TpPred: A Tool for Hierarchical Prediction of Transport Proteins Using Cluster of Neural Networks and Sequence Derived Features

Article *in* International Journal for Computational Biology · August 2012

DOI: 10.34040/IJCB.1.1.2012.18

4 authors:

Sankalp Jain
University of Vienna
15 PUBLICATIONS **138** CITATIONS
SEE PROFILE

Piyush Ranjan
Georgia Institute of Technology
20 PUBLICATIONS **742** CITATIONS
SEE PROFILE

Dipankar Sengupta
University of Westminster
34 PUBLICATIONS **113** CITATIONS
SEE PROFILE

Pradeep K Naik
Sambalpur University
174 PUBLICATIONS **2,499** CITATIONS
SEE PROFILE

TpPred: A Tool for Hierarchical Prediction of Transport Proteins Using Cluster of Neural Networks and Sequence Derived Features

Sankalp Jain, Piyush Ranjan, Dipankar Sengupta, and Pradeep Kumar Naik*

Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan 173234, Himachal Pradesh, India.

*Email: pknaik1973@gmail.com

Abstract: A top–down predictor, called TpPred, is developed which consists of 3 level of hierarchical classification using cascade of neural networks from sequence derived features. The 1st layer of the prediction engine is for identifying a query protein as transport protein or not; the 2nd layer for the main functional class; and the 3rd layer for the sub-functional class. The overall success rates for all the three layers are higher than 65% that were obtained through rigorous cross-validation tests on the very stringent benchmark datasets in which none of the proteins has 30% sequence identity with any other in the same class or subclass. TpPred achieved good prediction accuracies and could nicely complement experimental approaches for identification of transport proteins. TpPred is freely available to be use in-house as a standalone version and is accessible at http://www.juit.ac.in/attachments/tppred/Home.html.

Keywords: Transport proteins,hierarchical classification, neural networks, sequence derived features

1 Introduction

Transport proteins are biologically important and play indispensable roles in the fundamental cellular processes of all organisms. They are involved in the transport of ions and molecules across the membrane, play essential roles in cellular metabolism and activities. They mediate the entry of nutrients into cytoplasm and the extrusion of metabolite wastes, maintain a stable internal environment inside the cell by regulating the uptake and efflux of ions, protect cells from environmental insults, and enhance communications between cells through the secretion of proteins, carbohydrates and lipids [1-3]. Specific transporters have been explored as therapeutic targets [4-6]. A variety of transporters are responsible for the absorption, distribution and excretion of drugs within the human body which must be factored into pharmacological studies [7,8]. Different transport systems differ in their putative membrane topology, energy coupling mechanism and substrate specificities [9]. The immense importance of studying transport proteins and the enormity of the data available on these proteins has warranted the systematic annotation and classification of transport proteins for elucidating the functional mechanisms of proteins and biological processes.

Transport proteins have been identified by such experimental approaches as absorbance spectroscopy, gel electrophoresis, metal-affinity columns and shift assay, chromatography, mass spectroscopy, and combined spectroscopic studies. However, some of these methods generally require a purified or semi-purified target of interest, do not facilitate identification of unknown targets form complex protein mixtures, or require multi-step processes and very specialized equipment, which limit their application ranges. Therefore, there is need to explore other methods including computational approaches for facilitating the identification of transport proteins to complement these experimental methods. With the explosion of protein sequences entering into databanks, it is highly desirable to explore the feasibility of selectively classifying newly found protein sequences into their respective transport protein classes by means of an automated method [10, 11]. This is indeed important because knowing which protein belongs to which particular class may help to deduce its catalytic mechanism and specificity, giving clues to the relevant biological function. Primary sequence of these proteins are readily available, therefore a method using the sequence derived features will prove a much valuable and a cost effective process of determining and classifying these proteins into broader transporter/non-transporter and specifically into major classes and subclasses as defined by Transport Classification (TC) system (http://www.tcdb.org/browse.php) [12].
So far, sequence alignment and clustering are the primary method for predicting the TC family, as well as the function of transporters [13, 14]. Some transporters are known to have no or low homology to other proteins of known function [15-18]. A substantial portion of transporters in different TC families have been found to have very low sequence identity to other family members. For instance, a member of the multidrug transporter family, bmr3, has only 7% sequence identity and 17% similarity to another family member blt [18]. The potassium channel, TASK-2, has 18–22% sequence identity to other members of the two-pore domain K^+ channel family, such as TWIK-1, TREK-1, TASK-1, and TRAAK [19]. Two members of the major facilitator family, GlpT and LacY, are 21% identical to each other [21]. Thus, the function of some of these transporters may be difficult to assign based solely on homology, [21, 22] and methods that predict protein function without the use of sequence similarity are needed.

This work explored a machine learning method, artificial neural network (NN) that predicts transport proteins directly from sequence or sequence-derived properties. The sequence derived features that were used are amino acid composition, pseudo amino acid composition and physicochemical properties. Using these parameters and their combination we have developed a cluster of neural networks for the hierarchical classification of transport proteins in a “top-down” approach.

2 Materials and methods

2.1 Preparation of dataset

All transport proteins used in this study are taken from the Transport Classification Database (http://www.tcdb.org/) in which the proteins are classified on the basis of their function [12]. A total of 5,359 transport protein sequences taken together, have been classified into seven major classes as: channels/pores (1139), electrochemical potential-driven transporters (1456), primary active transporters (2045), group translocators (107), transmembrane electron carriers (106), accessory factors involved in transport (129) and incompletely characterized transport systems (377). With the aim of avoiding prejudiced learning in the networks, we scaled the sequences such that the inequality in the data points or number of protein sequences in each class may be compromised.We reduced the proteins in each class with a similarity cutoff of 30% using BLASTClust [23]. A negative dataset consisting of 2,907 protein sequences, representing non-transport members is also created from PDB database. These datasets are divided into separate training, testing and independent evaluation sets (Table 1).

2.2 Feature vector construction

Following three types of discrete feature vectors were constructed for each protein sequence.

1. *Amino acid composition*: Given the sequence of a protein, its amino acid composition was computed and then used to generate a set of 20 features representing composition of 20 standard amino acids in the protein sequences that include A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y. These features have been widely used in predicting different structural classes and subcellular localization of proteins [10,11,24]. The formula used to calculate amino acid composition is:

$$AAcomp(i) = \frac{AA(i)}{\sum_{j=1}^{20} AA(j)}$$

where AA(i) = Frequency of i^{th} amino acid in the sequence

2. *Physicochemical properties:* Twelve sequence derived properties for each protein sequence was calculated using EMBOSS (EBI) package [25]. The parameters include: molecular weight, totalcharge, isoelectric point, mole percentages of tiny (A+C+G+S+T), small (A+B+C+D+G+N+P+S+T+V), aliphatic (I+L+V), aromatic (F+H+W+Y), non-polar (A+C+F+G+I+L+M+P+V+W+Y), polar (D+E+H+K+N+Q+R+S+T+Z), charged (B+D+E+H+K+R+Z), acidic (B+D+E+Z) and basic (H+K+R) amino acids .

3. *Pseudo amino acid composition (PseAA):* This class of descriptor consists of a set of 37 features, 20 of which are weighted amino acid compositions and rest 17 are correlation factors calculated among amino acids for each protein sequence [26].

A protein sequence **P** with L amino acid resides can be represented as:

$$P = R_1 R_2 R_3 R_4 \ldots R_L \quad (1)$$

where R_1 represents the 1st residue of the protein **P**, R_2 the 2nd residue and so forth. According to the simplest discrete model, the amino acid composition of the protein **P**based on the equation (1) can be expressed as:

$$P = [f_1 \quad f_2 \quad \cdots \quad f_{20}]^T \quad (2)$$

where $f_u (u = 1,2,...,20)$ are the normalized occurrence frequencies for

the 20 native amino acids in **P** and **T** the transposing operator. The additional 17 features are a series of rank-different correlation factors along a protein chain and were calculated as follows.

A protein sequence **P** consisting of *L* amino acid resides can be represented as:

$$P=[p_1 \;\; p_2 \;\; \cdots \;\; p_{20} \;\; p_{20+1} \;\; \cdots \;\; p_{20+\lambda}]^T, (\lambda<L) \quad (3)$$

where $20 + \lambda$ components are given by

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda}\tau_k}, (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda}\tau_k}, (20+1 \le u \le 20+\lambda) \end{cases} \quad (4)$$

where*w* is the weight factor and τ_k is the k^{th} tier correlation factor that reflects the sequence order correlation between all the k^{th} most contiguous residues as formulated by

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,i+k}, (k < L) \quad (5)$$

with

$$J_{i,i+k} = \frac{1}{\Gamma}\sum_{g=1}^{\Gamma}\left[\Phi_\xi(R_{i+k}) - \Phi_\xi(R_i)\right]^2 \quad (6)$$

where$\Phi_\xi(R_i)$is the ξ-th function of the amino acid R_i, and $\Gamma$the total number of the functions considered. $\Phi_1(R_i)$, $\Phi_2(R_i)$ and $\Phi_3(R_i)$ represented respectively the hydrophobicity value [27], hydrophilicity value [28], and side chain mass of amino acid R_i (Table 2); while $\Phi_1(R_{i+k})$, $\Phi_2(R_{i+k})$ and $\Phi_3(R_{i+k})$ are the corresponding values for the amino acid R_{i+k}. Therefore, the total number of functions considered is Γ=3.

It can be seen from equation (3) that the first 20 components, i.e. p_1, p_2, …, p_{20} are associated with the conventional AA composition of protein, while the remaining components p_{20+1}, …, $p_{20+\lambda}$are the correlation factors that reflect the 1st tier, 2nd tier, …, and the λ^{th} tier sequence order correlation patterns. It is through these additional λ factors the important sequence-order information are incorporated.

2.3 System architecture and component of NN topology

The overall classification system consists of three layer of successive multilayer feed forward (acyclic) artificial NNs (Fig. 1), each one with a single hidden layer at which the computation takes place. Some common features shared by all NNs are the following:

1. There is full connectivity as every node in each network layer is connected to every other node in the adjacent forward layer.
2. There are a small number of nodes in the hidden layer responsible for the actual learning process carried out by each component network.
3. The activation function on each node is a nonlinear, sigmoid logistic function of the weighted sum of all synaptic weights (plus a constant bias).

NN1 is a binary classifier which classifies an input protein sequence as a transport protein or non-transport protein. If the input protein sequence is classified as a transport protein then it is processed by NN2 which gets classified into one of the seven main classes of transport proteins (channels/pores, electrochemical potential-driven transporters, primary active transporters, group translocators, transmembrane electron carriers, accessory factors involved in transport and incompletely characterized transport systems). Each class (except electrochemical potential-driven transporters) consists of an independent NN [channels/pores (NN3), primary active transporters (NN4), group translocators (NN5), transmembrane electron carriers (NN6), accessory factors involved in transport (NN7) and incompletely characterized transport systems (NN8)] for classification of input protein sequence specifically into its functional sub-class. We have used three categories of sequence derived features such as physicochemical

properties, amino acid composition and pseudo amino acid composition for training of NNs. Using these parameters independently and with combination we have developed seven neural network clusters:, NN_{AAcomp}, NN_{pseAA}, $NN_{prop,}$ $NN_{AAcomp+pseAA,}$ $NN_{AAcomp+prop}$, $NN_{pseAA+prop}$, and $NN_{AAcomp+pseAA+prop}$. Before the learning process, all network synaptic weights are initialized to small random values which have been optimized to final weights during learning process based on backpropagation algorithm [29].

An important issue in the design of a NN classification system is the network's generalization, that is, its ability to give correct predictions when it is presented with unseen examples. With a small number of training samples and a relatively large number of synaptic weights, there is always the possibility that the network's free parameters will adapt to the special features of the training data (overfitting). A straightforward way to overcome this problem is to use sufficient number of training examples (usually more than 30 times the number of adjustable network parameters). However, the protein classes are unbiased and it is not possible to have these many numbers. Therefore to control the over fitting in our application, we have employed nonconvergent criteria (early stopping method); the training process is stopped before the optimization procedure finished. We follow the common method which is to withhold and use part of the training data (20%) as an internal validation set. Training is stopped at the point at which the classification error on the holdout subset begins to rise.

In the prediction phase, just like the forward pass in learning, network weights are globally fixed (those obtained after the convergence of the training process) and the NN is presented with an unknown example for classification. In the same hierarchical manner, the input signal propagates once in the forward direction and the output value constitutes the network's decision based on the already studied training examples. The prediction accuracy of the models has been validated using self-consistency, jackknife and independent data set. For jackknife test we randomized the test set for 100 times and recorded average performance accuracy.

3 Results and discussion

Neural network has been successfully used previously for predicting the functional classes of proteins from sequence-derived structural and physicochemical properties and irrespective of sequence similarity [30-32]. However, transport proteins involve a substantially more diverse spectrum of proteins than most of the other classes of proteins. The diverse spectrum of proteins poses a more critical test for constructing a NN prediction system. In order to assess the performance of the TpPred, we applied several tests. We created a new independent test set with well-characterized protein sequences from all level of classes and sub-classes (Table 1) to evaluate the performance of the new integrated system. In addition we have also performed sub-sampling test (self consistency test) and jackknife test for evaluating the performance of TpPred. These validation tests are commonly used for measuring the accuracy of a classifier [10, 33-35]. The performance of neural networks with combined features (especially the one combined all three types of features) tend to perform better than the one using only a single type of features or less type of features.

3.1 Performance of 1st layer of neural network

The performance and validation results of NN1 are given in Table 3. The network achieved an overall accuracy of 97.3% and 88.4% for the training set and test set data using combination of sequence derived features—amino acid composition, pseudo amino acid composition and physicochemical properties. While considering the validation techniques by using an independent data set, self consistency test and jackknife test, the overall accuracy of the 1st layer of TpPred is 85.2%, 88.0% and 81.4% respectively. The details of the performance accuracy and validation results based on different types of sequence derived feature have been represented in supplementary Table S1.

3.2 Performance of 2nd layer of neural network

The overall success rate in identifying the transport proteins among their seven major functional classes is 97.5% (using training set) and 75.0% (using test set) (Table 4). Similarly the overall performance accuracy based on three types of validation tests has been found to be 79.8% (using independent data set), 84.2% (using self consistency test) and 68.5% (using jackknife test). The corresponding results by TpPred on the data set for seven major classes of transport proteins using different types of sequence derived features are given in supplementary Table S2.

3.3 Performance of 3rd layer of neural network

The performance accuracy and validation results of NNs in identifying subclasses of channels/pores (NN3), primary active transporters (NN4), group translocators (NN5), transmembrane electron carriers (NN6), accessory factors involved in transport (NN7) and incompletely characterized transport systems (NN8) using combination of all sequence derived features has been given in Table 5. The corresponding results by TpPred on the detection of α-type channels (1.A), β-barrel porins (1.B), pore-forming toxins (1.C) and holins (1.D) are 94.4% (training set), 83.2% (test set), 69.5% (independent data set), 70.0% (self consistency test) and 64.6% (jackknife test) on the data set 'S1'. Similarly for the data set 'S3' the performance accuracy for the detection of P-P-bond-hydrolysis-driven transporters (3.A), decarboxylation-driven transporters (3.B), oxidoreduction-driven transporters (3.D) and light absorption driven transporters (3.E) is 95.1% (training set), 95.0% (test set), 73.3% (independent data set), 79.3% (self consistency test) and 68.5% (jackknife test). For the data set 'S4', the performance accuracy for the detection of phosphotransfer-driven group translocators (4.A) and acyl CoA ligase-coupled transporters (4.C) is 100% (training set), 100% (test set), 80.4% (independent data set), 86.8% (self consistency test) and 73.0% (jackknife test). For the data set 'S5', the performance accuracy for the detection of transmembrane 2-electron transfer carriers (5.A) and transmembrane 1-electron transfer carriers (5.B) is 100% (training set), 100% (test set), 95.4% (independent data set), 96.8% (self consistency test) and 82.7% (jackknife test). For the data set 'S6', the performance accuracy for the detection of auxiliary transport proteins (8.A) and ribosomally synthesized protein-peptide toxins (8.B) that target channels and carriers proteins is 97.4% (training set), 100% (test set), 83.0% (independent data set), 86.7% (self consistency test) and 76.3% (jackknife test). The overall accuracy of detection of recognized transporters of unknown biochemical mechanism (9.A) and putative transport proteins (9.B) is 100% (training set), 97.8% (test set), 82.2% (independent set), 87.8% (self consistency test) and 73.7% (jackknife test) for the data set 'S7'. The details of the performance accuracy have been represented in supplementary Table S3.

For the current data sets in which none of the protein sequence has ≥30% sequence identity to any others in a same class or subclass, the overall success rates by the TpPred in identifying the main functional classes of transport proteins and their subclasses is very high. In an earlier study, contribution of individual feature property to protein classification is investigated by separately conducting classification by the use of each feature property [36-38]. The same method was employed here. An analysis on the classification of the group of all transport proteins seems to suggest that, in order of prominence, the hydrophobicity and hydrophilicity play more prominent role than other feature properties. Hydrophobicity has been shown to be important for its membrane binding properties. It was also found that polarity and solvent accessibility of the binding site influences the functional properties of proteins [39]. Therefore, our prediction results are consistent with these experimental findings. Overall TpPred is a very powerful predictor in identifying transport proteins, their main classes, and their subclasses.

4 Conclusion

From a practical point of view, the most important aspect of a prediction model is its ability to make correct predictions. Till date most of the available methods use the 3-D

structure of the protein to predict and classify transport proteins. This is a very tedious job and requires much costlier endeavors. The sequence of a protein is an important determinant for the detailed molecular function of proteins and would consequently also be useful for prediction of transport protein and classes. Additionally much encouraging results have been predicted using the sequence derived features. Therefore, a much accurate and reliable method is that which predicts the transport proteins and their classes based on both strategies. Cascade of neural networks used in this study appears to be a potentially useful tool for the prediction of transport proteins of different classes. The prediction accuracy may be further enhanced with the further expansion of our knowledge about transport proteins particularly for those small transport classes, more refined representation of the structural and physicochemical properties of proteins, and the improvement of prediction algorithms such as the better treatment of imbalanced dataset.

References

[1] Hediger MA, "Structure, function and evolution of solute transporters in prokaryotes and eukaryotes", J. Exp. Biology, 196: 15–49 (1994).

[2] Borst P and Elferink RO, "Mammalian ABC transporters in health and disease", Annu. Rev. Biochemistry, 71: 537–592 (2002).

[3] Seal RP and Amara SG, "Excitatory amino acid transporters: a family in flux", Annu. Rev. Pharmacol. Toxicology, 39: 431–456 (1999).

[4] Joet T, Morin C, Fischbarg J, Louw AI, Eckstein-Ludwig U, Woodrow C and Krishna S, "Why is the Plasmodium falciparum hexose transporter a promising new drug target?", Expert. Opin.Ther. Targets, 7: 593–602 (2003).

[5] Birch PJ, Dekker LV, James IF, Southan A and Cronk D, "Strategies to identify ion channel modulators: current and novel approaches to target neuropathic pain", Drug Discov. Today, 9: 410–418 (2004).

[6] Dutta AK, Zhang S, Kolhatkar R and Reith ME, "Dopamine transporter as target for drug development of cocaine dependence medications", Eur. J. Pharmacology, 479: 93–106 (2003).

[7] Lee W and Kim RB, "Transporters and renal drug elimination", Annu. Rev. Pharmacol. Toxicology, 44: 137–166 (2004).

[8] Kunta JR and Sinko PJ, "Intestinal drug transporters: in vivo function and clinical importance", Curr. Drug Metabolism, 5: 109–124 (2004).

[9] Driessen AJ, Rosen BP and Konings WN, "Diversity of transport mechanisms: common structural principles", Trends Biochem. Science, 25: 397–401 (2000).

[10] Chou KC and Zhang CT, "Prediction of protein structural classes", Crit. Rev. Biochem. and Mol. Biology, 30: 275-349 (1995).

[11] Klein P, "Prediction of protein structural class by discriminant analysis", Biochem. Biophys. Acta, 874: 205-215 (1986).

[12] Saier MH, Tran CV and Barabote RD, "TCDB: the transporter classification database for membrane transport protein analyses and information", Nucl. Acids Research, 34: 181-186 (2006).

[13] Zhou X, Hvorup RN and Saier MH Jr, "An automated program to screen databases for members of protein families", J. Mol. Microbiol. Biotechnology, 5: 7–10 (2003).

[14] Campbell RS, Brearley GM, Varsani H, Morris HC, Milligan TP, Hall SK, Hammond PM and Price CP, "Development and validation of a robust specific enzyme mediated assay for phenylalanine in serum", Clin. Chim. Acta, 210: 197–210 (1992).

[15] Howard EM, Zhang H and Roepe PD, "A novel transporter, Pfcrt, confers antimalarial drug resistance", J. Membr. Biology, 190: 1–8 (2002).

[16] Sano Y, Inamura K, Miyake A, Mochizuki S, Kitada C, Yokoi H, Nozawa K, Okada H, Matsushime H and Furuichi K, "A novel two-pore domain K_channel, TRESK, is localized in the spinal cord", J. Biol. Chemistry, 278: 27406–27412 (2003).

[17] Zhang Y, Jock S and Geider K, "Genes of Erwinia amylovora involved in yellow color formation and release of a low-molecular-weight compound during growth in the presence of copper ions", Mol. Gen. Genetics, 264: 233–240 (2000).

[18] Ohki R and Murata M, "bmr3, a third multidrug transporter gene of Bacillus

subtili", J. Bacteriology, 179: 1423–1427 (1997).
[19] Reyes R, Duprat F, Lesage F, Fink M, Salinas M, Farman N and Lazdunski M, "Cloning and expression of a novel pH-sensitive two pore domain K_channel from human kidney", J. Biol. Chemistry, 273: 30863–30869 (1998).
[20] Vardy E, Arkin IT, Gottschalk KE, Kaback HR and Schuldiner S, "Structural conservation in the major facilitator superfamily as revealed by comparative modelling", Protein Science, 13: 1832–1840 (2004).
[21] Enright AJ and Ouzounis CA, "GeneRAGE: a robust algorithm for sequence clustering and domain detection", Bioinformatics, 16: 451–457 (2000).
[22] Whisstock JC and Lesk AM, "Prediction of protein function from protein sequence and structure", Q. Rev. Biophysics, 36: 307–340 (2003).
[23] Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ, "Basic local alignment search tool", J. Mol. Biology, 215: 403-410 (1990).
[24] Anfisen CB, "Principles that govern the folding of protein chains", Science, 181: 223-230 (1973).
[25] Rice P, Longden I and Bleasby A, "EMBOSS: The European Molecular Biology Open Software Suite", Trends in Genetics, 16: 276-277 (2000).
[26] Chou KC, "Prediction of protein cellular attributes using pseudo amino acid composition" Proteins, Structure, Function and Genetics, 43: 246-255 (2001).
[27] Tanford C, "Contribution of hydrophobic interactions to the stability of the globular conformation of proteins", J. American Chem. Society, 84: 4240-4247 (1962).
[28] Hopp TP and Woods KR, "Prediction of protein antigenic determinants from amino acid sequences", Proc. Natl. Acad. Science, 78: 3824-3828 (1981).
[29] Rumelhart DE, Hinton GE and Williams RJ, "Learning internal representations by error propagation", In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: Foundations, Cambridge, MA: MIT Press, 318-362 (1986).
[30] Jaiswal K, Kumar C and Naik PK, "Prediction of EF-hand calcium-binding proteins and identification of calcium-binding regions using machine learning techniques", J. Cell Mol. Biology 8(2): 41-49 (2010).
[31] Patel A, Patel S and Naik PK, "Binary classification of uncharacterized proteins into DNA binding/non-DNA binding proteins from sequence derived features using ANN", Digest J. of Nanomateria and Biostructure 4(4): 775-782 (2009).
[32] Naik PK, Mishra VS, Gupta M and Jaiswal K, "Prediction of enzymes and non-enzymes from protein sequences based on sequence features and PSSM matrix using artificial neural network", Bioinformation 2(3): 107-112 (2007).
[33] Zhou GP, "An intriguing controversy over protein structural class prediction", J. Protein Chemistry 17: 729–738 (1998).
[34] Chou KC and Cai YD, "Using functional domain composition and support vector machines for prediction of protein subcellular location", J. Biol. Chemistry 277: 45765–45769 (2002).
[35] Huang Y and Li Y, "Prediction of protein subcellular locations using fuzzy k-NN method", Bioinformatics 20: 21–28 (2004).
[36] Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B and Chen YZ, "Prediction of the functional class of lipid-binding proteins from sequence derived properties irrespective of sequence similarity", J. Lipid Research 47: 824-831 (2006).
[37] Fierro-Monti I and Mathews MB, "Proteins binding to duplexed RNA: one motif, multiple functions", Trends in Biochem. Science 25: 241-246 (2000).
[38] Perez-Canadillas JM and Varani G, "Recent advances in RNA-protein recognition", Curr. Opin.Struct. Biology, 11: 53-58 (2001).
[39] Maglio O, Nastri F, Calhoun JR, Lahr S, Wade H, Pavone V, DeGrado WF and Lombardi A, "Artificial di-iron proteins: solution characterization of four helix bundles containing two distinct types of inter-helical loops", J. Biol. Inorganic Chemistry 10: 539-549 (2005).

Table 1 Number of transport proteins according to their class and subclass used for training and validation of TpPred.

Classes & Subclasses	Number of proteins	Training set	Test set	Independent set
1. Channels/pores (S1)	1139	545	157	164
1.A	481	386	95	50
1.B	269	212	57	52
1.C	309	246	63	57
1.E	38	34	4	5
2. Electrochemical potential-driven transporters (S2)	1456	558	148	73
3. Primary active transporters (S3)	2045	896	210	134
3.A	1612	1280	332	67
3.B	22	20	2	3
3.D	370	301	69	61
3.E	27	24	3	3
4. Group translocators (S4)	107	90	17	20
4.A	91	73	18	17
4.C	12	10	2	3
5. Transmembrane electron carriers (S5)	106	81	25	21
5.A	61	50	11	11
5.B	45	35	10	10
8. Accessory factors involved in transport (S6)	129	109	20	26
8.A	94	78	16	17
8.B	35	26	9	9
9. Incompletely characterized transporters (S7)	377	268	75	49
9.A	211	168	43	26
9.B	164	132	32	23

The transport proteins are classified at two levels (TC class, and TC subclass) as indicated by a specific TC number TC I.X. Here I = 1,…..,9 represents each of the 9 TC classes, X = A, B, C, D, E,… represents each of the TC subclasses that belong to a TC class.

Table 2 Hydrophobicity, hydrophilicity and mass of side chain scales for 20 amino acids used in calculating pseudo amino acid composition (PseAA).

Amino acid	Hydrophobicity[a]	Hydrophilicity[b]	Mass of side chain
A	0.62	-0.5	15
C	0.29	-1	47
D	-0.9	3	59
E	-0.74	3	73
F	1.19	-2.5	91
G	0.48	0	1
H	-0.4	-0.5	82
I	1.38	-1.8	57
K	-1.5	3	73
L	1.06	-1.8	57
M	0.64	-1.3	75
N	-0.78	0.2	58
P	0.12	0	42
Q	-0.85	0.2	72
R	-2.53	3	101
S	-0.18	0.3	31
T	-0.05	-0.4	45
V	1.08	-1.5	43
W	0.81	-3.4	130
Y	0.26	-2.3	107

[a]Hydrophobicity values are from reference [27]
[b]Hydrophilicity values are from reference [28]

Table 3 Performance accuracy and validation results of 1st layer of TpPred based on combination of pseudo amino acid composition, amino acid composition and physicochemical properties**.**

Classes of	Train Set	Test Set	Validation of NN1 (%)		
proteins	(%)	(%)	Independent data set	Self consistency test	Jackknife test
Transport	98.2	87.7	70.4	81.6	75.9
Non-transport	96.4	89.1	100.0	94.4	86.9
Average	97.3	88.4	85.2	88.0	81.4

Table 4 Performance accuracy and validation results of 2nd layer of TpPred based on combination of pseudo amino acid composition, amino acid composition and physicochemical properties.

Classes of proteins	Training set (%)	Test set (%)	Validation of NN2 (%)		
			Independent data set	Self consistency test	Jackknife test
1.Channels/pores	95.8	66.9	57.9	59.9	46.0
2.Electrochemical potential-driven transporters	93.9	73.6	84.9	89.9	71.6
3.Primary active transporters	93.1	70.9	64.2	68.2	52.7
4.Group translocators	100.0	76.5	90.0	96.0	81.9
5.Transmembrane electron carriers	100.0	68.0	95.2	97.2	78.6
8.Accesory factors involved in transport	100.0	85.0	84.6	89.6	73.3
9.Incompletely characterized transporters	100.0	84.0	81.6	88.6	75.1
Average	97.5	75.0	79.8	84.2	68.5

Table 5 Performance accuracy and validation results of 3rd layer of TpPred based on combination of pseudo amino acid composition, amino acid composition and physicochemical properties.

Classes and subclasses	Training Set (%)	Test Set (%)	Validation of NNs (%)		
			Independent data set	Self consistency test	Jackknife test
1. Channels / pores (NN3)					
1.A	98.7	86.3	40.0	43.9	41.3
1.B	96.2	86.0	82.7	89.9	76.9
1.C	94.3	85.7	75.4	81.4	72.3
1.E	88.2	75.0	80.0	85.0	77.9
Average	94.4	83.2	69.5	75.0	67.1
3. Primary active transporters (NN4)					
3.A	99.4	90.1	64.2	70.9	67.0
3.B	90.0	100.0	66.7	76.2	61.7
3.D	95.0	89.8	62.3	70.1	59.9
3.E	95.8	100.0	100.0	100.0	85.6
Average	95.1	95.0	73.3	79.3	68.5
4. Group translocators (NN5)					
4.A	100.0	100.0	94.1	97.3	79.5
4.C	100.0	100.0	66.7	76.2	66.5
Average	100.0	100.0	80.4	86.8	73.0
5. Transmembrane electron carriers (NN6)					
5.A	100.0	100.0	90.9	93.6	76.4
5.B	100.0	100.0	100.0	100.0	89.0
Average	100.0	100.0	95.4	96.8	82.7
8. Accessory factors involved in transport (NN7)					
8.A	98.7	100.0	88.2	91.2	75.3
8.B	96.1	100.0	77.8	82.2	77.2
Average	97.4	100.0	83.0	86.7	76.3
9. Incompletely characterized transporters (NN8)					
9.A	100.0	98.0	85.0	90.4	77.8
9.B	100.0	97.6	79.5	85.2	69.6
Average	100.0	97.8	82.2	87.8	73.7

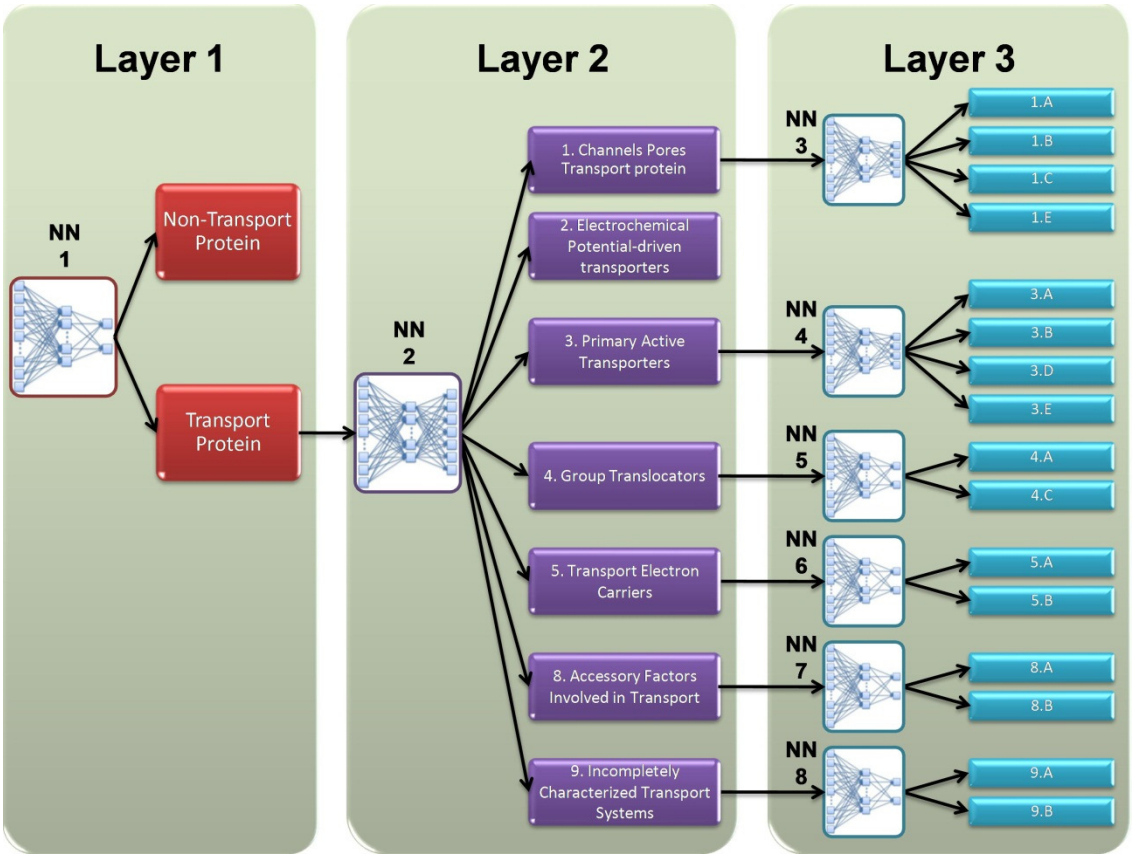

Fig. 1 A schematic drawing to classify transport proteins into their seven main functional classes and subclasses. The notation for different subclasses are: 1.A, α-type channel; 1.B, ß-barrel porins; 1.C, pore-forming toxins; 1.H, holins; 3.A, P-P-bond-hydrolysis-driven transporters; 3.B, decarboxylation-driven transporters; 3.D, oxidoreduction-driven transporters; 3.E, light absorption-driven transporters; 4.A, phosphotransfer-driven group translocators; 4.C, acyl CoA ligase-coupled transporters; 5.A, transmembrane 2-electron transfer carriers; 5.B, transmembrane 1-electron transfer carriers; 8.A, auxiliary transport proteins ; 8.B, protein-peptide toxins targeted to channels and carriers; 9.A, recognized transporters of unknown biochemical mechanism; 9.B, putative transport proteins.