6658
Others	79.77	84.56	81.96	0.6324
Independent test set				
Sugar	69.23	96.39	92.71	0.6789
Ion	82.86	88.52	83.33	0.7097
Amino acid/protein	79.17	84.72	89.58	0.5968
Others	66.67	97.22	86.46	0.7088

Table 3 The performance of uniform model with best feature set on cross validation set and independent test set for different substrate-target ABC transporter classes

Transporter class	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
Cross validation dataset				
Sugar	42.59	100.0	92.03	0.5961
Ion	79.61	78.17	78.22	0.5283
Amino acid/protein	49.00	96.00	84.36	0.5405
Others	73.09	73.96	73.74	0.4572
Independent test set				
Sugar	38.46	98.80	90.63	0.5266
Ion	75.00	81.94	80.21	0.5273
Amino acid/protein	58.33	97.22	87.50	0.6455
Others	82.86	77.05	79.17	0.5798

3.4 Results on the independent test dataset

We evaluated the practical performance of our models with an independent dataset of 96 ABC transporter sequences, including 13 sugar transporters, 24 ion transporters, 24 amino acid/protein transporters and 35 other transporters. The result proves that our method by considering three subunit proteins individually has yielded a good performance with an accuracy of 92.71%, 83.33%, 89.58% and 86.46% for sugar transporter, ion transporter, amino acid/protein transporter and other transporter respectively on the independent dataset. Detailed results can be seen in Table 2. Three subunit models and best features for each set are also http://cic.scu.edu.cn/bioinformatics/ABCtrans_pred.zip and results can be easily got with them using LIBSVM in MATLAB.

3.5 Comparison with uniform model on the all protein set and multiple sequence alignment.

To further evaluate the superiority of our individually considered models, we used the above 15 different feature sets to develop uniform models on the all protein set. Result of each model on the training set of all proteins with 5-fold cross validation test is shown in supplementary (Table S4) and the best model is the one based on PSSM+AAindex+AAC feature set. Using the best uniform model, the detailed performance of the cross validation set and independent test set for different substrate-target ABC transporter classes is shown in Table 3. So we can make a comparison between the best uniform model on all proteins and the combined model by three individual subunit-specific models (shown in Figure 7). It is obvious that individually considered models perform much better than uniform model in a whole especially for other transporter class. Figure 7 demonstrates that both the accuracy and MCC values of the combined model are higher than those of the uniform model for four substrate-specific classes of ABC transporter.

By comparison between the results listed in Table 2 and Table 3, although individually considered model show a slightly lower sensitivity than uniform model for ion transporter, it has improved sensitivity for sugar transporter, amino acid/protein transporter and other transporter by 32.61%, 27.9% and 6.68% on cross validation test, respectively. And same results can be seen in independent test set. Individually considered model has achieved higher accuracy for all four classes and much higher

sensitivity can be seen especially for sugar, ion and amino acid/protein transporters. By comparing the individually considered model with the uniform model, we can further believe that individual representation and model construction for each type of subunit proteins is a more reasonable way in predicting the substrate specificity of ABC transporters.

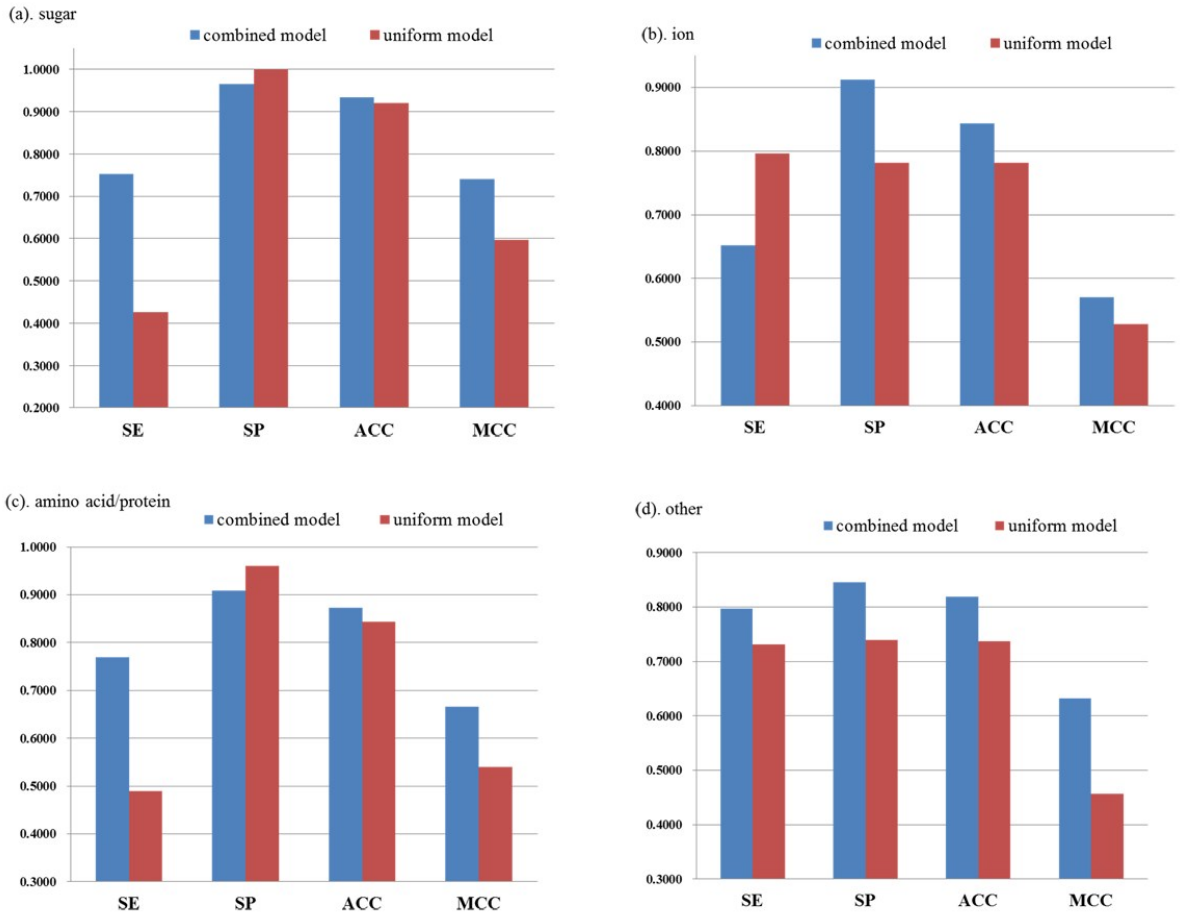

Fig.7 Comparison of performances between combined model and uniform model for four transport target classes on cross validation dataset. Sensitivity, specificity, accuracy and Mathew's correlation coefficient are used to compare the performance of combined model and uniform model for sugar transporter (a), ion transporter (b), amino acid/protein transporter (c) and other transporter (d). SE, SP, ACC and MCC stand for sensitivity, specificity, accuracy and Mathew's correlation coefficient, respectively.

In addition, multi-sequence alignment methods are simple and effective for classification research. These methods have shown good performance [20], while they may obtain failed results when the query protein has very low identity within other known members. ABC transporters have some family resemblance especially for conserved NBD. Therefore it is probable to predict substrate-specificity of transporter using multi-sequence alignment methods such as BLAST. To make a comparison between our method and the multi-sequence alignment methods on ABC transporter system, we used BLAST to predict our test dataset.

Considering alignment within the family and some ABC transporters have conserved NBD as mentioned above, we ran BLAST with E-value of 1e-20 to match transporters in test data set and 76 of 96 transporters are matched. It got an average accuracy of 75% (57/76) which is lower than ours (80%) and more than 20% transporters cannot be matched. The results are list in Supplementary Table S5. When the similarity threshold was set higher permissively, more matching numbers were got but false predictions also increased. And likewise, with higher E-value, the accuracy increased with more unclassified members. Previous works [20] have got the same conclusions. In all, our method has its own advantages and shows better performance compared to simple multiple sequence alignment.

3.6 component and biological function of other transporters class

Substrates transported by the other class transporters have a wide variety such as lipids, drugs, haemins and other biological macromolecules which are very important for life. While numbers of transporters for each substrate are so limited that we cannot build the valid statistical models and make predictions individually for them. Here we give an analysis of the component and biological function of other transporter class particularly. Lipids are fundamental to life and probably, apart from amino acids and possibly ribose [48]. Our present dataset includes 5 ABC lipid transporters which are all divided into training set randomly during model building. We make statistics of prediction results for these 5 transporters and find that our model can correctly predict 4 of them. Another important component is drug

transporter, most of they are associated with distribution and concentration of drug in the organism especially for the multidrug resistance transporters that can reduce effective drug concentration during the cancer therapy. Works on predicting substrates of these multidrug transporters can be seen in [27-31]. Our present dataset contains 12 multidrug transporters, they all have been correctly predicted in cross validation test. It suggests that our descriptions can effectively distinguish multidrug transporters from sugar, ion and amino acid /protein transporters. So we can expect that if more members of one specific class will be discovered in the future, our method would be used specially for these new transporters and be tested in a wider range.

4. Conclusions

In this work, our study focused on the functional classification of ABC transporter family based on substrate-specificity. We take account of family properties for ABC transporters rather than considering them just as common transporters like previous studies. Since ABC transporters have two or three subunit proteins with unique structures and properties, we analysed the differences among three type subunit proteins of ABC transporters and divided all ABC transporters into three protein sets responding to three subunit proteins. And individual representation and model construction was implemented for each type of subunit proteins. Then we developed three multi-classification models for sugar transporter, ion transporter, amino acid/protein transporter and other transporter on each subunit protein set. Finally, we summed up all results of three subunit-specific models to make systematic functional classification of ABC transporters. This procedure yields three multi-class classification models which are assigned according to the greatest preferences for three subunit protein sets (Figure 6). Our method has achieved accuracy in range of 81% to 93% on training set of 377 ABC transporters. Compared with uniform models on all proteins, it has improved sensitivity for sugar transporter, amino acid/protein transporter and other transporter by 32.61%, 27.9% and 6.68% on training set, respectively. Our method also performed well on the independent test dataset with the accuracy of 92.71%, 83.33%, 89.58% and 86.46% for sugar transporter, ion transporter, amino acid/protein transporter and other transporter respectively. These results suggest that considering three subunit proteins separately highlights the difference between ABC transporter classes and developing individual models for three substrate protein groups is a recommendable and effective way for functional annotation of ABC transporters.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (No. 21175095, 21273154, 21375090).

Notes

College of Chemistry, Sichuan University, Chengdu 610064, People's Republic of China. Tel.: +86-028-85412290; Fax: +86-028-85412290; E-mail address: yzguo@scu.edu.cn; xmpuscu@scu.edu.cn

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

References

1. W. Busch and M. H. Saier, Crit. Rev. *Biochem. Mol. Biol*., 2002, **37**, 287-337.
2. T. J. Lee, I. Paulsen and P. Karp, *Bioinformatics*, 2008, **24**: 259-267.
3. K. J. Linton, C. F. Higgins, *Mol. Microbiol.*, 1998, **28**, 5-13.
4. F. J. Medrano, C. S. de Souza and A. Romero, *Acta Crystallogr F.*, 2014, **70**, 564-571.
5. M. Dean, Y. Hamon and G. Chimini, *J. Lipid Res.*, 2001, **42**, 1007-1017.
6. K. Tomii and M. Kanehisa, *Genome Res.*, 1998, **8**, 1048-1059.
7. M. L. Oldham, D. Khare and F. A. Quiocho, *Nature*, 2007, **450**, 515-521.
8. R. J. Dawson, K. Hollenstein and K. P. Locher, *Mol. Microbiol.*, 2007, **65**, 250-257.
9. D. C. Rees, E. Johnson and O. Lewinson, *Nat. Rev. Mol. Cell Bio.*, 2009, **10**, 218-227.
10. G. F. Ames, Annu. Rev. *Biochem.*, 1986, **55**, 397-425.
11. F. Klepsch, P.Vasanthanathan and G. F. Ecker, *J. Chem. Inf. Model*, 2014, **54**, 218-229.
12. Q. Ren, K. H. Kang and I. T. Paulsen, *Nucleic Acids Res.*, 2004, **32**, D284-D288.
13. S. F. Altschul, W. Gish and W. Miller, *J. Mol. Biol.*, 1990, **215**, 403-410.
14. A. Krogh, M. Brown and I. S. Mian, *J. Mol. Biol.*, 1994, **235**, 1501-1531.
15. E. Eskin, W. S. Noble and Y. Singer, *J. Comput. Biol.*, 2003, **10**, 187-213.
16. B. Heil, J. Ludwig and H. Lichtenberg-Fraté, *Bioinformatics*, 2006, **22**, 1562-1568.
17. H. H. Lin, L. Y. Han and C. Z. Cai, *Proteins*, 2006, **62**, 218-231.
18. H. Li, X. Dai and X. Zhao, *Bioinformatics*, 2008, **24**, 1129-1136.
19. H. Li, V. A. Benedito and M. K. Udvardi, *BMC bioinformatics*, 2009, **10**, 418.
20. A. Barghash and V. Helms, *BMC bioinformatics*, 2013, **14**, 343.
21. M. M. Gromiha and Y. Yabuki, *BMC bioinformatics*, 2008, **9**, 135.
22. N. S. Schaadt and V. Helms, *Biopolymers*, 2012, **97**, 558-567.
23. N. S. Schaadt, J. Christoph and V. Helms, *J. Chem. Inf. Model*, 2010, **50**, 1899-1905.
24. Y. Y. Ou, S. A. Chen and M. M. Gromiha, *Proteins*, 2010, **78**, 1789-1797.
25. S. A. Chen, Y. Y. Ou and T. Y. Lee, *Bioinformatics*, 2011, **27**, 2062-2067.
26. N. K. Mishra, J. Chang and P. X. Zhao, *PloS one*, 2014, **9**, e100278.
27. L. Zhong, C. Y. Ma and H. Zhang, *Comput. Biol. Med.*, 2011, **41**, 1006-1013.
28. E. Hazai, I, Hazai and I. Ragueneau-Majlessi, *BMC bioinformatics*, 2013, **14**, 130.
29. Z. Wang, Y. Chen and H. Liang, *J. Chem. Inf. Model*, 2011, **51**, 1447-1456.
30. J. Huang, G. Ma and I. Muhammad, *J. Chem. Inf. Model*, 2007, **47**, 1638-1647.
31. Z. Bikadi, I. Hazai and D. Malik, *PLoS One*, 2011, **6**, e25815.
32. A. Selvaraj, V. Sumantran and N. Chowdhary, *Curr. Bioinform.*, 2014, **9**, 166-172.
33. M. H. Saier, V. S. Reddy and D. G. Tamang, *Nucleic Acids Res.*, 2014, **42**, D251-D258.
34. W. Li and A. Godzik, *Bioinformatics*, 2006, **22**, 1658-1659.
35. C. Z. Cai, L. Y. Han, Z. L. Ji, *Nucleic Acids Res.*, 2003, **31**, 3692-3697.
36. H. H. Lin, L. Y. Han and C. Z. Cai, *Proteins*, 2006, **62**, 218-231.
37. B. A. van den Berg, M. J. Reinders and J. A. Roubos, *BMC bioinformatics*, 2014, **15**, 93.
38. H. B. Rao, F. Zhu and G. B. Yang, *Nucleic Acids Res.*, 2011, **39**, W385-W390.
39. S. Ding, S. Yan and S. Qi, *J. Theor. Biol.*, 2014, **353**, 19-23.
40. L. Zou, C. Nan and F. Hu, *Bioinformatics*, 2013, **29**, 3135-3142.
41. C. Fang, T. Noguchi and D. Tominaga, *BMC bioinformatics*, 2013, **14**, 300.
42. M. M. Gromiha and S. Selvaraj, *Biophys. Chem.*, 1999, **77**, 49-68.

43. M. M. Gromiha, A. M. Thangakani and S. Selvaraj, *Nucleic Acids Res.*, 2006, **34**, W70-W74.
44. M. M. Gromiha, S. Selvaraj and A. M. Thangakani, *J. Chem. Inf. Model*, 2006, **46**, 1503-1508.
45. V. Vapnik, Statistical learning theory, 1998, *Wiley*, New York.
46. C. C. Chang and C. J. Lin, *ACM. T. Intel. Syst. Tec.*, 2011, **2**, 27.
47. K. Hollenstein, D. C. Frei and K. P. Locher, *Nature*, 2007, **446**, 213-216.
48. R. P. Bywater, K. Conde-Frieboes, Astrobiology, 2005, **5**, 568-574.