Computational prediction and analysis of human transporters using physicochemical properties of amino acids

Hui-Ling Huang, Ming-Che Li, Tamara Vasylenko, Shinn-Ying Ho

***Abstract*—Membrane transport proteins (named transporters) play key roles in transporting cellular molecules across cells and cellular compartment boundaries, mediating the absorption and removal of unwanted materials from cells, and establishing electrochemical gradients across membranes. A variety of transporters are responsible to absorption, distribution, and excretion of drugs. The immense importance of developing computational prediction methods arises mainly from two aspects: 1) specific transporters have been explored as therapeutic targets, and 2) few effective prediction method exists. This study proposes an optimization method HTPred to identify and analyze human transporters from protein sequences, mainly comprising the four steps: 1) constructing a new dataset (named HITSET) consisting of 5,176 reviewed human transporters and non-transporters, 2) encoding the sequences of human transporters by using physicochemical properties of amino acids in the AAindex database, 3) utilizing an optimal feature selection to identify informative physicochemical properties by maximizing prediction accuracy of a support vector machine based classifier, and 4) ranking and analyzing the identified properties according to prediction contribution to gain insight into human transporters. The blind test accuracies on the datasets without and with reducing sequence identity (<25%) were 84.71% and 74.59% using 18 and 29 physicochemical properties, respectively. The top-one property is the free energy change of alpha (Ri) to alpha (Rh), which shows that the inside/outside preferences of the amino acids in a polypeptide chain which presumably reflect the interactions of the residues with water. Transporters are membrane-spanning molecules that mostly form pore-like structures which interact with water so that their interior side is originally formed from the polar groups.**

***Index Terms*—About four key words or phrases in alphabetical order, separated by commas.**

I. Introduction

Integral membrane proteins, which are embedded into biological membranes, include several functional classes such as receptors, enzymes, and transporters [1]. Membrane transport proteins (simply transporters) are composed by one or more protein subunits. Transporters play key roles in transporting cellular molecules across cells and cellular compartment boundaries, mediating the absorption and removal of unwanted materials from cells, and establishing electrochemical gradients across membranes [2-5]. The immense importance of studying transporters has clinical basis. Specific transporters have been explored as therapeutic targets [6-7]. Moreover, a variety of transporters are responsible to absorption, distribution, and excretion of drugs, as the electro-chemical gradients which establish across membranes influence drug partitioning into and out of cells and cell organelles [8, 9]. Thus, computational prediction of transporters is vital for facilitating functional study of genomes and searching for new therapeutic targets and pharmacologically relevant transporters, and would help in the design of novel anti-microbial drugs [10].

The human genome, with an estimated total of 35,000 genes, contains numerous gene families that encode membrane transporters. Currently, our knowledge of the relevant human transporters is limited due to the very nature of membrane-bound proteins, as they are technically demanding to be crystallized for structural characterization by spectroscopic methods. That being the case, a bioinformatics approach offers an attractive alternative [2].

Several databases have already been built as comprehensive repositories of transporter sequences for the purpose of helping computational biologists to develop and test their prediction methods. In a context of drug discovery and development the most immediately accessible database is Human ABC-Transporter Database which gives key information on expression, function and substrate for ABC family members only [11]. On the other hand, web-accessible relational Human membrane transporter database (HMTD) is a good resource to identify other transporter families which plays a role in drug absorption, distribution, metabolism and excretion. It performs indexing of transporters in a number of ways and provides information on sequence variants, altered functions caused by polymorphisms/mutations, and the (patho) physiological role and associated disease [12]. It can be queried to list all the known transporters in a given tissue or the tissue distribution of a given transporter.

The most widely adopted TC-classification system of all transport proteins is provided by Transport Classification Database (TCDB) [10]. It contains comprehensive information on experimentally-characterized transporters which are organized within a simple tree structure based on both function and homology, and contains over 550 transporter families. However, the sequences are from various organisms and human sequences are not available for all transporters [1]. Employing homology searches (ex., BLAST) against experimentally-determined transporters in the TCDB, a putative transporter database named TransportDB was constructed for hundreds of completely sequenced genomes [13, 14]. However, some categories of transporters were manually excluded here. The comparison of above-mentioned databases (shown in Table 1) shows that TCDB has the largest number of entries. TransportDB contains computationally predicted sequences, which are not all experimentally verified. TCDB provides widely adopted TC-classification system for all transporters. TCDB and its associated tools, such as SSEARCH, have been widely used to predict and classify putative transporters. Lin et al. explored a support vector machine (SVM) method for transporter proteins prediction using training set, composed of TCDB sequences and their homologs searched by BLAST against SWISS-PORT database [15]. Li et al. used training data from TCDB to propose a Nearest Neighbor approach which integrate homology and motif search methods in order to discover transporter families [16]. Vagner et al. used BLASTP to analyze predicted plant transporters in Medicago against TCDB database [17].

Table 1: The comparison of existing Transporters Databases

Feature	TCDB	TransportDB	HMTD	ABC-Transporter Database	HITSET
Total Number of entries	~ 5,600	~ 4,650	~ 250	~ 45	5,176
Reviewed entries	+ (reviewed)	+/- (predicted)	+ (reviewed)	+ (reviewed)	+ (reviewed)
TC-system	+	+/-	-	-	+/-
Human\All organisms database	All organisms	All organisms	Human	Human	Human
Gene description	-	+	+	+	-

+/- in “Reviewed entries” stands for the databases of experimentally proved transporter proteins along with the proteins with computationally predicted annotations; +/- in “TC-system” stands for the databases, which have classification another from the TC-system, but some classes remain the same.

To facilitate the study of transporters, we have constructed a human integral transporter dataset (named HITSET). HITSET creates a view for all human proteins containing membrane-spanning regions, deposited in the SwissProt database. All entries of HITSET were extracted from Swiss-prot using keyword “reviewed”. HITSET and TransportDB follow only main classes of TC-system. The ABC-Transporter Database and HITSET both include only human protein subunits. HITSET has the largest number of reviewed human sequences.

These proteins are clustered into families. We identified 5,176 human transmembrane proteins and divided them into family-belonged and orphan groups according to general annotation of UniProtKB/Swiss-Prot. Family-belonged groups had 916 families, totally 4,030 proteins and orphan groups had 1,146 proteins. 728 reviewed transporters were arranged into six major functional classes: Human Alpha-type channels, Human Beta-type porins, Human Toxins, Human Secondary active transporters, Human Primary active transporters, and Human Unclassified transporters. HITSET also contains 18 proteins from six specific families. Our dataset does not include auxiliary transport proteins that modulate the activity of other transporters rather than performing the transport themselves.

The availability of HITSET will allow further development of computational methods for novel human whole-protein and 3D-structure transporter predictions, as well as identify candidates for further experimental investigation [18, 19]. With the advancement of research about transporters and the increase of our knowledge, more data will be added to the database. The progression of research may also enable us to include the secondary and tertiary structures (as more membrane transport proteins are crystallized) of membrane transporter genes in our database.

This study proposes an optimization method HTPred to identify and analyze human transporters in HITSET from protein sequences. HTPred utilizes SVM and informative physicochemical properties selected by maximizing prediction accuracy of the SVM-based classifier. The blind test accuracies on the datasets without and with reducing sequence identity (<25%) were 84.71% and 74.59%, respectively, using 18 identified physicochemical properties. The 18 physicochemical properties were further analyzed to gain insight into human transporters.

The style will adjust your fonts and line spacing. **Do not change the font sizes or line spacing to squeeze more text into a limited number of pages.** Use italics for emphasis; do not underline.

II. Materials and Methods

A. HITSET dataset

The procedure for construction of the HITSET dataset is given in the following six steps.

Step 1: Obtaining human transmembrane transporter proteins.

We collected 5,176 proteins by using the keyword “transmembrane” with the query – reviewed:yes AND organism: "Homo sapiens (Human) [9606] in UniProtKB/Swiss-Prot (version 2011 11, www.expasy.org).

Step 2: Dividing into family-belonged and orphan groups.

We categorized the 5,176 human transmembrane transporters into 916 families (4,030 proteins) and orphan groups (1,146 proteins) according to their sequence similarities and general annotations of each protein in UniprotKB/Swiss-Prot.

Step 3: Distinguishing the confirmed, potential and non-transporter proteins among orphan groups.

Orphan group proteins were manually checked on four main UniprotKB/Swiss-Prot annotations: protein names, gene names, function (general annotation) and sequence similarities (general annotation). If these annotations contained transporter-related names such as channel, pore, leak, porin, facilitator, gap junction, transporter, porter, uniporter, cotransporter, symporter, exchanger, antiporter, exporter, importer, carrier, shuttle, pump, translocator, translocon, permease, translocase and so on, or transport actions related vocabularies such as release, facilitate, transport, cotransport, symport, exchange, antiport, uptake, efflux, influx, import, export, intake, outlet, translocate, partition, extrusion, intrusion, accumulate, diffusion and so on, corresponding proteins were divided into confirmed and potential transporter proteins. We used PubMed to search for the related works and curated literatures on confirmed transporters. The remaining proteins were classified into non-transporters.

Step 4: Distinguishing the confirmed, potential and non-transporter proteins among family- belonged groups.

Family-belonged transporters were analyzed in the same way, as described in Step 3. According to Swiss-Port annotations, they were divided into confirmed, potential and non-transporters.

Step 5: establishing HITSET

There were totally 9 superfamilies, 916 families and 4030 transmembrane transporters in human transmembrane transporter dataset. These 4030 family-belonged proteins were composed of 713 confirmed, 170 potential and 3147 non-transporter proteins. Totally 358 literatures were curated. The orphan human transmembrane transporters included 15 confirmed, 20 potential and 1111 non-transporter proteins. Finally, after combining previous family-belonged and orphan groups, HITSET contained 5176 human transmembrane transporters, 9 superfamilies, 916 families, 728 confirmed, 190 potential, and 4258 non-transporter proteins. The numbers of superfamilies, families, proteins, literatures, human transporters, potential human transporters and non-human-transporters in previous five steps are shown in Table 2.

Table 2. The statistics of previous five steps in HITSET construction

steps	Superfamily number	Family number	Protein number	Literature number	HTS	PHTS	NHTS
1	-	-	5176	-	-	-	-
2	9	916	5176	-	-	-	-
3	6	0	1146	21	15	20	1111
4	8	916	4030	362	713	170	3147
5	9	916	5176	379	728	190	4258

HTS: human transporter; PHTS: potential human transporter ; NHTS: non-human-transporter.

Step 6: Designing “HT” (“Human Transporter”) code.

We referred to TC-system [10] to classify human transmembrane transporter subunits into six main classes. The classification criterion of TC-system is based on the mode of transport, energy- coupling mechanisms and transmembrane structure of transporters. The definition of six main classes and corresponding Human Transporter code (HT code) are given below:

1. Human Transporter Alpha-type channels (HTA)

The transmembrane regions of this type of transporters are mainly composed of alpha helixes, and the transport processes of these transporters are usually energy independent and do not require carriers to regulate the transport.

2. Human Transporter Beta-type porins (HTB)

In this type of human transporter subunits, the transmembrane region is composed exclusively of β-strands, and form β-barrels in general. The transport processes are usually energy independent and do not require the regulation of carriers as HTA.

3. Human Transporter Toxins (HTT)

Transporter toxins (PFT) are synthesized and secreted by one cell; it could insert itself into the membrane of a target cell and form a pore which the target cell usually cannot handle it. HTT is energy independent but doesn't require the regulation of carriers for transport.

4. Human Transporter Secondary active transporters (HTS)

HTS is energy independent and require the regulation of carriers for transport.

5. Human Transporter Primary active transporters (HTP)

HTP is energy dependent and require the regulation of carriers for transport.

6. Human Transporter Unclassified transporters (HTU)

Human transporters which were not classified into previous five main classes were assigned to this class. If the latest researches will prove that a HTU member can be classified to one of the previous five classes, then this code would be discarded and changed into a new one of the updated classes for it. The numbers of superfamilies, families and proteins in step 6 are shown in Table 3.

Table 3. The statistics of step 6 in HITSET construction

HT code	Superfamily number	Family number	Protein number	HTS number	PHTS number
HTA	0	44	334	288	46
HTB	0	1	3	3	0
HTT	0	2	3	3	0
HTS	3	68	448	334	114
HTP	1	18	100	85	15
HTU	0	9	30	15	15
Total	4	142	918	728	190

Human Transporter Alpha-type channels (HTA); Human Transporter Beta-type porins (HTB); Human Transporter Toxins (HTT); Human Transporter Secondary active transporters (HTS); Human Transporter Primary active transporters (HTP); Human Transporter Unclassified transporters (HTU).

The confirmed and potential transporters in HITSET were classified as follows: HTWX1X2Y1Y2Z1Z2Z3. Here, “HT” is an abbreviation for “Human Transporter”; letter “W” corresponds to one of six main transporter classes: A (Alpha-type channels), B (Beta-type porins), T (Toxins), S (Secondary active transporters), P (Primary active transporters), and U (Unclassified transporters). Positions of “X1X2”, “Y1Y2” and “Z1Z2Z3” correspond to superfamily, family and member, respectively. HT codes were specialized for confirmed and potential transporter subunits, not for non-transporter subunits. Potential transporters were highlighted by adding “P” (Potential) at the end of HT code, so that HT code became “HTWX1X2Y1Y2Z1Z2Z3P”. If a potential transporter subunit will be proved as an actual transporter subunit in the future, the “P” in the end of the HT code would be deleted; inversely, if a potential transporter subunit were proved as a non-transporter subunit, the HT code of this subunit would be discarded.

To develop a powerful statistical predictor, we divided HITSET into training (1A, 2A) and independent test (1B, 2B) subsets. HITSET subset 1 contains 728 human transporters, 190 human potential transporters and 4258 non-transporters, totally 5176 human transmembrane proteins. In order to balance the transporter and non-transporter of subset 1, we randomly selected 728 proteins from 4258 non-transporters, so the final training dataset (HITSET subset 1A) of HITSET contains 728 transporters and 728 non-transporters, as shown in Table 4. The remaining 3530 human non-transporters and 190 human potential transporters were used to construct an independent test set (HITSET subset 1B), as shown in Table 4.

To reduce the sequence redundancy of HITSET, we used USEARCH with identity threshold set to 25%. As a result we got 3250 human transporters. Among these 3250 transporters, 366 were transporters, 144 were potential transporters, and 2740 were non-transporters. Then we randomly selected 366 of the 2740 non-transporters to balance the dataset, and the final subset 2A contains 366 transporter and 366 non-transporters, as shown in Table 5. The remaining 2374 human non-transporters and 144 human potential transporters were used to construct an extra independent test set (subset 2B), as shown in Table 5.

Table 4. The sequence numbers of HITSET subset 1 in various stages

Classes	Total	Subset 1A	Subset 1B
Transporter	728	728	0
Potential transporter	190	0	190
Non-human-transporter	4258	728	3530

Table 5. The sequence numbers of HITSET subset 2 in various stages

Dataset	Total	25%identity	3 classes	Total	Subset 2A	Subset 2B
HITSET	5176	3250	HTS	366	366	0
			PHTS	144	0	144
			NHTS	2740	366	2374

B. *Sequence representation using physicochemical properties*

AAindex is a database developed by Kanehisa et al., which collects numerical indices representing physicochemical and biochemical properties of amino acids [20]. By removing the properties with an amino acid value ‘NA’, the number of properties in AAindex 9.0 is reduced from 544 to 531. The 531 properties were used as initial features to construct an SVM classifier for the discrimination between cancerlectins and non-cancerlectins. The original sequences in the dataset HITSET were transformed to the numerical indices according to the corresponding values of amino acids of each feature. To calculate a feature vector value of a protein, the feature values of every amino acid in a protein sequence were summed up, and divided by the sequence length of the protein. In this way, every protein was represented as a 531 feature vector for later machine learning classification. The average values of each physicochemical property form a feature vector of the protein sequence. These values were normalized to the scale between -1 and 1 for using SVM.

C. *Feature selection by using IBCGA*

An efficient inheritable bi-objective combinatorial genetic algorithm IBCGA [21] based on an intelligent genetic algorithm IGA [22] is utilized to solve the feature selection problem while maximizing prediction accuracy. IGA based on orthogonal experimental design uses a divide-and-conquer strategy and a systematic reasoning method instead of the conventional generate-and-go method to efficiently solve the combinatorial optimization problem C(n, m) having a huge search space of size n!/(m!(n-m)!)), where n=531 in this study. IBCGA can efficiently search the space of C(n, r±1) by inheriting a good solution in the space of C(n, r) [14]. Therefore, IBCGA can economically obtain a complete set of high-quality solutions in a single run.

The normalized protein sequences of the training data sets were the input for SVM. Fitness function is the only guide for genetic algorithms to obtain desirable solutions. The fitness function of IBCGA is the overall accuracy five-fold cross-validation (5-CV). IBCGA with the fitness function f(X) can simultaneously obtain a set of solutions, Xr, where r=rstart, rstart+1, …, rend in a single run. The algorithm of IBCGA with the given values rstart and rend is described as follows:

Step1. (Initiation) Randomly generate an initial population of Npop individuals. All the n binary GA-genes have r 1’s and n-r 0’s where r = rstart.

Step2. (Evaluation) Evaluate the fitness values of all individuals using f(X).

Step3. (Selection) Use the traditional tournament selection that selects the winner from two randomly selected individuals to form a mating pool.

Step4. (Crossover) Select pc•Npop parents from the mating pool to perform orthogonal array crossover on the selected pairs of parents where pc is the crossover probability.

Step5. (Mutation) Apply the swap mutation operator to the randomly selected pm•Npop individuals in the new population where pm is the mutation probability. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.

Step6. (Termination test) If the stopping condition for obtaining the solution Xr is satisfied, output the best individual as Xr. Otherwise, go to Step 2. In this study, the stopping condition is to perform 40 generations.

Step7. (Inheritance) If r < rend, randomly change one bit in the binary GA-genes for each individual from 0 to 1; increase the number r by one, and go to Step 2. Otherwise, stop the algorithm.

D. *Prediction Method HTPred*

The selected m physicochemical properties and the associated parameter set of SVM by using HTPred are used to implement the computational system and analyze the physicochemical properties to further understand the human transporters. Since the HTPred is a non-deterministic method, it should make more effort to identify an efficient and robust feature set of informative physicochemical properties in five aspects. The procedure is as the following steps:

Step 1 : We prepare the independent data sets where each set is used as the training data set of 5-CV.

Step 2 : HTPred is performed R independent runs for each of independent data sets. In this study, R = 30. There are total 30 sets of m physicochemical properties for each of independent data sets.

Step 3 : Choose the set of selected physicochemical properties with a maximal accuracy.

HTPred will automatically determine a set of informative physicochemical properties and an SVM-model for predicting human transporters and non-transporters. The prediction performances were evaluated in terms of the test accuracy, Mathew's correlation coefficient (MCC), Specificity and Sensitivity.

III. Results and Discussion

If you are using *Word*, use either the Microsoft Equation Editor or the *MathType* add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). "Float over text" should *not* be selected.

A. *Identified properties by IBCGA*

The statistical result of IBCGA in selecting property sets from R = 30 independent runs on subsets 1A and 2A is given in Fig. 1. The highest scores were obtained on the 20th run for subset 1A and 20th run for subset 2A. These runs were selected and their prediction accuracies for different numbers of features are given in Fig. 2. The best property numbers for the 20th run of subset 1A and 20th run of subset 2A were m=37 and 32, respectively, with accuracies of 86.63% and 88.52%. The feature numbers and corresponding accuracies are shown in Fig. 2. However, the 26th independent run of subset 1A and the 29th independent run of subset 2A showed the most "stable solution" of all runs (shown in Fig. 3). The training accuracies on the subsets 1A and 2A were 86.42% and 86.68% with the feature numbers of m=18 (shown in Table 7) and 29, respectively.

Based on the result of 30 independent runs of subset 1A's and subset 2A's training sets, we found that the accuracies were similar to each other. In this case, we decided to use the most stable solutions for subsets 1A and 2A to establish human transporter prediction models, HTPred_a and HTPred_b, respectively.

Table 6. The prediction accuracies (%) on the datasets without (subset 1) and with reducing sequence identity 25% (subset 2).

	Training Accuracy	Test Accuracy	Sensitivity	Specificity	MCC
Subset 1	86.42	84.71	82.64	86.78	0.69
Subset 2	86.68	74.59	72.13	77.05	0.49

Table 7. The 18 properties of the 26th independent run on the training subset 1A

Feature ID	AAindex identity	Description
20	BURA740102	Normalized frequency of extended structure (Burgess et al., 1974)
54	CIDH920101	Normalized hydrophobicity scales for alpha-proteins (Cid et al., 1992)
65	DAYM780201	Relative mutability (Dayhoff et al., 1978b)
94	FINA910103	Helix termination parameter at posision j-2,j-1,j (Finkelstein et al., 1991)
105	GEIM800109	Aperiodic indices for alpha-proteins (Geisow-Roberts, 1980)
133	JOND750102	pK (-COOH) (Jones, 1975)
171	MAXF760101	Normalized frequency of alpha-helix (Maxfield-Scheraga, 1976)
174	MAXF760104	Normalized frequency of left-handed alpha-helix (Maxfield-Scheraga, 1976)
207	NAKH920106	AA composition of CYT of multi-spanning proteins (Nakashima-Nishikawa, 1992)
220	OOBM850103	Optimized transfer energy parameter (Oobatake et al., 1985)
236	PALJ810114	Normalized frequency of turn in all-beta class (Palau et al., 1981)
310	RACS820111	Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga, 1982)
357	ROSM880103	Loss of Side chain hydropathy by helix formation (Roseman, 1988)
364	SUEM840102	Zimm-Bragg parameter sigma x 1.00E+04 (Sueki et al., 1984)
379	VELV850101	Electron-ion interaction potential (Veljkovic et al., 1985)
386	WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)
388	WOEC730101	Polar requirement (Woese, 1973)
506	DIGM050101	Hydrostatic pressure asymmetry index, PAI (Di Giulio, 2005)

B. *Prediction performance evaluation*

The effectiveness of the identified informative physicochemical and biochemical properties was evaluated by implementing HTPred_a and HTPred_b predictors. The independent test accuracies of HTPred_a and HTPred_b on subsets 1B and 2B were 84.71% and 74.59%, respectively. The analysis focused on the identified properties obtained from HTPred_a.

For the 3530 sequences of the subset 1B, HTPred_a predicted 2931 non-human-transporters and 599 human transporters with the accuracy of 83.03%. Among 190 potential human transporters, 128 sequences were predicted as human transporters and 62 as non-human-transporters. Then we used PubMed to search for the related works of 190 potential human transporters and curated the related works of human transporters. We found out that 14 proteins were experimentally identified lately as human transporters. Among the above-mentioned proteins, 10 sequences were predicted as human transporters and 4 sequences were predicted as non-human-transporters by HTPred_a. The protein name, gene name, HT code, predicted result and related reference are shown in Table 8. The experimental results reveal that the identified set of physicochemical properties is promising for the transporter prediction.

Table 8. Experimentally identified human transporters and the prediction result of HTPred.

Protein name	Gene	HT code	Result	Reference
Piezo-type mechanosensitive ion channel component 1	PIEZO1	HTA0044001P	0	Coste *et al.* [23]
Piezo-type mechanosensitive ion channel component 2	PIEZO2	HTA0044002P	1	Coste *et al.* [23]
Zinc transporter 10	SLC30A10	HTS0006006P	0	Tuschl *et al.* [24] Quadri *et al.* [25]
Putative small intestine	SLC17A4	HTS0107008P	1	Togawa *et al.* [26]

sodium-dependent phosphate transport protein				
Probable cationic amino acid transporter	SLC7A14	HTS0201001P	1	Jaenecke *et al.* [27]
Choline transporter-like protein 5	SLC44A5	HTS0010001P	1	Sugimoto *et al.* [28]
Anoctamin-10	Ano10	HTA0002010P	1	Tian *et al.* [29]
Anoctamin-7	Ano7	HTA0002007P	0	Tian *et al.* [29]
Anoctamin-4	Ano4	HTA0002004P	1	Tian *et al.* [29]
Anoctamin-6	Ano6	HTA0002006P	1	Tian *et al.* [29]
Anoctamin-8	Ano8	HTA0002008P	0	Tian *et al.* [29]
Anoctamin-9	Ano9	HTA0002009P	1	Tian *et al.* [29]
Anoctamin-5	Ano5	HTA0002005P	1	Tian *et al.* [29]
Anoctamin-3	Ano3	HTA0002003P	1	Tian *et al.* [29]

C. Main effect difference analysis

The HTPred_a classifier distinguishes transporters from non-transporters. The main effect difference (MED) [22] was used to estimate effects of individual features. The principle of the MED analysis is to calculate MED scores according to the prediction effectiveness. The most effective property has the largest value of MED. Due to the function diversities of transporters, we mainly focused on the top-ranked features which have high MED scores. The 18 properties ranked by using MED obtained from HTPred_a on the training feature set are shown in Fig. 4.

D. Relationship between transporters and selected physicochemical properties

Some typical properties in the set of the selected 18 features in the AAindex database are discussed below. The property WERD780103, described as "free energy change of alpha (Ri) to alpha (Rh)" is the most meaningful feature as it shows the inside/outside preferences of the amino acids in a polypeptide chain which presumably reflect the interactions of the residues with water [30]. Transporters are membrane-spanning molecules that mostly form pore-like structures which interact with water so that their interior side is originally formed from the polar groups. Just as in the molecule of acetylcholine receptor (shown in Fig. 5), residues surrounding a pore which is shown with a red axe might be abundant with polar groups. In this case, the conformational preferences of the residues in a protein could reveal whether or not it is wholly or partially in contact with solution and thus, show its membrane connection. The DIGM050101 property, described as "hydrostatic pressure asymmetry index", or PAI gives values to individual amino acids that are positively correlated to the polarity as well. Giulio et al. [31] showed that on average, the more polar amino acids possess a higher PAI value, that is to say they are more barophilic. In this respect, barophily contributes to the distinction of protein structures, which are membrane-spanning.

The NAKH920106 feature, named "AA composition of CYT of multi-spanning proteins" expands the list of features by pointing out the differences in amino acid composition between cytoplasmic (CYT) and extracellular (EXT) protein domains [32]. Thus, transporters possessing transmembrane regions, have both CYT and EXT domains presented and the difference (CYT – EXT) becomes a meaningful characteristic.

The OOBM850103 feature, which corresponds to the "optimized transfer energy parameter", investigates tertiary structure of a protein [33]. This information might be important if several subunits are forming a pore-like structure in a membrane as 5 subunits of acetylcholine receptor (Fig. 5).

The relationship between protein sequence and structure in integral membrane associated transporter proteins is emphasized by the frequencies of amino acid exchanges in their transmembrane segments. As pointed out by Jones D.T. et al. [34], the transmembrane protein mutation data matrix is quite different from the matrix calculated from a general sequence set. Consequently, relative mutability feature (DAYM780201) contains valuable information for classification task [35]. The FINA910103 feature, described as "Helix termination parameter at position j-2, j-1, j" characterizes the α-helical contents of transporter proteins and can serve as a distinguishing feature as long as transporters posses α-helical regions most of which have defined localizations [36].

Two features found to be presented in the subsets 1A and 1B. These are MAXF760101 and CIDH920101, described as "Normalized frequency of alpha-helix" and "Normalized hydrophobicity scales for alpha-proteins", respectively. Normalized frequency of alpha-helix is a conformational information collected from each single residue in a peptide chain to represent the whole backbone conformation [37]. It is significant in α-helics recognition in transporters. The normalized hydrophobicity scales for alpha-proteins feature reflect the predominance of alpha-helixes in the whole set of transmembrane transporters, thus alpha class proteins hydrophobicity scale can represent transporter proteins in our model [38].

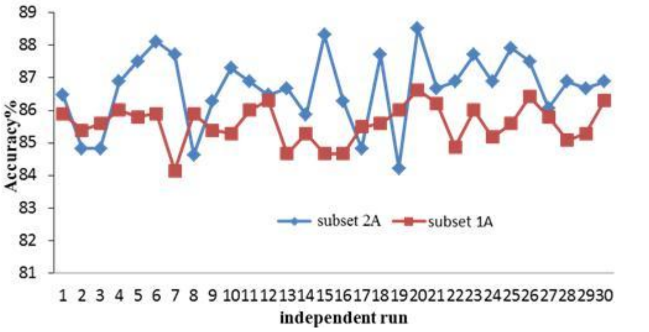

Figure 1. Accuracies of training subsets 1A and 2A for 30 independent runs.

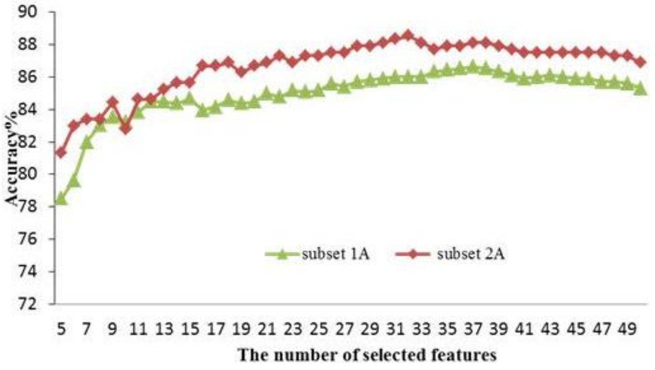

Figure 2. The number of selected features and accuracies of subsets 1A and 2A at 20th independent run

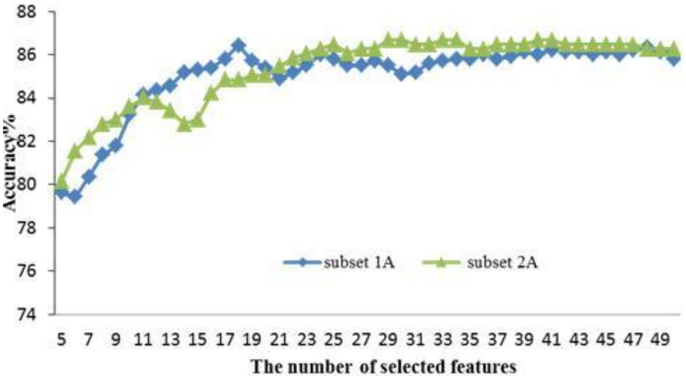

Figure 3. The number of selected features and accuracies of subsets 1A and 2A at 26th and 29th independent runs, respectively.

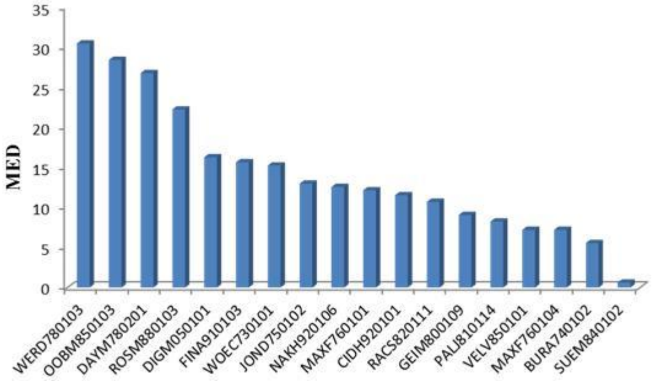

Figure 4. The histogram of 18 properties with MED scores obtained from HTPred_a on training subset 1A.

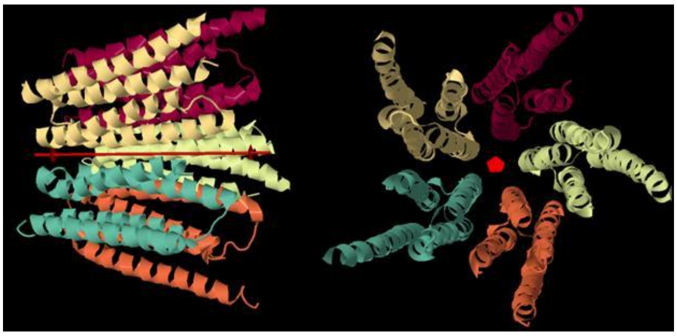

Figure 5. The 3D Structure of acetylcholine receptor pore. The red axes are going through the pore of the transporter. Five subunits, forming the pore are shown in different colors.

IV. Conclusions

Membrane transport proteins (named transporters) play key roles in 1) transporting cellular molecules across cells and cellular compartment boundaries, 2) mediating the absorption and removal of unwanted materials from cells, and 3) establishing electrochemical gradients across membranes. A variety of transporters are responsible to absorption, distribution, and excretion of drugs. Therefore, it is desirable to develop prediction methods for discovery of transporters and their functions. This study has established a dataset of reviewed human transporters and non-transporters, named HITSET. Consequently, we have proposed an optimization method HTPred to identify and analyze human transporters from protein sequences based on the SVM classifier.

We used the physicochemical properties of amino acids in the AAindex database and the feature selection algorithm IBCGA to select features which are important to identify whether a protein is a transporter or not. We ranked and analyzed the identified properties according to prediction contribution to gain insight into human transporters. The results suggested that free energy change, transfer energy parameter, mutability and other secondary structure properties of residues play important roles in the transportation function.

Acknowledgment

This work was funded by National Science Council of Taiwan under the contract number NSC-102-2221-E-009-16 9-, and "Center for Bioinformatics Research of Aiming for the Top University Program" of the National Chiao Tung University and Ministry of Education, Taiwan, R.O.C. for the project 102W962. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Almén, M.S., Nordström, K., Fredriksson, R. and Schiöth, H.B. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biology. 2009; 7:50.

[2] Lee, V.H. Membrane transporters. Eur J Pharm Sci. 2000; 11:S41–50.

[3] Hediger, MA. Structure, function and evolution of solute transporters in prokaryotes and eukaryotes. J Exp Biol. 1994; 196:15–49.

[4] Borst P., Elferink R.O. Mammalian ABC transporters in health and disease. Annu Rev Biochem. 2002; 71:537–592.

[5] Seal, R.P., Amara S.G. Excitatory amino acid transporters: a family in flux. Annu Rev Pharmacol Toxicol. 1999; 39:431–456.

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

[7] Birch, P.J., Dekker, L.V., James, I.F., Southan, A., Cronk, D.. Strategies to identify ion channel modulators: current and novel approaches to target neuropathic pain. Drug Discov Today. 2004; 9:410–418.

[8] Larsen, A.K., Escargueil, A.E., Skladanowski, A.. Resistance mechanisms associated with altered intracellular distribution of anticancer agents. Pharmacol Ther. 2000; 85:217–29.

[9] Lee, W., Kim, R.B.. Transporters and renal drug elimination. Annu Rev Pharmacol Toxicol. 2004; 44:137–166.

[10] Saier, M.H., Tran, C.V., Barabote, R.D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res. 2005; 34:181-186.

[11] Barnes, M.R. Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data, 2nd Edition. ISBN: 978-0-470-02619-9.

[12] Yan, Q. and Sadée, W.. Human membrane transporter database: A web-accessible relational database for drug transport studies and pharmacogenomics. AAPS PharmSci. 2000; 2(3): 11–17.

[13] Ren, Q., Kang, K., Paulsen, I.. TransportDB: a relational database of cellular membrane transport systems. Nucleic Acids Res. 2004; 32: D284–D288.

[14] Ren, Q., Chen, K., Paulsen, I.. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res. 2007; 35: D274–D279.

[15] Lin, H.H., Han, L.Y., Cai, C.Z., Ji, Z.L., Chen, Y.Z.. Prediction of Transporter Family From Protein Sequence by Support Vector Machine Approach. PROTEINS: Structure, Function, and Bioinformatics. 2006; 62: 218–231.

[16] Li, H.Q., Dai, X.B., Zhao, X.Ch.. A nearest neighbor approach for automated transporter prediction and categorization from protein sequences. Bioinformatics. 2008; 24: 1129–1136.

[17] Benedito, V.A., Li, H.Q., Dai, X.B., Wandrey, M,. Ji, He, Kaundal, R., Torres-Jerez, I., Gomez, S.K., Harrison, M.J, Tang, Y.H., Zhao, P.X.,

Udvardi, M.K.. Genomic Inventory and Transcriptional Analysis of Medicago truncatula Transporters. Plant Physiology. 2010; 152: 1716–1730.

[18] Yan, N.. Structural advances for the major facilitator superfamily (MFS) transporters. Volume 38, Issue 3, March 2013, Pages 151–159

[19] Bai, H., Euring, D., Volmer, K.,Janz, D., Polle, A.. The Nitrate Transporter (NRT) Gene Family in Poplar. PLoS ONE. 01/2013; 8(8):e72126

[20] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M.. AAindex: amino acid index database. progress report 2008. Nucleic Acids Res 2008; 36 (Database issue):D202-205.

[21] Ho, S.-Y. et al.,"Inheritable genetic algorithm for bi-objective 0/1 combinatorial optimization problems and its applications," IEEE Trans. Syst. Man Cybern. Part B-Cybern., vol. 34, pp. 609-620, 2004a.

[22] Ho, S.Y., Shu, L.S., Chen, J.H.: Intelligent evolutionary algorithms for large parameter optimization problems. IEEE Trans Evolut Comput 2004; 8(6):522-541.

[23] Coste, B., Xiao, B., Santos, J.S., Syeda, R., Grandl, J., Spencer, K.S., Kim, S.E., Schmidt, M., Mathur, J., Dubin, A.E. et al. (2012) Piezo proteins are pore-forming subunits of mechanically activated channels. Nature, 483; 176-181.

[24] Tuschl, K., Clayton, P.T., Gospe, S.M., Jr., Gulab, S., Ibrahim, S., Singhi, P., Aulakh, R., Ribeiro, R.T., Barsottini, O.G., Zaki, M.S. et al.. Syndrome of hepatic cirrhosis, dystonia, polycythemia, and hypermanganesemia caused by mutations in SLC30A10, a manganese transporter in man. American journal of human genetics, 2012;90, 457-466.

[25] Quadri, M., Federico, A., Zhao, T., Breedveld, G.J., Battisti, C., Delnooz, C., Severijnen, L.A., Di Toro Mammarella, L., Mignarri, A., Monti, L. et al. Mutations in SLC30A10 cause parkinsonism and dystonia with hypermanganesemia, polycythemia, and chronic liver disease. American journal of human genetics, 2012; 90, 467-477.

[26] Togawa, N., Miyaji, T., Izawa, S., Omote, H. and Moriyama, Y.. A Na+-phosphate cotransporter homologue (SLC17A4 protein) is an intestinal organic anion exporter. American journal of physiology. Cell physiology, 2012; 302, C1652-1660.

[27] Jaenecke, I., Boissel, J.P., Lemke, M., Rupp, J., Gasnier, B. and Closs, E.I.. A Chimera Carrying the Functional Domain of the Orphan Protein SLC7A14 in the Backbone of SLC7A2 Mediates Trans-stimulated Arginine Transport. The Journal of biological chemistry, 2012; 287, 30853-30860.

[28] Sugimoto, M., Watanabe, T. and Sugimoto, Y.. The molecular effects of a polymorphism in the 5'UTR of solute carrier family 44, member 5 that is associated with birth weight in Holsteins, PloS one, 2012; 7, e41267.

[29] Tian, Y., Schreiber, R. and Kunzelmann, K.. Anoctamins are a family of Ca2+ activated Cl- channels, Journal of cell science, 2012.

[30] Wertz, D.H., Scheraga, H.A. Influence of Water on Protein Structure. An Analysis of the Preferences of Amino Acid Residues for the Inside or Outside and for Specific Conformations in a Protein Molecule. Macromolecules, 1978; 11 (1): 9–15.

[31] Giulio, M.D.. A comparison of proteins from Pyrococcus furiosus and Pyrococcus abyssi: barophily in the physicochemical properties of amino acids and in the genetic code. Gene, 2005, 346: 1-6.

[32] Nakashima, h., Nishikawa, K. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS Lett, 1992; 303: 141-146.

[33] Oobatake, M., Kubota, Y. and Ooi, T. Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteuins. Bull. Inst. Chem. Res., Kyoto Univ, 1985; 63: 82-94.

[34] Jones D.T., Taylor W.R., Thornton J.M. A mutation data matrix for transmembrane proteins. 1994. FEBS Letters, 339: 269-275.

[35] Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure, 1978; Vol.5, Suppl.3: 345-352.

[36] Finkelstein, A.V., Badretdinov, A.Y. and Ptitsyn, O.B. Physical reasons for secondary structure stability: alpha-helices in short peptides. Proteins, 1991; 10: 287-299.

[37] Maxfield, F.R. and Scheraga, H.A. Status of empirical methods for the prediction of protein backbone topography. Biochemistry 1976; 15: 5138-5153.

[38] Cid, H., Bunster, M., Canales, M., Gazitúa, F. Hydrophobicity and structural classes in proteins. Protein Eng., 1992; Jul, 5(5):373-5.

Hui-Ling Huang was born in Taiwan in 1968. She received the M.S. and Ph.D. degrees in computer science and information engineering from Feng Chia University, Taichung, Taiwan, in 1998, and 2002, respectively. She is currently an associate professor in the Department of Biological Science and Technology, and Institute of Bioinformatics and Systems Biology, at National Chiao Tung University, Hsinchu, Taiwan. Her research interests include evolutionary algorithms, system optimization, bioinformatics, computational biology, bioimage informatics, neuroscience, medical informatics, and biomedical engineering.

Ming-Che Li was born in Taiwan, R.O.C. He received the M.S. degree in the Institute of Bioinformatics and Systems Biology from National Chiao Tung University, Hsinchu, Taiwan, in 2012. His research interests include computational biology and bioinformatics.

Tamara Vasylenko was born in Ukraine. She is currently a M.S. student in the Institute of Bioinformatics and Systems Biology, at National Chiao Tung University, Hsinchu, Taiwan. Her research interests include computational biology and bioinformatics.

Shinn-Ying Ho was born in Taiwan in 1962. He received the B.S., M.S., and Ph.D. degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1984, 1986, and 1992, respectively. From 1992 to 2004, he was with the Department of Information Engineering and Computer Science at Feng Chia University, Taichung, Taiwan. He is currently a professor in the Department of Biological Science and Technology, and Institute of Bioinformatics and Systems Biology, and the director of Institute of Bioinformatics and Systems Biology at National Chiao Tung University, Taiwan. His research interests include evolutionary algorithms, image processing, pattern recognition, data mining, computer vision, fuzzy classifier, large parameter optimization problems, and system optimization. From 2004, the major research fields are bio-inspired optimization methodologies, bioinformatics, computational biology, bioimage informatics, neuroscience, medical informatics, synthetic biology, biomedical engineering, and genetic engineering. Prof. Ho is the editorial board member and associate editor of several international peer-reviewed journals. He has produced more than 50 papers in peer-reviewed journals as a corresponding author.