BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 9 2008, pages 1129–1136
doi:10.1093/bioinformatics/btn099

Sequence analysis

A nearest neighbor approach for automated transporter prediction and categorization from protein sequences

Haiquan Li, Xinbin Dai and Xuechun Zhao*

Bioinformatics Lab, Plant Biology Division, The Samuel Roberts Noble Foundation, Inc., 2510 Sam Noble Parkway, Ardmore, OK 73401, USA

Received on November 21, 2007; revised on March 10, 2008; accepted on March 11, 2008

Advance Access publication March 12, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Membrane transport proteins play a crucial role in the import and export of ions, small molecules or macromolecules across biological membranes. Currently, there are a limited number of published computational tools which enable the systematic discovery and categorization of transporters prior to costly experimental validation. To approach this problem, we utilized a nearest neighbor method which seamlessly integrates homologous search and topological analysis into a machine-learning framework.

Results: Our approach satisfactorily distinguished 484 transporter families in the Transporter Classification Database, a curated and representative database for transporters. A five-fold cross-validation on the database achieved a positive classification rate of 72.3% on average. Furthermore, this method successfully detected transporters in seven model and four non-model organisms, ranging from archaean to mammalian species. A preliminary literature-based validation has cross-validated 65.8% of our predictions on the 11 organisms, including 55.9% of our predictions overlapping with 83.6% of the predicted transporters in TransportDB.

Availability and Supplementary information: http://bioinfo.noble.org/manuscript-support/transporter/

Contact: pzhao@noble.org

1 INTRODUCTION

Membrane transport proteins, or simply transporters, support basic biological processes in living cells by moving essential nutrients and metabolites, such as ions, small molecules and macromolecules across biological membranes. As a consequence of transport protein activity, cells are able to maintain physiological concentrations of ions for essential physicochemical potential of cells, import and export signaling molecules to mediate intercellular communications, and prevent the accumulation of toxins since transporters function as toxin pumps (Yan, 2003).

Various biomedical, biological and biophysical techniques have been developed to screen transporters and to determine their transport mechanisms (Yan, 2003). For instance, membrane proteins are inserted into lipid bilayer membranes by reconstitution methods and the resulting liposomes are capable of recapitulating the transporters' function for further analysis (Rigaud *et al*., 1995). The patch clamp techniques have been applied to identify transported substrates and to study the transport mechanisms of ionic channels in excitable membranes (Sakmann and Neher, 1984). However, current experimental techniques are often inefficient in labor and cost, and require sophisticated skills. Therefore, computational methods are desired to select candidates in high confidence for experimental study to maximize the outcome of benchwork.

Computational methods, especially machine-learning methods, require a large set of curated data for training. The IUBMB-endorsed transporter classification database (TCDB), which employs a hierarchical, functional and phylogenetic classification system, is suitable for this purpose (Busch and Saier, 2002; Saier, 2000; Saier *et al*., 2006). The transporters in the database are hierarchically categorized into classes, subclasses, superfamilies, families and subfamilies. In particular, the classification of families is based on their phylogeny (Chang *et al*., 2004), hydropathy (Zhai and Saier, 2001), substrate specificity (Paulsen *et al*., 1998) and transmembrane topology. The strength of this system lies in the fact that the homologous transporters in a common family share the similar transport functions. Therefore, if a novel protein is classified into a family, its transport mechanism or pathway can be postulated. Other systems such as Pfam (Sonnhammer *et al*., 1997) are not suitable to use homology to make such an inference.

Sequence homology search, motif search and machine-learning techniques have been employed to predict membrane transporters from their primary amino acid sequences based on the catalogued information about known transporter proteins. Using the sequence homology approach, unknown proteins are characterized as putative transporters if their sequences are homologous to previously identified transporter sequences. BLAST (Altschul *et al*., 1990) has been widely applied in the homology search. For example, TransportDB is a putative transporter database annotated through the assistance of BLAST search for hundreds of completely sequenced organisms (Ren *et al*., 2004, 2007). Although BLAST-based methods may accurately identify many real targets, these methods also incorrectly identify numerous false positives because homologous sequences, especially paralogs, may evolve in quite

*To whom correspondence should be addressed.

distinct transport functions (Doolittle, 1981). Therefore, extensive human annotations are still required after the BLAST search.

In contrast to sequence homology approach which searches individual sequences in the transporter database, the motif-based approach relies on the use of motifs or profiles using traditional family modeling methods such as Hidden Markov Model (HMM) (Krogh *et al*., 1994), PST (Bejerano and Yonam, 2001; Eskin *et al*., 2003) and PROTOMAT (Henikoff and Henikoff, 1994) in order to characterize transporter families in the database. Motif-based methods often require a minimal number of transporters for modeling and may suffer from the low levels of conservation amongst transporter families such as potassium channels (Heil *et al*., 2006) and Na^+/H^+ exchangers (Dibrov and Fliegel, 1998). Moreover, although Pfam (Sonnhammer *et al*., 1997) and TIGRFAMS (Haft *et al*., 2001) contain some motifs about certain transporter families/superfamilies, there are a limited number of dedicated and comprehensive database of transporter motifs. As a result, tools such as INTERPROSCAN (Zdobnov and Apweiler, 2001) often miss a large percentage of putative targets, resulting in very low search efficiency.

The machine-learning approach is quite dissimilar to each of the two previously discussed methodologies. This method relies on predictions made from the rules that have been learned from curated data. A support vector machine (SVM) method was reported by Lin *et al*. (2006) with 60–97.1% accuracy on five transport superfamilies and three families. Although the SVM method can successfully identify putative transporters in some transporter families, like many other machine-learning methods, this technology requires a large number of trained transporters, which is impractical for most existing transporter families.

In addition to the three major approaches discussed, a number of other methods also exist including some which are based simply on the numbers of transmembrane segments (TMS) (Schwacke *et al*., 2003). These methods often lead to very high false positive rates since the transmembrane segments taken into consideration do not generally indicate transport functions, thus making the TMS-based methods alone ineffective in predicting transporter proteins. For instance, more than 40% of proteins in the model plant organism, *Arabidopsis thaliana*, have putative TMS, but <20% of these proteins are transporters (Schwacke *et al*., 2003).

In summary, although TCDB and TransportDB and their associated tools, such as SSEARCH (Saier *et al*., 2006), have been widely used to predict and classify putative transporters, their prediction performance has been seldom reported. In this article, we propose a simple Nearest Neighbor (NN) approach which seamlessly integrates homology searches, motif searches and topological analysis into a machine-learning framework in order to cover as many transporter families as possible. We integrate homology and motif search methods by combining their scores in the similarity measurement of the NN approach. To circumvent the absence of credible and comprehensive non-transporters in the transporter training datasets, we further integrate transmembrane segment information in preprocessing in order to filter out unlikely transporters. We demonstrated the effectiveness of the integration using both five-fold cross-validation and literature-based validation on seven model organisms and four non-model organisms including archacean, bacterial, plant, insect and mammalian species.

2 METHODS

2.1 Training and prediction of transporters

Our training data was downloaded from the TCDB website (Saier *et al*., 2006) in December 2007 and contained 4155 transporters within 740 TC families. We excluded the 256 families that contained a single family member and included the remaining 3899 transporters and 484 TC families in our study. Our approach utilized NN classification of the TCDB, as shown in the Transporter Prediction module (Fig. 1). Using the simplest NN classification, the unknown protein is assigned the family of a previously described transporter in the TCDB which is most similar to the unknown protein (Cover and Hart, 1967), based on global or local scores, such as BLAST *e*-values. The use of this methodology, however, may result in inaccurate predictions since the unknown protein can only be assigned to the family of the most similar protein if it is present in the database. Therefore, the possibility exists that without checking the universal protein databases, such as NCBI RefSeq (Pruitt *et al*., 2005) and SWISS-PROT (Apweiler, 2001), a putative transport protein will be assigned to the wrong family of transporters. However, in most cases, the protein most similar to the unknown protein in the universal database is not annotated and therefore does not provide sufficient insight into the functional annotation of the unknown protein. Moreover, in utilizing the NN classification, the poential exists that the true family of the unknown protein may not be with the NN, but will most likely be within the k-NN (k is a small constant number), since the homologous sequences may have evolved to perform different transport functions. Therefore, K-NN methods thus may be applied in this case to improve the accuracy with a weighting strategy to identify the most likely function-identical neighbors in the protein database (Horton and Nakai, 1997). Our strategy did not depend on analysis of the universal protein databases and did not use KNN directly, but rather followed its weighting strategy. The sequence similarity between an unknown protein and every transporter protein in the training database TCDB was weighted by comparing the unknown protein to the family of the compared classified transporter. Consequently, the unknown protein was classified into a TC family that contained both a member that was highly similar to the unknown protein in sequence

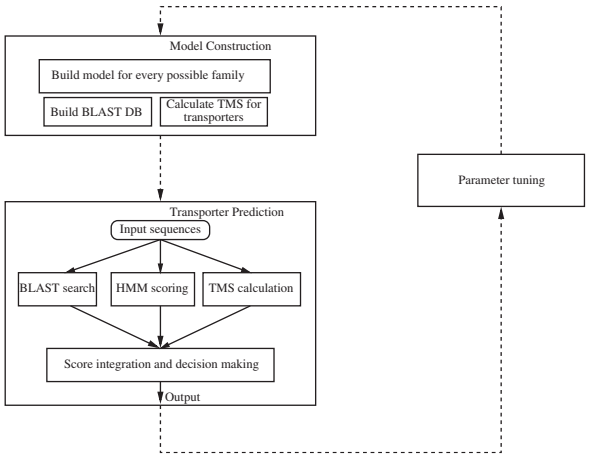

Fig. 1. The flow of our transporter modeling and prediction.

as well as members that contained characteristics/motifs matching the unknown protein.

In our approach, the extent to which an unknown protein was matched to a family of transporters was measured by a HMM program known as the Sequence Alignment and Modeling System (SAM) (Krogh *et al.*, 1994). This methodology was utilized since HMM models are able to efficiently capture the conserved features of a transport family and, more importantly, the SAM program does not require pre-alignments of transporter sequences in the family. The option of pre-alignments in SAM facilitates the weighting process since many transporter families do not possess well-accepted alignments.

In the weighting process, two practical issues were considered. First, a stratgy had to be developed by which to overcome the problem of weighting every transporter in TCDB. For some transporters, it is impossible to apply the weighting due to 1) The families of the transporters may contain too few members to be modeled by HMM and 2) The transporters may be too dissimilar to the query protein, such that most BLAST programs will not report them. For transporters with hits by a single search method, their scores under the single measurement should be comparable with the ones with weighted measurement. In order to compare these transporters, we applied the same measurements in the HMM scoring and BLAST scoring, namely *e*-value scores, and assigned the square root of their product as the weighted score, if the weighting was possible.

The second issue that was necessary to address was associated with the negative training data. Usually, transporter databases such as TCDB only contain transport proteins without specifying non-transport proteins, such as membrane proteins without transport functions. Lacking the negative information, many traditional classification methods would not work because positive features are undistinguishable from the negative ones. Therefore, we used the putative transmembrane segment (TMS) information, generated by either TMPRED (Hofmann and Stoffel, 1993) or HMMTOP (Tusnady and Simon, 2001), which enabled us to circumvent this problem, to some extent. It was observed that the numbers of transmembrane segments in most transporter families tended to vary slightly depending on the specific transport requirements and duplication processes that occurred during evolution. Therefore, our strategy set a minimal requirement for family membership; the number of transmembrane segments for an unknown protein must be within 1 SD of the mean number of transmembrane segments in that family for the candidate to be included in the family. Since non-transporters have either no transmembrane segments or the numbers of TMSs outside of the normal distributions of many transporter families, these non-transporters, including the membrane proteins without transport functions, were likely to have been filtered out during preprocessing. TMS number was not chosen as a filter simply because there are many proteins in TCDB with no reported TMS, but rather because more than 10% of transporters in TCDB failed to detect any TMS in both HMMTOP and TMPRED methods. Therefore, this simple filter will eliminate most non-transporters while retaining most transporters for analysis.

In order to demonstrate the proposed approach, a prediction web server, as well as a command line interface, was implemented. The Decypher hardware system, provided by Active Motif Incorporation, was utilized to conduct the BLAST search and multiple CPU/computers were run in parallel to perform HMM modeling and scoring. HMM models for all transporter families were pre-calculated and stored in order to improve the prediction speed. For the automatization of the predictive pipeline, default and uniform parameters were used for all TC families in all procedures, in particular, the threshold for Blast search was set to 100 and the threshold for family sizes applicable to HMM modeling and TMS filtering was set to 5. The final *e*-value threshold for transporter inclusion or exclusion was examined from 10 down to 0.0001, the frequently used *e*-values in sequence homologous searches. A detailed description of the parameters and the mathematical representation of our NN approach are present in the Supplementary Materials.

2.2 Five-fold cross-validation and literature-validation

To self-evaluate the classification performance of our NN approach, a five-fold cross-validation was conducted (Schaffer, 1993). The training transporter database, TCDB, was partition into five folds for training and testing. The fold partition was performed family by family. In each family with at least two members, the transporters were selected randomly and distributed into the five folds by natural order, until no transporter was available. For the natural order, the smaller folds had a higher priority than the larger folds in acquiring remaining transporters.

In order to evaluate the performance of an individual TC family in cross-validation, the precision and recall of each family was calculated; precision was defined as the proportion of correctly classified transporters among the total number of predictions in the family and recall was defined as the proportion of correctly classified transporters among the total number of transporters in the family. To evaluate the performance for multiple families, measurements such as the area under Receiver Operating Characteristic curve (Hand and Till, 2001) should be utilized; however, due to complication in calculation and explanation, the average precision and recall among the multiple families were calculated as an approximation. The performance of the entire cross-validation was estimated by applying correct classification rate, or for short, classification rate, a simple and well-accepted measurement. The classification rate of a cross-validation was defined as the proportion of testing transporters which had been correctly classified. Since there were no negative samples in the dataset, it should be strictly termed as positive classification rate. In addition, we also used the measurement to compare two classification methods, under fixed or nearly fixed prior distributions among distinct transporter families, since the positive classification rate was partially dependent on the prior probabilities of the families.

To further examine the performance of our approach on real data where negative data are present, seven model organisms and four non-model organisms were chosen in our prediction and literature-validation. These organisms included: *Escherichia coli O157:H7 EDL933*, *Saccharomyces cerevisiae S288C*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *A.thaliana*, *Oryza sativa*, *Homo sapiens*, *Picrophilus torridus DSM 970*, *Photobacterium profundum SS9*, *Desulfotalea psychrophila LSv54* and *Aspergillus fumigatus*. It is expected that the average performance of each organism will serve as a benchmark for the overall performance of our approach since there are no represented transporters in the TCDB for non-model organisms. Sequence data for each of the 11 organisms included in our analysis was acquired from NCBI, with the exception of *S.cerevisia* and *O.sativa*, which were download from SGD (ftp://ftp.yeastgenome.org/yeast/data_download/sequence/genomic_sequence/orf_protein/archive/) and TIGR (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa, version 4.0) and each genome version acquired was the one that mostly matched those in TransportDB, based on its reported amounts of proteins (http://www.membranetransport.org/complete_list.php).

The literature-validation of each of the 11 organisms was performed in two steps. Firstly, we compared the categorized transporters with those in TransportDB (Ren *et al.*, 2004, 2007) and calculated the proportion of proteins that overlapped. The comparison was convincing since the predictions in TransportDB had been annotated by biologists. Secondly, we compared the functional annotations accompanying with the predicted transporter sequences with the descriptions of the predicted TC families. Since the annotation information was

blinded during the prediction, we then analyzed the percentage of significant words in the description of TC families. In order to perform this analysis, a simple text mining program was developed which automatically identified overlapping words and then removed obviously insignificant English words. In addition, the program was designed to identify compatible words when calculating the percent of overlap, based on a series of compatibility rules generated on the basis of biological activity; for example: the abbreviation K^+ was compatible with words such as potassium, ions and metal.

3 RESULTS AND DISCUSSION

3.1 Results of five-fold cross-validation

The 3899 transporters within 484 TC families of TCDB were partitioned into five folds. At least 219 TC families were tested in all folds, including at least 160 ones which were built with HMM models (see the Supplementary Material for partition details). These data indicated a much larger coverage than those of SVM approaches utilized by Lin *et al.* (2006) which analyzed five TC superfamilies and three TC families.

Among the 484 families tested in at least one fold, 192 (~40%) had precisions and recalls of over 90% at the *e*-value threshold of 0.0001, a typical threshold in homologous searches. Of these 192 families, 147 had a perfect performance (Fig. 2). These results significantly outperformed those of other competitive approaches such as SVM (Lin *et al.*, 2006) and probabilistic suffix trees (PSTs) (Leonardi, 2006). In the SVM approach (Lin *et al.*, 2006), an average precision of 81.0% was only achieved on the five superfamilies and three families. The PST approach (Leonardi, 2006) only achieved comparable results to ours on about 10 families or superfamilies (details shown in the Supplementary Material).

Clustering the 484 TC families by size, our approach performed worse in smallest families, with a recall and precision of approximately 57.8 and 56.6%, respectively; both of which were dominated by the BLAST search (Fig. 3). Using our methodology, the recall and the precision increased in correlation with family size with larger families yielding higher recall and precision; however, precision increased more rapidly than the recall, with a range of 16.4–40.2% as compared with a range of 9.3–23.0% for increased recall. We believe that the observed difference in precision levels can be attributed to the more training sequences in the larger families and more importantly, by the removal of false positives through the integrated HMM weighting and TMS filtering.

The overall positive classification rates of the five-fold cross-validation is 72.3%, as determined by average classification rates of 74.4%, 73.8%, 72.6%, 71.7%, 71.3%, 70.2% with respect to *e*-value thresholds 10, 1, 0.1, 0.01, 0.001 and 0.0001, respectively. For families containing at least five transporters (corresponding to fold five), the average positive classification rate was 76.1% with a range from 72.9–77.9% among all the *e*-value thresholds. On the fixed *e*-value thresholds, the variation of positive classification rates was within 8.6% among the five folds, with smaller folds generally worse than larger folds. This variation on the classification rates is mainly caused by the inapplicability of HMM models in the small families of the smaller folds.

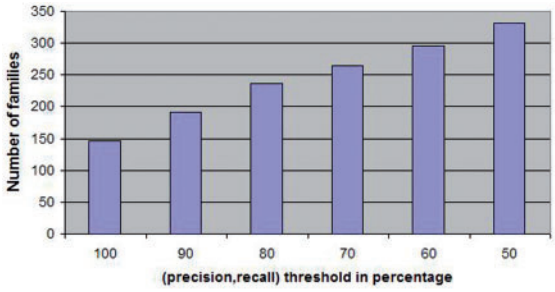

Fig. 2. The number of TC families with respect to specified performance ranges at *e*-value threshold 0.0001.

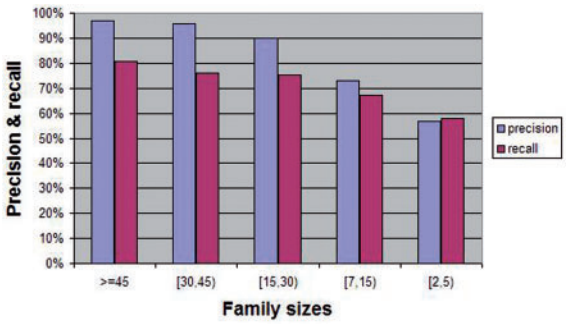

Fig. 3. The average precision and recall at different ranges of family sizes, where [a,b) means family sizes are from a to b, including a but exluding b.

The results in the five-fold cross-validation suggest that our NN approach satisfactorily distinguished the hundreds of transporter families in the TCDB through the integration of BLAST searches, HMM modeling and TMS filtering.

3.1.1 Comparative results on alternative combining strategies To examine the combined effectiveness of our approach, we compared alternative combinations of strategies among BLAST search, HMM modeling and TMS filtering, and evaluated them by the positive classification rates in the five-fold cross-validation at the *e*-value thresholds from 10 to 0.0001.

Firstly, we compared the performance of BLAST search and HMM modeling used in our combined approach. We ignored the TMS filtering since it was unable to singularly model TC families acceptably (Schwacke *et al.*, 2003). The positive classification rates of the large TC families, with more than five members, were analyzed using each of the two described methods (Fig. 4, compare upper and lower curves). The results of this study conclusively demonstrate that BLAST search outperformed HMM modeling in separating large TC families, under the variations of the *e*-value thresholds. It is also interesting to note that the BLAST search on all TC families (Fig. 4, middle curve) even outperformed the HMM modeling on large families in term of positive classification rates. These results are consistent with previously reported studies (Zhou *et al.*, 2003) and demonstrate why HMM modeling is not widely accepted as an independent method for transporter categorization.

Next, we examined the impact of combining HMM modeling into BLAST search on modeling performance. The integration

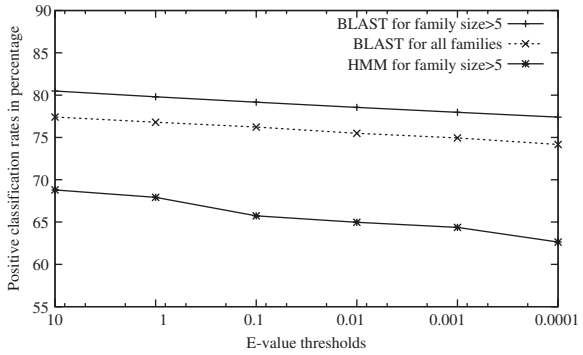

Fig. 4. Performance comparison between BLAST search and HMM modeling on classifying transporter proteins, evaluated by positive classification rates.

of HMM weighting resulted in a slight decrease of positive classification rates, which was 1.6% on average at the *e*-value thresholds, although in some cases, these rates demonstrated a slight increase. Since these analysis did not include negative samples, we next investigated the overall classification rate of the weighted strategy which we hypothesized would be higher than that of the BLAST search alone, since HMM modeling can filter out many false positives that are not characteristic of a specific family (Krogh *et al*., 1994). The hypothesis was confirmed in our literature-based validation of the 11 organisms, where the overall validation rates increased 7.0% on average. These results indicated that our methodology combines the strengths of the two methods.

Finally, we examined the impact of integrating TMS filtering into BLAST search and HMM modeling on performance. Again, the integration of TMS filtering resulted in a marginal decrease of positive classification rates, which was 1.9% on average at the *e*-value thresholds. When we performed the literature-based validation of the 11 organisms, the overall validation rates increased 3.7% on average after the TMS filtering. The two-fold impact of the TMS filtering was left as an option to users in our web server implementation, since in some cases, such as in partial EST sequences, the TMS filtering is not significant. The impacts of integrating TMS into BLAST search or into HMM modeling had similar results and can be referred to the Supplementary Materials.

The results of the comparative performance study suggest that our combined strategy is effective, due to the successful tradeoffs among BLAST search, HMM modeling and TMS filtering.

3.2 Literature-validation results on model and non-model organisms

The *e*-value threshold was set to 0.0001 in the prediction of transporters in the examined organisms since previous studies have concluded that this is an acceptable value for homologous searches (John and Sali, 2004). The complete results are shown in the Supplementary Materials. Overall, the predicted percentage of transporters was between 3.4 and 9.2% for eukaryotic species, and between 12.6 and 16.3% for archaean and bacterial species (refer to the second and the third column of Table 1). The predicted percentages were consistent with the overall range in TransportDB, especially considering the higher percentages of transporters in archaean and bacterial species as compared to eukaryotic species (Paulsen *et al*., 1998; Ren *et al*., 2007). In addition, the predicted transporters were observed to contain significantly larger numbers of transmembrane segments (3.7 to 6.0 time more) than those in the rejected non-transporters, which implied that our predicted proteins were more likely to be transporters than the rejected ones (Supplementary Materials).

The literature-validation results of the predictions are summarized in Table 1. An average of 65.8% of the predicted transporters were confirmed to be correctly categorized through our literature-based validation, including 55.9% of these categorized transporters was validated through TransportDB. The validated percent demonstrated an 83.6% overlap with putative transporters in the corresponding part of TransportDB. Among the overlaps, ~97.8% of them were exact matches, specifically these proteins were predicted as the same TC families or superfamilies. This is significant considering that TransportDB only reports superfamilies for some transporters. The remaining 2.2% of the overlapping predictions, so-called partial matches, were the proteins being predicted as transporters by both approaches but with conflicting superfamilies or families. Finally, it was believed that a significant portion of the unvalidated populations of putative transporter proteins were correctly categorized based on manual examinations of annotation information, along with the protein sequence data, through a random selection (Supplementary Material).

When we compared model and non-model organisms, where latter had no represented transporters in TCDB, there were no substantial differences in the overlapping rates and total validation rates. In analyzing the validation rates, we determined that there was less than a 2% difference, on average, between the two groups of organisms. Comparing our approach with alternative combined strategies, the integration of HMM modeling into BLAST search resulted in an average validation rate increase of 7.0%. Additionally, the overlapping rates with TransportDB demonstrated an average increase of 1.6% (Table 2). These data indicated an essential enhancement of HMM modeling to the BLAST search. On the other hand, TMS filtering had a two-fold impact on BLAST search and HMM modeling, with decreased overlapping rate but increased validation rate (Table 2).

Further examination of the performance of individual TC families demonstrated that average validation rates of 81.31 and 88.22% were achieved for families belonging to the two largest superfamilies among all the organisms (Chang *et al*., 2004), specifically facilitator superfamily (MFS) and ATP-binding-cassette superfamily (ABC). Of the 51 largest TC families whose predictions occurred in at least 2% of one or more organisms, 26 families were validated at least 80%. Furthermore, it was determined that an average of 91.70% of all predicted carrier families were validated. Finally, of the 395 TC families with at least one predictions, 248 (~63%) families were validated at least 60% (Supplementary Material).

One of the major concerns with regard to our validation results was that the number of predictions in our approach,

Table 1. The validation results of our predictions by TransportDB and our preliminary text mining program at *e*-value threshold 0.0001

Organism	Number of proteins	Our predictions	TransportDB predictions	Exact matches	Partial matches	TransportDB unique	Our validated unique	Overlap rate (%)	Total validated rate (%)
E.coli	5324	865	579	484	18	77	101	86.70	69.71
S.cerevisiae	6310	582	341	303	5	33	19	90.32	56.19
D.melanogaster	13 779	971	647	539	7	101	60	84.39	62.41
C.elegans	20 051	1214	669	561	12	97	115	85.50	56.67
A.thaliana	26 536	1593	976	845	15	116	162	88.11	64.16
O.sativa	55 890	1668	1285	968	36	281	133	78.13	68.17
H.sapiens	27 960	1801	948	822	15	111	349	88.29	65.85
P.torridus	1535	193	171	123	5	43	27	74.85	80.31
P.profundum	5491	800	582	463	7	112	70	80.76	67.50
D.psychrophila	3234	442	305	230	7	68	47	77.70	64.25
A.fumigatus	9923	864	620	518	7	95	69	84.68	68.75

Exact matches: proteins predicted both by our approach and TransportDB, with the same TC families or superfamilies. Partial matches: proteins predicted as transporters by both methods but with conflicted TC superfamilies if existed or families if not existed. TransportDB unique: the proteins occurring in TransportDB but absent in our predictions. Our validated unique: the proteins missed in TransportDB but validated by annotations together with protein sequences. Overlap rate: the sum of exact matches and partial matches divided by the number of predictions by TransportDB. Validation rate: the sum of exact matches, partial matches and our validation unique divided by the number of our predictions.

Table 2. Comparison of prediction and validation results among alternative combinations of methods at *e*-value threshold 0.0001

Organism	Predicted transporters			Overlap rate with TransportDB (%)			Validation rate (%)		
	ALL[a]	BLAST & HMM	BLAST	*ALL*[a]	BLAST & HMM	BLAST	*ALL*[a]	BLAST & HMM	BLAST
E.coli	865	916	933	86.70	90.50	89.29	69.71	68.34	66.67
S.cerevisiae	582	609	681	90.32	92.67	91.20	56.19	55.01	49.05
D.melanogaster	971	1095	1289	84.39	87.79	85.94	62.41	57.44	48.10
C.elegans	1214	1357	1534	85.50	88.49	85.05	56.67	52.47	44.85
A.thaliana	1593	1757	2177	88.11	90.27	90.88	64.16	59.36	48.78
O.sativa	1668	1957	2575	78.13	85.37	86.69	68.17	63.67	50.14
H.sapiens	1801	2236	2769	88.29	93.04	93.57	65.85	56.93	47.85
P.torridus	193	199	182	74.85	77.19	63.16	80.31	79.90	73.63
P.profundum	800	871	925	80.76	84.02	84.02	67.50	64.52	61.08
D.psychrophila	442	493	538	77.70	80.00	81.64	64.25	59.43	56.32
A.fumigatus	864	980	1080	84.68	92.58	92.42	68.75	66.12	60.28
Average				83.58	87.45	85.81	65.82	62.11	55.16

[a]The approach we applied in this article which combined BLAST searches, HMM modeling and TMS filtering.

in some instances, was nearly twice that of TransportDB. We believe that there are three likely causes for his observation: (1) TransportDB only focuses on solute and cytoplasmic membrane ion transporters and artificially excludes some transporter families such as sodium ion-transporting carboxylic acid decarboxylases (Ren *et al*., 2004); (2) Our approach may identify many novel members of transporter families or superfamilies that TransportDB does not. For example, among the 349 predictions in *H.sapiens* that were validated by our methodology, but were not identified by TransportDB (Table 1, eighth column), 117 were in families or superfamilies that TransportDB studied. Therefore, some of the novel predictions may have been identified as a consequence of the integration of HMM models. For example, a tumor suppressing protein in *H.sapiens* named GI:34734073/NP_899056 was identified as a member of transporter family 2.A.1.2 by our approach but excluded from the superfamily 2.A.1 that TransportDB reported. Its BLAST *e*-value of 24.76 to the superfamily was somewhat large and would therefore be excluded from consideration by most BLAST searches. However, in our approach the weighted *e*-value was 5.15*e*−10 and could be detected by most *e*-value thresholds, which was attributed to an HMM *e*-value of 1.07e−20 and (3) Our approach may identify some novel transporters without TMS and this population of putative transporters may encompass up to 11–20% of predictions (Supplementary Material), since our method uses the SD of TMS on transporter families to filter unlikely proteins rather than used non-zero TMS numbers as filters. Although the proportion of predictions was in low-confidence, a number of true transporters were identified and

confirmed by manual validation. For example, a nicotinamide ribonucleoside (NR) uptake permease in *E.coli* named GI:5804961/NP_291003 had no detected TMS by either HMMTOP or TMPRED, but was correctly categorized into the corresponding TC family 4.B.1 using our approach.

The literature-validation results of model and non-model organisms indicated that our NN approach is effective and therefore shows great promise as a methodology to screen putative transporters from unknown protein sequences on a genome scale. We are currently using these approaches to generate predictions and verify these predictions using wet-lab techniques on the model legume organism *Medicago truncatula* and fungal organism *Epichloe festucae*.

4 CONCLUSION

In summary, we have developed an automated approach for transporter prediction and categorization through seamlessly integrating the analysis of three types of information from protein sequences within a modified NN framework. Specifically this approach integrates the analysis of (1) TMS information, which was used to preprocess data by filtering out sequences based on conflicted TMS numbers; (2) BLAST *e*-value scores, which were used as the primary measurement of similarity in the NN classification and (3) HMM *e*-value scores, which were weighted to reflect the extent to which an unknown protein matched to the overall feature of a family. By combining HMM scores, an unknown protein was categorized into a TC family based on how well the unknown protein matched one of the transporters in the family as well as how well the features of the putative family member protein matched the overall features of the family. To examine the effectiveness of our proposed approach, we conducted a five-fold cross-validation on TCDB and literature-based validations to the predictions of both model and non-model organisms. The validation of our methodology by these two criteria demonstrated the effectiveness of our approach in detecting unknown transporter proteins and postulating their unknown transport mechanisms on a genome scale.

Our described approach is distinct from related methods such as TCDB and TransportDB in a number of ways including: (1) neither TCDB nor TransportDB utilize HMM in modeling the characteristics of TC families or in predicting novel transporters, although TCDB uses HMM in topological analysis following the preliminary predictions; (2) while our methodology is independent of potential human biases, both TCDB and TransportDB involve human intervention, especially TCDB is highly dependent upon an intensive human curation of data and (3) in TCDB, other information including substrates were analyzed; however, our approach depends solely upon primary sequences data (Busch and Saier, 2002; Ren *et al*., 2004, 2007; Saier, 2000; Saier *et al*., 2006).

Nevertheless, our approach has some constraints. First, it may suffer from lack of credible and comprehensive negative samples in the training data. Secondly, since it is confined by the features presented in the TCDB, the potential exists that certain novel types of transporters will not be detected. We believe that the constraints of the novel methodology described in this study have the potential to be overcome by integrating the analysis of additional reference databases such as PFAM (Sonnhammer *et al*., 1997), TIGRFAMS (Haft *et al*., 2001), GO (Ashburner *et al*., 2000) and SWISS-PROT (Apweiler, 2001) which will result in more confidence and enable clustering of the unknown sequences in order to detect additional novel types of transporters. Nevertheless, the computational methodology described in this work represents a novel strategy for the identification and categorization of putative transport proteins based on protein sequence analysis.

ACKNOWLEDGEMENTS

Funding support for the work is provided by the Noble Foundation. The authors are grateful for the comments and suggestions raised by their colleagues, Drs. Michael Udvardi, Carolyn Young, Ji He, Ranamalie Amarasinghe, Vagner Benedito, Rakesh Kaundal and the anonymous reviewers of this article.

Conflict of Interest: none declared.

REFERENCES

Altschul,S. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Apweiler,R. (2001) Functional information in swiss-prot: the basis for large-scale characterisation of protein sequences. *Brief. Bioinform*., **2**, 9–18.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet*, **25**, 25–29.

Bejerano,G. and Yonam,G. (2001) Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics*, **17**, 23–43.

Busch,W. and Saier,M.J. (2002) The transporter classification (tc) system, 2002. *Crit. Rev. Biochem. Mol. Biol*., **37**, 287–337.

Chang,A. *et al*. (2004) Phylogeny as a guide to structure and function of membrane transport proteins. *Mol. Membr. Biol*., **21**, 171–181.

Cover,T. and Hart,P. (1967) Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*, **13**, 21–27.

Dibrov,P. and Fliegel,L. (1998) Comparative molecular analysis of NA^+/H^+ exchangers: a unified model for NA^+/H^+ antiport? *FEBS Lett*, **424**, 1–5.

Doolittle,R. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.

Eskin,E. *et al*. (2003) Protein family classification using sparse markov transducers. *J. Comput. Biol*., **10**, 187–213.

Haft,D. *et al*. (2001) Tigrfams: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*., **29**, 41–43.

Hand,D.J. and Till,R.J. (2001) A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn*., **45**, 171–186.

Heil,B. *et al*. (2006) Computational recognition of potassium channel sequences. *Bioinformatics*, **22**, 1562–1568.

Henikoff,S. and Henikoff,J. (1994) Protein family classification based on searching a database of blocks. *Genomics*, **19**, 97–107.

Hofmann,K. and Stoffel,W. (1993) Tmbase—a database of membrane spanning proteins segments. *Biol. Chem*., **374**, 166.

Horton,P. and Nakai,K. (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology* vol. 5. Halkidiki, Greece, pp. 147–152.

John,B. and Sali,A. (2004) Detection of homologous proteins by an intermediate sequence search. *Protein Sci*., **13**, 54–62.

Krogh,A. *et al*. (1994) Hidden markov models in computational biology: applications to protein modeling. *J. Mol. Biol*., **235**, 1501–1531.

Leonardi,F. (2006) A generalization of the pst algorithm: modeling the sparse nature of protein sequences. *Bioinformatics*, **22**, 1302–1307.

Lin,H. *et al*. (2006) Prediction of transporter family from protein sequence by support vector machine approach. *Proteins*, **62**, 218–231.

Paulsen,I. *et al*. (1998) Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol*., **277**, 573–592.

Pruitt,K. *et al*. (2005) Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33** (Suppl. 1), D501–D504.

Ren,Q. *et al*. (2004) Transportdb: a relational database of cellular membrane transport systems. *Nucleic Acids Res*., **32** (Database issue), D284–D288.

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

Rigaud,J. *et al*. (1995) Reconstitution of membrane proteins into liposomes: application to energy-transducing membrane proteins. *Biochim. Biophys. Acta*, **1231**, 223–246.

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

Saier,M.J. *et al*. (2006) Tcdb: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res*., **34** (Database issue), D181–D186.

Sakmann,B. and Neher,E. (1984) Patch clamp techniques for studying ionic channels in excitable membranes. *Annu. Rev. Physiol*., **46**, 455–472.

Schaffer,C. (1993) Selecting a classification method by cross-validation. *Mach. Learn*., **13**, 135–143.

Schwacke,R. *et al*. (2003) Aramemnon, a novel database for arabidopsis integral membrane proteins. *Plant Physiol*, **131**, 16–26.

Sonnhammer,E. *et al*. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.

Tusnady,G. and Simon,I. (2001) The hmmtop transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.

Yan,Q. (2003) *Membrane Transporters: Methods and Protocols*. vol. 227 of *Methods in Molecular Biology*. Humana Press, Totowa, New Jersey, USA.

Zdobnov,E. and Apweiler,R. (2001) Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, **17**, 847–848.

Zhai,Y. and Saier,M.J. (2001) A web-based program (what) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. *J. Mol. Microbiol. Biotechnol*., **3**, 501–502.

Zhou,X. *et al*. (2003) An automated program to screen databases for members of protein families. *J. Mol. Microbiol. Biotechnol*., **5**, 7–10.