Current Topics in Medicinal Chemistry. 2013; 13(4): 1681-1691

Kernel-based feature selection techniques for transport proteins based on star graph topological indices

Carlos Fernandez-Lozano[1], Marcos Gestal[1], Nieves Pedreira-Souto[1], Lucian Postelnicu[2], Julián Dorado[1] and Cristian Robert Munteanu[1]

[1] Computer Science Faculty, University of A Coruña, 15071 A Coruña, Spain; [2] Infomed Fluids S.R.L., Bd. Th. Pallady nr. 50, sector 3, 032266 Bucharest, Romania

Abstract

The transport of the molecules inside cells is a very important topic, especially in Drug Metabolism. The experimental testing of the new proteins for the transporter molecular function is expensive and inefficient due to the large amount of new peptides. Therefore, there is a need for cheap and fast theoretical models to predict the transporter proteins. In the current work, the primary structure of a protein is represented as a molecular Star graph, characterized by a series of topological indices. The dataset was made up of 2,503 protein chains, out of which 413 have transporter molecular function and 2,090 have no transporter function. These indices were used as input to several classification techniques to find the best Quantitative Structure Activity Relationship (QSAR) model that can evaluate the transporter function of a new protein chain. Among several feature selection techniques, the Support Vector Machine Recursive Feature Elimination allows us to obtain a classification model based on 20 attributes with a true positive rate of 83% and a false positive rate of 16.7%.

Keywords:

QSAR; Star Graph; Support Vector Machine; Topological indices; Transport protein

l. INTRODUCTION

The transport proteins are vital to the growth and life of all living things due to their function of moving the molecules within an organism, including the administrated drugs. There are several different kinds of transport proteins such as carrier proteins that are involved in the movement across a biological membrane of ions, small molecules, or macromolecules, such as another protein [1]. Each carrier protein is designed to recognize only one substance or one group of very similar substances. Therefore, specific diseases are correlated with the errors in transport proteins. Another example of transport protein is the vesicular transporter, a transmembrane or membrane associated protein that regulates or facilitates the movement by vesicles of the contents of the cell.

During the 1950s, the researchers reported in literature that the movement of a number of substances, both uncharged and ionic, across cell membranes was catalyzed by specific proteins. Transporters are divided into passive and active transporters: passive transporters (facilitated transporters) allow passage of solutes (e.g., glucose, amino acids, urea) across membranes down their electrochemical gradients; active transporters create ion/solute gradients across membranes, utilizing diverse energy-coupling mechanisms.

Several mechanisms of transport have been presented. First, the electrochemical potential gradients of cations were observed, which allow the transport of nonelectrolytes against their concentration gradients [2, 3]. Another mechanism is used by the proteins called ion pumps that are driven by the energy of hydrolysis of ATP through the action of specific adenosinetriphosphatases [4], in particular for Na+ and K+, for Ca2+ and for H+ and K+ [5, 6]. Other sources of energy for transport, other than hydrolysis of A TP, were discovered: light in halobacteria [7], oxidative decarboxylation in various bacteria [8], terminal oxidation of substances in inner mitochondrial and chloroplast membranes, etc. The discoveries of selective channels in cell membranes, such as would allow the passage of Na+, K+ and Ca2+ ions across nerve plasma membranes and a number of intracellular membranes in other types of cells, without any energy supply being required and hence no accumulation of such ions against their electrochemical potential gradient being achieved. Another transport pathway was discovered that not only did not require energy coupling but it was either fully or in great measure nonselective, sometimes in the absence of proteins. Therefore, a Transporter Classification (TC) system was defined during the 1990's, in analogy with the Enzyme Commission (EC) system for classification of enzymes. The difference consists in the fact that the TC system incorporates. both functional and phylogenetic information [9]. The TC system provides descriptions, TC numbers, and examples of over 600 families of transport proteins [10]. Transport systems are classified on the basis of five criteria, and each of these criteria corresponds to one of the five numbers or letters within the TC# for a particular type of transporter. Thus, a TC# normally has the following five components: V.W.X.Y.Z. V (a number) corresponds to the transporter class such as channel, carrier/porter, primary active transporter, group translocator or transmembrane electron flow carrier; W (a lettter) corresponds to the transporter subclass which in the case of primary active transporters refers to the energy source used to drive transport; X (a number) corresponds to the transporter family (superfamily); Y (a number) corresponds to the subfamily in which a transporter is found, and Z corresponds to a specific transporter with a particular substrate or range of substrates transported. The TC database (TCDB) can be accessed at http://www.tcdb.org/ [11, 12]. TCDB is a curated database of factual information from over 10,000 published references and it contains about 5,600 unique protein sequences. Another transporter collection consists of TransportDB, a relational database describing the predicted cytoplasmic membrane transport protein complement for organisms whose complete genome sequence is available at http://www.membranetransport.org/ [13-16]. The RCSB Protein Data Bank (http: //www.rcsb.org) is classifying the transporter proteins using the same TC system in: (1) Channels/Pores, (2) Electrochemical Potential-driven Transporters, (3) Primary Active Transporters, (4) Group Translocators, (5) Transmembrane Electron Carriers, (8) Accessory Factors Involved in Transport and (9) Incompletely Characterized Transport Systems. The gene ontology for the transport proteins (G0:0015031) is presented in (Fig. l) and it is available online at http ://www.yeastrc.org.

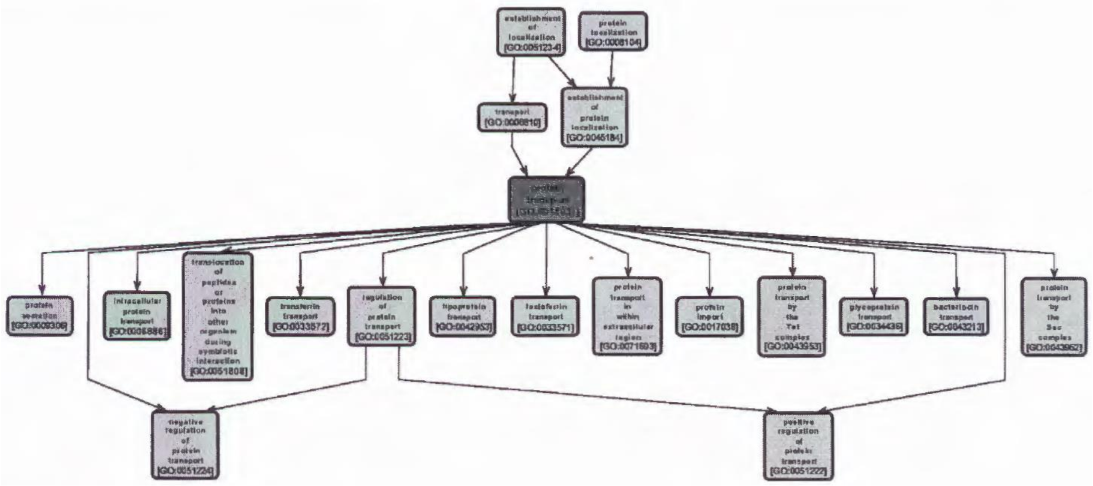
establishment of localization [GO:0051234]
protein localization [GO:0008104]
transport [GO:0006810]
establishment of protein localization [GO:0045184]
protein transport
protein secretion [GO:0009306]
intracellular protein transport [GO:0006886]
translocation of peptides or proteins into other organism during symbiotic interaction [GO:0051808]
transferrin transport [GO:0033572]
regulation of protein transport [GO:0051223]
lipoprotein transport [GO:0042953]
lactoferrin transport [GO:0033571]
protein transport in within extracellular region [GO:0071693]
protein import [GO:0017038]
protein transport by the Tat complex [GO:0043953]
glycoprotein transport [GO:0034436]
bacteriocin transport [GO:0043213]
protein transport by the Sec complex [GO:0043952]
negative regulation of protein transport [GO:0051224]
positive regulation of protein transport [GO:0051222]

Fig. (1). Transport protein gene ontology (GO:OOI5031).

Drug transporters (DTs) are among the many ATPasetype active transporters germane to the pharmaceutical industry. DTs pump drug molecules out of the (primarily) liver cells, in order to maintain low/nontoxic concentrations of a drug molecule inside cells and in the body. Consequently, if the DT activity is over increased, the efficacy of the drugs can be reduced. The drug toxicity and drug clearance are two common reasons for a drug to be withdrawn from the market, costing developers anywhere from hundreds of millions to billions of dollars. Therefore, eliminating these concerns can greatly reduce the cost of and time to market approval. The authors created liver cells with high levels of the enzyme responsible for the liver's metabolic properties. By increasing the levels of oxygen to these cultures, they were able to predict clearance rates from drugs as diverse as anxiolytics and anticonvulsant weeks in advance [17].

The importance of drug transport and drug transporters goes beyond the individual transporter proteins and beyond the assays to measure their activity. Increased research efforts to characterize current transporters and identify new ones are vital to an understanding of drug resistance as well as to determining drug efficacy and safety, in general. Concomitant development of drug transporter assays is equally important in meeting these challenges. Drugs that are substrates or inhibitors of drug transporters have been implicated in drug-drug interactions. Thus, the 2012 FDA DDI draft guidance [18] and the 2012 EMA guideline on investigation of drug interactions [19] mention several specific transporters to investigate and assay conditions to follow. In addition, the European Drug Initiative on Channels and Transporters (EDICT, http://www.edict-project.eu) [20] is helping to determine the structures of clinically significant membrane protein channels and transporters for the initial development of drugs.

Drug transporters play an important role in the absorption, distribution, and excretion (ADE) of many drugs [21]. It was pointed out the significant differences between rodents, dog, monkey, and human in the substrate specificity, tissue distribution, and relative abundance of transporters. These differences complicate cross-species extrapolations, which is important when attempting to predict human pharmacokinetics (PK) of drug candidates and assess risk for drug-drug interactions (DDIs). The quantitative knowledge of species differences of transporters, especially at the protein and functional level is still limited. Therefore, there is a challenge to extrapolate and integrate data from both preclinical species and humans to quantitatively predict the impact of transporters on drug absorption, disposition, and drug-drug interactions. The penetration of drugs into the human brain through the blood-brain barrier (888) represents a major obstacle to the development of successful neuropharmaceuticals. This restricted permeability is due to the delicate intercellular junctions, efflux transporters and metabolizing enzymes present at the BBB [22]. The pharmaceutical industry and academic research rely heavily on permeability studies conducted on animals and *in vitro* models of the BBB. Different drug transporters are involved in specific biological processes in several organs such as brain (ex.: in epilepsy, the P-glycoprotein and multidrug resistance associated proteins) [23], intestine [24], kidney [25, 26], placenta [27], sexual gonads [28] or central nervous system [29].

The human transport proteins for drugs and endogenous substances in plasma membrane domains of intestinal epithelia, hepatocytes, kidney proximal tubules and brain capillary endothelial cells [30] are depicted in (Fig. 2):

a) The *intestinal epithelia* contain in their apical (luminal) membrane several uptake transporters, including one or more members of the organic anion-transporting polypeptide (OATP) family, peptide transporter 1 (PEPT 1; SLCI5Al), ileal apical sodium/bile acid co-transporter (ASBT; SLC10A2), and monocarboxylic acid transporter 1 (MCTl; SLC16AI). The apical ATP-dependent efflux pumps include multidrug resistance protein 2 (MRP2; ABCC2), breast cancer resistance protein (BCRP; ABCG2) and P-glycoprotein (P-gp; MDRl, ABCBl). The baso lateral membrane of intestinal epithelia contains organic cation transporter 1 (OCTl; SLC22Al), heteromeric organic solute transporter (OSTα-OSTβ) and MRP3 (ABCC3).

b) Human *hepatocyte* uptake transporters in the basolateral (sinusoidal) membrane include the sodium/taurocholate co-transporting peptide (NTCP; SLCIOAl), three members of the OATP family - OATPlBl (SLCOJBI), OATPIB3 (SLCOJB3) and OATP2Bl (SLC02BI), organic anion transporter 2 (OAT2; SLC22A7) and OAT7 (SLC22A9), and OCTl. Efflux pumps in the hepatocyte basolateral membrane include MRP3, MRP4 (ABCC4) and MRP6 (ABCC6). Apical (canalicular) efflux pumps of the hepatocyte comprise P-gp, bile-salt export pump (BSEP or SPGP; ABCBil), BCRP (ABCG2) and MRP2. In addition, multidrug and toxin extrusion protein 1 (MATE1; SLC47 Al) is located in the apical hepatocyte membrane.

c) *Kidney proximal tubules* contain OAT4 (SLC22AII), urate transporter 1 (URATJ ; SCL22Al2), PEPTl and PEPT2 (SLCI5A2), MRP2 and MRP4, MATE! and MATE2-K (SLC47A2), P-gp, organic cation/ergothioneine transporter (OCTN 1; SLC22A4), and organic cation/carnitine transporter (OCTN2; SLC22A5) in the apical (luminal) membrane. Basolateral uptake transporters in proximal tubule epithelia include OATP4CI (SLC04Cl), OCT2, OATl, OAT2 and OAT3 (SLC22A8).

d) *Blood- brain barrier:* Apical (luminal) transport proteins of brain capillary endothelial cells contributing to the function of the blood- brain barrier include the uptake transporters OATP1A2 and OATP2B1, the efflux pumps P-gp, BCRP, MRP4 and MRP5 (ABCC5).

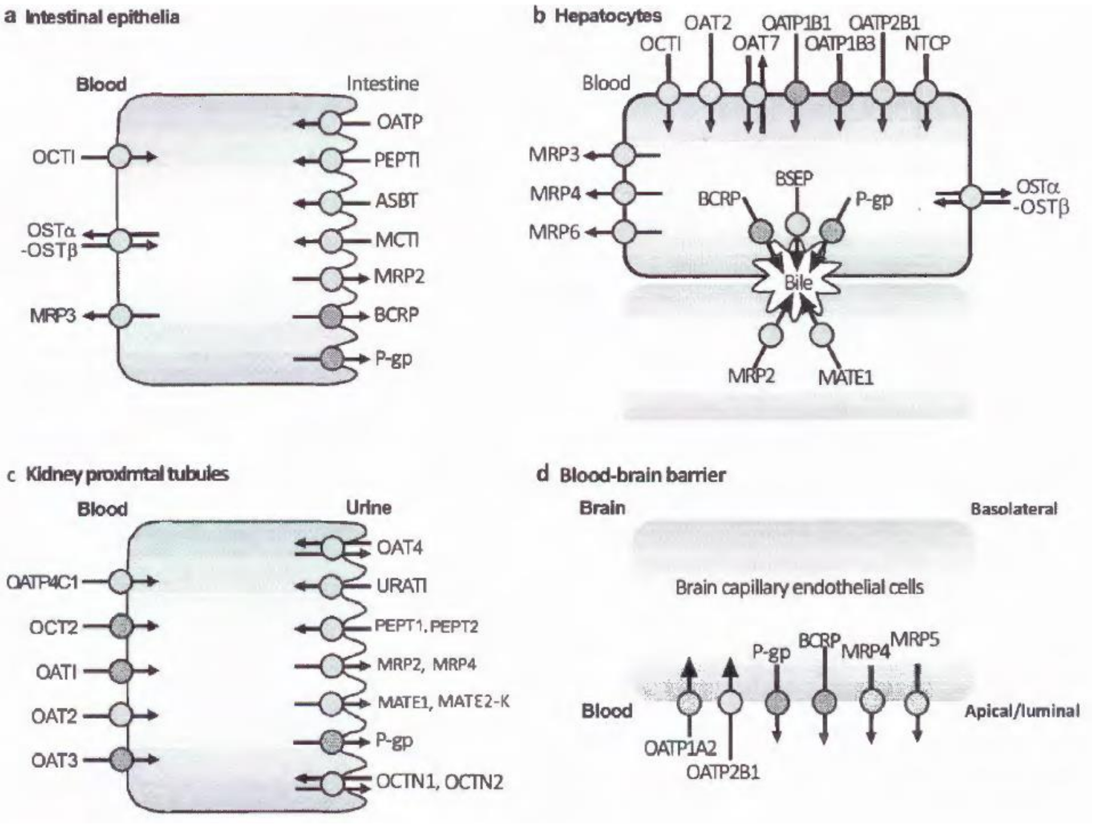

Fig. (2). The human transport proteins for drugs and endogenous substances.

The localization of transporters to particular membranes and tissues is sometimes controversial and, consequently, the International Transporter Consortium is showing only the localization of transporters for which good evidence exists.

The importance of transporters in drug metabolism creates the necessity of cheap and fast theoretical models to predict the transporter function for a new amino acid sequence. This model can reduce the number of molecules for tests in different experiments and increase the success rates when molecules are tested looking for transporters. The current work is putting forward the first theoretical model for transporter prediction using the primary sequences of protein chains. The obtained model is a Quantitative Structure - Activity Relationship (QSAR) [31] between primary amino acid structure of the protein chain and the transporter molecular function. The base of the model consists of the topological indexes (TIs) of macromolecular Star graphs with amino acids as nodes. The graph theory is a branch of mathematical chemistry and it is currently an intense area of research, generating new information regarding DNN/proteins (represented as graphs) [32-35]. The QSAR models are found using several Machine Leaning techniques of classification [36].

The QSAR models have been intensively used with applications in different problems [37, 38] such as the prediction of analgetic agents [39], anti-cancer agents [40-42], anti-inflamatory agents [43], taxane analogues in colon cancer [44], anti-parasitic drugs [45, 46], anti-Alzheimer agents [47] or human breast biomarkers [48]. The graphs/complex networks have been used to graphically process information from molecules (ex: drugs, proteins, nucleic acids) [49-55] to biological systems (ex: parasite-host networks, protein-protein interaction network, gene-gene interaction networks) [56-58].

Protein models obtained with Star Graph TIs have been published and permitted to prediction of cancer [59], DNA promoters [60], human breast and colon cancer-related proteins [61], natural/random proteins [62] or anti-oxidant proteins [63]. The kernel-based machine learning protocols have been previously used to predict DNA-binding proteins [64], the protein function [65, 66], the protein fold [67], the protein functional sites [68], and the drug - protein interaction [69].

In the current work, the authors present the first transport/ non-transport protein classification model based on embedded/non-embedded Star Graph TIs including the trace of connectivity matrices, Harary number, Wiener index, Gutman index, Schultz index, Moreau-Broto indices, Balaban. distance connectivity index, Kier-Hall connectivity indices and Randic connectivity index. Several feature selection methods were tested and the best model was obtained with the Support Vector Machine Recursive Feature Elimination technique.

2. MATERlALS AND METHODS

The description of the methodology followed in this work is presented in (Fig. 3). The database consists of amino acid sequences (primary structure) of transporter and nontransporter proteins in F ASTA format. In the first step, the sequences of amino acids are transformed with S2SNet [70] into topological indices of the protein Star graphs. The resulting numbers that characterized each graph (protein graphical representation) are then used in Weka [71] to find the best QSAR classification model with Machine Learning methods. A final model is suggested to predict the transporter molecular function for new amino acid sequences .

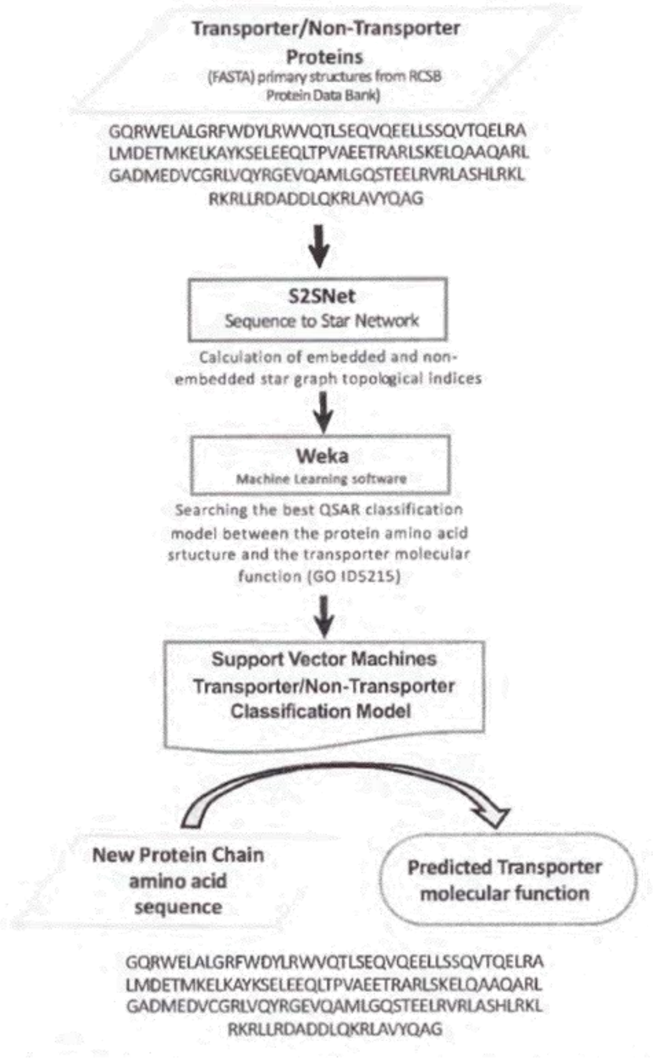
Transporter/Non-Transporter Proteins
(FASTA) primary structures from RCSB Protein Data Bank)
GQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELRA
LMDETMKELKAYKSELEEQLTPVAEETRARLSKELQAAQARL
GADMEDVCGRLVQYRGEVQAMLGQSTEELRVRLASHLRKL
RKRLLRDADDLQKRLAVYQAG
S2SNet
Sequence to Star Network
Calculation of embedded and non-embedded star graph topological indices
Weka
Machine Learning software
Searching the best QSAR classification model between the protein amino acid srtucture and the transporter molecular function (GO ID5215)
Support Vector Machines
Transporter/Non-Transporter
Classification Model
New Protein Chain amino acid sequence
Predicted Transporter molecular function
GQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELRA
LMDETMKELKAYKSELEEQLTPVAEETRARLSKELQAAQARL
GADMEDVCGRLVQYRGEVQAMLGQSTEELRVRLASHLRKL
RKRLLRDADDLQKRLAVYQAG

Fig. (3). General flow chart for obtaining a transporter protein classification model using the topological indices of the protein sequence Stargrapbs.

In statistics, two events are independent when the fact that one occurs does not modify the probability of the other one. An observation is normal when its behavior follows a normal or Gaussian distribution with a certain value of mean and variance. In order to respect this independence condition, the authors have tested the different classification techniques using 10-fold cross-validation to split data [72]. Dataset is randomly partitioned into 10 equal-sized bins. 9 bins were picked 10 times to train the models and the remaining bin was used to test them, each time leaving out a different bin.

The performance of prediction models for a two-class problem (i.e. transport or not transport) is typically evaluated using a confusion matrix. There are several numbers of well-known accuracy measures for a two-class classifier in the literature, such as: classification rate, precision, sensitivity, specificity, F-measure, Area Under the Receiver Operating Characteristic (ROC) Curve, etc. An experimental comparison of performance measures for classification could be found in [73]. The higher the precision, the less effort wasted in testing and inspection; and the higher the recall, the fewer defective modules go undetected. However, there is a tradeoff between precision and recall and therefore a combination of both is needed in a single efficiency measure, known as F-measure, which considers both precision and recall equally important [74]. Jin *et al.* [75] suggested that AUC is a better measure than accuracy when comparing classifiers and in general.

The ROC is also known as a relative operating characteristic curve, because it is a comparison of two operating characteristics (a true positive rate and a false positive rate) as the criterion changes [76]. ROC curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TP Rate = true positive rate) vs. the fraction of false positives out of the negatives (FP Rate = false positive rate), at various threshold settings. TP Rate is also known a~ sensitivity, and FP Rate is one minus the specificity or true negative rate. ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently of (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

The test set is divided randomly into two parts (training and validation) extracting a total of 20% of the training data. The objective of this paper is to evaluate the ability of SVM in predicting transport/non-transport proteins classification.

Several experiments have been performed in order to select the best models. The final compared models are a linear classifier techniques: LIBLINEAR (LL) [77], a neural network technique - Multilayer Perceptrons (MLP), a Bayesian technique - Naïve Bayes, and two tree-structured classifier techniques - Random Forests (RF) and J48. The parameters for each of the models were initialized mostly with the default setting of the Weka toolkit and using the experimenter, they were optimized in order to obtain the best parameter combination for each model.

2.1. Protein Set

The current work is based on datasets extracted from two protein databases. The sets of protein primary sequences are made up of 413 proteins with the transporter molecular function and 2,090 non-transporters. The transporter protein FASTA sequences (positive group) have been downloaded from the Protein Databank [78], the "Transporter (GO ID5215)" list obtained with the "Molecular Function Browser" in the "Advanced Search Interface" (protein identity cut-off= 30%). The negative group was created using the PISCES CulledPDB [79] list of proteins (downloaded on November 16th, 2012) with identity (degree of correspondence between two sequences) less than 20%, resolution of 1.6 A and R-factor 0.25 (non-transporter proteins with any other possible biological function). The sequence identities for PDB sequences in the PISCES server (http://dunbrack.fccc.edu/PISCES.php) have been determined using Combinatorial Extension (CE) structural alignment [80]. This server used a Z-score of 3.5 as the threshold to accept possible evolutionary relationships. PISCES' alignments are local, so that two proteins which share a common domain with sequence identity above the threshold are not both included in the output lists. No list has been post-filtered for any source organism.

2.2. Star Graph Topological lndices

The protein primary scquences were graphically transformed into Star Graphs, where the amino acids are the vertices (nodes), connected in a specific sequence by the peptide bonds. The Star Graph is a special type of tree with *N* vertices where one has got N-1 degrees of freedom and the remaining N-1 vertices have got one single degree of freedom [81] . Each of the 23 possible branches ("rays") of the star contains the same amino acid type and the star centre is a non-amino acid vertex [82]. We used 20 groups for the standard amino acids (A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V), 2 groups with two nonstandard ones (U, O) and 1 group with four ambiguous amino acids (B, Z, J, X). Thus, the following information of the protein primary structure is encoded into the Star Graph connectivity: amino acid type, sequence and frequency.

A protein can be represented by diverse forms of graphs, which can be associated with distinct distance matrices. The best method to create a standard Star Graph is described subsequently: each amino acid/vertex holds the position in the original sequence and the branches are labeled alphabetically by the 3-letter amino acid code. The graph is embedded if the initial sequence connectivity in the protein chain is included.

Graphs are compared using the corresponding connectivity matrix, distance matrix and degree matrix. These matrices and the normalized ones are the basis of the TIs calculation. The calculation of the Star Graph TIs for the amino acid sequences was performed using the Sequence to the Star Networks (S2SNet) application [70]. This tool was programmed in Python/wxPython [83] and has a *Graphviz* [84] plotting back-end. The TIs are calculated using the embedded and non-embedded Star graphs, without weights, using Markov normalization and a power of matrices/indices *(n)* up to 5. The following list depicts the S2SNet TIs [85]:

– Trace of the *n* connectivity matrices (Tr_n):

$$Tr_n = \sum_i (M^n)_{ii}, \tag{1}$$

where n = 0 – power limit, M = graph connectivity matrix (i^*i dimension); ii = i^{th} diagonal element;

– Harari number (H):

$$H = \sum_{i<j} m_{ij}/d_{ij}, \tag{2}$$

where d_{ij} are elements of the distance matrix and m_{ij} are elements of the M connectivity matrix;

– Wiener index (W):

$$W = \sum_{i<j} d_{ij}, \tag{3}$$

– Gutman topological index (S_6):

$$S_6 = \sum_{ij} deg_i * \ deg_j / \ d_{ij}, \tag{4}$$

where deg_i are the elements of the degree matrix;

– Schultz topological index (nontrivial part) (S):

$$S = \sum_{i<j} \left(deg_i + deg_j\right) {}^* \ d_{ij}, \tag{5}$$

– Balaban distance connectivity index (J):

$$J = (edges - nodes + 2) {}^* \sum_{i<j} m_{ij} {}^* \mathrm{sqrt}\left(\sum_k d_{ik} {}^* \sum_k d_{kj}\right), \tag{6}$$

where *nodes* + 1 = AA numbers/node number in the Star Graph + origin, $\sum_k d_{ik}$, k--,d-ik...is the node distance degree;

– Kier-Hall connectivity indices (${}^{n}X$):

$$ {}^{0}X = \sum_{i} 1 \ / \operatorname{sqrt}(deg_i), \tag{7} $$

$$ {}^{2}\mathrm{X} = \sum_{\mathrm{i<j<k}} \mathrm{m_{ij}} \ {}^{*}\mathrm{m_{jk}} \ / \ \mathrm{sqrt}(\mathrm{deg_i} \ {}^{*}\mathrm{deg_j} \ {}^{*}\mathrm{deg_k}), \tag{8} $$

$$ {}^{3}\mathrm{X} = \sum_{\mathrm{i<j<k<m}} \mathrm{m_{ij}} \ {}^{*}\mathrm{m_{jk}} \ {}^{*}\mathrm{m_{km}} \ / \ \mathrm{sqrt}(\mathrm{deg_i} \ {}^{*}\mathrm{deg_j} \ {}^{*}\mathrm{deg_k} \ {}^{*}\mathrm{deg_m}), \tag{9} $$

$$ {}^{4}\mathrm{X} = \sum_{\mathrm{i<j<k<m<o}} \mathrm{m_{ij}} \ {}^{*}\mathrm{m_{jk}} \ {}^{*}\mathrm{m_{km}} \ {}^{*}\mathrm{m_{mo}} \ / \ \mathrm{sqrt}(\mathrm{deg_i} \ {}^{*}\mathrm{deg_j} \ {}^{*}\mathrm{deg_k} \ {}^{*}\mathrm{deg_m} \ {}^{*}\mathrm{deg_o}), \tag{10} $$

$$ {}^{5}\mathrm{X} = \sum_{\mathrm{i<j<k<m<o<q}} \mathrm{m_{ij}} \ {}^{*}\mathrm{m_{jk}} \ {}^{*}\mathrm{m_{km}} \ {}^{*}\mathrm{m_{mo}} \ {}^{*}\mathrm{m_{oq}} \ / \ \mathrm{sqrt}(\mathrm{deg_i} \ {}^{*}\mathrm{deg_j} \ {}^{*}\mathrm{deg_k} \ {}^{*}\mathrm{deg_m} \ {}^{*}\mathrm{deg_o} \ {}^{*}\mathrm{deg_q}), \tag{11} $$

– Randic connectivity index (${}^{1}X$):

$$ {}^{1}X = \sum_{ij} m_{ij} / \operatorname{sqrt}(deg_i \ {}^{*}deg_j), \tag{12} $$

These Tls and other derivate ones will be used in the next step to create an antioxidant *1* non-antioxidant classification model using Machine Leaming methods.

2.3. Support Vector Machines

Vapnik introduces Support Vector Machines (SVMs), a kernel based learning algorithms, in the late 1970s on the foundation of a statistical learning theory [86], using the Structural Risk Minimization (SRM) principle which minimizes the generalization error (i.e. true errors on unseen examples). The basic implementation deals with two-class problems in which data are separated by a hyperplane defined by a number of support vectors. This hyperplane separates the positive from the negative examples, to be oriented in such a way that the distance between the boundary and the nearest data point in each class is maximal; the nearest data points are used to define the margins, known as support vectors [87]. Support vectors are a subset of training data used to define the boundary between the two classes. These classifiers have also proven to be exceptionally efficient in classification problems of higher dimensionality [88]. SVM uses different nonlinear kernel functions, like polynomial, sigmoid and radial basis function (RBF) which yields better prediction performance [89], where the nonlinear SVM maps the training samples from the input spaces into a higher dimensional feature space via a mapping function [87, 90]. Gunn introduced SVMs as an effective technique for solving both classification and regression problems. The main characteristics of SVMs [91] are the following:

- The ability of SVMs to learn can be independent of the feature space dimensionality under small training sample conditions.

- SVMs are formulated as a quadratic programming problem; it gives a global optimum solution.

- They are robust to outliers. Using the margin parameter C, SVMs control the misclassification error and prevent the effect of outliers.

- They can model nonlinear functional relationships that are difficult to model with other techniques.

3. RESULTS

The dataset used in this paper is made up of 2,503 protein sequences, out of which 413 have proved to have the molecular function of transporters (positive group) with maximum identity of 30%. The remaining 2,090 proteins (negative group) are sequences from the CulledPDB server with identity less than 20%, without the transporter function. In order to resample this dataset, a synthetic minority oversampling technique [92] can be applied in Weka. These protein sequences have been processed with the S2SNet application [70] in order to obtain the different topological indexes used in this study. Specifically, from each sequence 42 attributes are calculated for the corresponding embedded/nonembedded Star Graph as shown in Table l.

Table l. Attributes Extracted from Embedded/Non-Embedded Star Graph.

	Attributes (Tis)	
	Non-embedded	Embedded
Sh	Sh0,Shl, Sh2, Sh3,Sh4, Sh5	eSh0, eShl, eSh2,eSh3, eSh4, eSh5
Tr	Tr0, Tr2, Tr4	eTr0, eTr2, eTr3, eTr4, eTr5
X	X0, X1R, X2, X3 , X4 , X5	eX0, eX1R, eX2, eX3, eX4, eX5
Remaining	H, W, S6, S, J	eH, eW, eS6, eS, eJ

Table 2 shows the results of the different classification models tested, considering all the attributes extracted from tl1e Star Graph (42 attributes). This table presents for each model the classification scores obtained for the different classes, true and false positive rates (TP/FP Rates), as well as F-measure, the ROC values for both training and validation and the number of attributes that were considered. The SVM implementation [93] in Weka seems to be the best model because it obtains an AUC-ROC value of 0.911 with 42 features. For this model, the regularization parameter is set to 10; the kernel function used was Gaussian (RBF) with gamma set to 10.

Table 2. Classification Model Results.

	Training				Validation				
	TP Rate	FP Rate	F-Measure	ROC Area	TP Rate	FP Rate	F-Measure	ROC Area	No. of Features
SVM	**0.833**	**0.167**	**0.833**	**0.909**	**0.832**	**0.167**	**0.832**	**0.911**	**42**
MLP	0.607	0.393	0.593	0.633	0.649	0.359	0.643	0.677	42
NB	0.603	0.396	0.595	0.628	0.614	0.395	0.606	0.654	42
RF	0.810	0.190	0.809	0.891	0.812	0.185	0.812	0.898	42
LL	0.616	0.384	0.613	0.616	0.655	0.350	0.653	0.653	42
J48	0.682	0.318	0.682	0.734	0.691	0.310	0.691	0.739	42

In order to reduce the number of features necessary to classify proteins in transport/non-transport, we performed three types of techniques for feature selection.

Technique 1: In Table 1, we presented the grouping of the variables corresponding to the embedded and non-embedded Star graphs. In addition, each one of them is divided into different subsets: a subset called *Sh*, which includes the attributes related to the entropy; a subset called *Tr*, which includes the attributes related to the traces; a subset called *X*, which includes the attributes related to the polygon indexes; and the *remaining* attributes regarding the general shape of the graphs. We have performed different experiments with these subsets separately and in combination with the others with the best technique (SVM) shown in Table 2. Results are shown in Table 3. Using 22 variables, all embedded, SVM obtained an AUC-ROC value of 0.894 with a false positive rate of 0.18. Using 20 variables, *Tr* and *X* subsets, the model obtained an AUC-ROC value of 0.89 with a false positive rate of 0.166. Both are very close to the best result comparing the AUC-ROC values, but both also have a slightly higher false positive rate.

Table 3. Feature Selection Grouping by Membership and Using Weka 's Classifler Subset Evaluator.

	Training				Validation				
	TP Rate	FP Rate	F-Measure	ROC Area	TP Rate	FP Rate	F-Measure	ROC Area	Feat. No.
All	**0.833**	0.167	0.833	**0.909**	**0.832**	0.167	0.832	**0.911**	42
All emb	**0.831**	0.199	0.801	0.873	**0.818**	0.180	0.818	**0.894**	22
All non-emb	0.708	0.292	0.708	0.767	0.711	0.287	0.711	0.765	20
Sh emb	0.614	0.386	0.613	0.660	0.632	0.371	0.631	0.688	6
Sh non-emb	0.636	0.364	0.636	0.680	0.649	0.355	0.647	0.702	6
Sh & Tr emb	0.768	0.232	0.768	0.844	0.785	0.212	0.784	0.861	11
Sh & Tr non-emb	0.591	0.322	0.621	0.701	0.652	0.351	0.651	0.705	9
Sh & Tr	**0.840**	0.231	0.812	0.870	**0.806**	0.194	0.806	0.874	20
Sh	0.650	0.350	0.650	0.703	0.659	0.343	0.659	0.724	12
Sh & X emb	0.655	0.345	0.655	0.712	0.650	0.352	0.650	0.713	12
Sh & X non-emb	0.673	0.327	0.673	0.730	0.686	0.315	0.686	0.736	12
Sh & X	0.718	0.282	0.718	0.781	0.723	0.276	0.723	0.787	24
Tr emb	0.736	0.264	0.736	0.799	0.741	0.257	0.741	0.818	5
Tr non-emb	0.606	0.394	0.606	0.641	0.617	0.386	0.617	0.658	3
Tr	0.757	0.243	0.756	0.822	0.770	0.227	0.770	0.843	8
Tr & X emb	0.776	0.224	0.776	0.841	0.793	0.205	0.793	0.866	11
Tr & X non-emb	0.639	0.361	0.639	0.701	0.646	0.355	0.646	0.703	9
Tr & X	**0.813**	0.187	0.814	0.885	**0.832**	0.166	0.831	**0.890**	20
X emb	0.630	0.370	0.630	0.676	0.627	0.377	0.625	0.674	6
X non-emb	0.633	0.367	0.633	0.683	0.634	0.369	0.633	0.687	6
X	0.691	0.332	0.681	0.720	0.664	0.335	0.664	0.723	12

Technique 2: In Weka, there is a classifier subset eva1uator. This technique evaluates attribute subsets on training data or a separate hold out testing set using a 1 0-fold cross validation, and it uses a classifier to estimate the *merit* of a set of attributes. In this case, this technique obtained a solution with 27 variables, an AUC-ROC value of 0.909 and a false positive rate of 0.161. In this case, the solution has an AUC-ROC value very similar to the best model in Table 2 and it uses only 27 variables. Results are shown in Table 4.

Table 4. Feature Selection Results Using Technique 2.

	Training				Validation				
	TP Rate	FP Rate	F-Measure	ROC Area	TP Rate	FP Rate	F-Measure	ROC Area	No. of Features
SVM	0.827	0.173	0.827	0.902	0.866	0.161	0.866	0.909	27

The final model contains the following embedded and non-embedded Tls: Sh0, Sh1, Sh3, Sh4, Tr0, Tr2, H, W, S6, S, J, X1R, X3, X4, X5, eSh2, eSh5, eTr0, eTr2, eTr3, eTr4, eTr5, eH, eS6, eX1R, eX3 and eX5.

Technique 3: SVM seems to be the fittest model according to results in Table 2. There is a technique for feature selection using SVMs known as Support Vector Machine Recursive Feature Elimination (mSVM-RFE), the output is a feature ranking. SVM-RFE is an iterative algorithm that works backwards an initial set of features [94]. At each round this algorithm fits a simple linear SVM, ranks the features based on their weights in the SVM solution and eliminates the feature with the lowest weight. mSVM-RFE [95] extends this idea by using resampling techniques at each iteration to stabilize the feature ranking; in this work we use a 10-fold cross validation to this end. We estimate the generalization error and use it because when feature selection is performed on a data set with many features, it will pick some features that will be generalized, but it will also pick some useless features. Feature ranking is shown in Table 5. In this case, the features will give good performance if the error is estimated from this training set itself; Guyon *et al.* originally made this mistake, but this issue is outlined in Ref. [96]. Finally, we estimate the generalization error we can expect if we were to train a final classifier using the ranking of features for each of the 10 training sets. Generalization error for 20 variables is shown in (Fig. 4).

Table 6. mSVM-RFE Feature Selection Results.

Average Ranking Value	Attributes
Lower than or equal to 15	eTr4(5.1), eX5(5.1), J(5.2), eTr5(5.2), eTr3(5.7), Shl(8.0), X5(8.8), XIR(9.2), eJ(10.0), eShl(10.9), X0(12.5), S6(13.2), eSh2(13.3), Tr2(14.7), eTr2(15.0)
Higher than 15	eS(16.9), Sh0(18.9), Tr0(19.8), W(19.9), eH(20.7), eX3(22.7), Tr4(23.5), Sh5(24.5), Sh4(25.1), eX2(25.1), eSh4(26.2), Sh2(26.3), eSh3(26.9), eW(27.7), eSh0(29.6), eTr0(30.3), X3(30.9), X2(31.9), S(32.0), eX4(33.0), eX0(33.8), eS6(33.9), eX1R(33.9), eSh5(35.3), H(35.9), X4(37.1), Sh3(39.3)

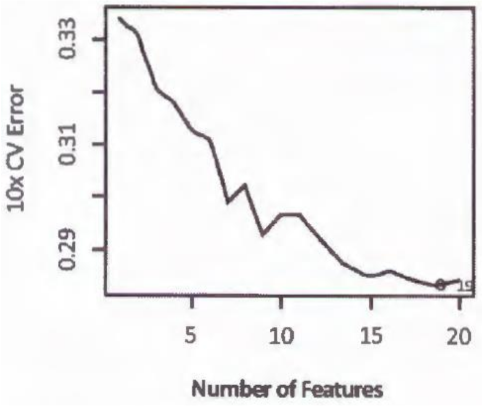

Fig. (4). Generalization error for the 20 first attributes in the ranking.

mSVM-RFE feature selection used a Radial Basis Function (RBF) kernel SVM and a grid search for SVM parameters (cost and gamma) and it was carried out a 10-fold cross validation error estimation. The optimal parameters were used to train the SVM on the entire training set and the generalization error was determined by predicting the corresponding test set. Results are shown in Table 6. We evaluate the top 10, 15, 20 and 2 7 features from the ranking and the top 10, 15 features from the embedded and non-embedded feature groups. We use R [97], a free software environment for statistical computing and graphics and an mSVM-RFE implementation (http://www.colbyimaging.com/wiki/statistics/msvm-rfe).

Table 6. mSVM-RFE Feature Selection Results.

		Training				Validation			
No. of Attr.	Type	TP Rate	FP Rate	F-Measure	ROC Area	TP Rate	FP Rate	F-Measure	ROC Area
10	Rank	0.798	0.203	0.797	0.870	0.813	0.183	0.813	0.881
15	Rank	0.808	0.192	0.808	0.884	0.824	0.173	0.824	0.894
20	**Rank**	**0.809**	**0.191**	**0.809**	**0.888**	**0.830**	**0.167**	**0.830**	**0.900**
27	Rank	0.824	0.176	0.824	0.899	0.836	0.162	0.836	0.906
10	Embedded	0.779	0.221	0.778	0.845	0.806	0.191	0.806	0.867
15	Embedded	0.785	0.215	0.785	0.858	0.807	0.191	0.807	0.883
10	Non-embedded	0.667	0.333	0.677	0.730	0.684	0.314	0.684	0.743
15	Non-embedded	0.694	0.306	0.694	0.751	0.692	0.307	0.692	0.754

These three techniques for feature selection place the accurate number of attributes for solving this problem between 20 and 27. There are hardly differences in the AUC-ROC between them; in order to select the best technique and model we are going to observe the combination of AUCROC, false positive rate and feature number combination, to assert that the 20 first variables in the mSVM-RFE ranking are the best subset of attributes to solve this problem. As you can see in Table 5, the final model contains the following embedded and non-embedded Tls: eShl, eSh2, eTr2, eTr3, eTr4, eTr5, eX5, eH, eJ, eS, Sh0, Shl, Tr0, Tr2, W, J, S6, X0, X1R, X5.

4. DISCUSSION

This paper presents the first theoretical model designed to identify proteins that have transport activity by using Star Graph Tls obtained from protein amino acid sequences (primary structure). SVM seems to be the most accurate model to solve this problem according to the results that reveal the effectiveness of this technique in predicting transport proteins. Three different feature selection techniques were performed in order to reduce the dimensionality of the problem. There are hardly differences in the AUC-ROC between these techniques, but mSVM-RFE gives the best results due to the lowest number of variables (20) and false positive rate combination. These results can help in the prediction of new drug transporters using peptide amino acid sequences with unknown biological functions as a first step of molecular screening in drug metabolism studies.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This work is supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. 10SIN 105004PR). Cristian R. Munteanu acknowledges the funding support for a research position by the "Isidro Parga Pondal" Program from Xunta de Galicia, Spain (supported by the European Social Fund).

REFERENCES

[1] Sadava, D.; Heller, H. C.; Orians, G. H.; Purves, W. K.; Hillis, D. M. *Life, the Science of Biology,* Macmillan Publishers 2009.

[2] Armstrong, C. M. Voltage-gated K channels, *Sci STKE,* 2003, *2003,* re10.

[3] Decoursey, T. E. Voltage-gated proton channels and other proton transfer pathways, *Physiol Rev,* 2003, 83, 475-579.

[4] Borst, P.; Elferink, R. O. Mammalian ABC transporters in health and disease, *Annu Rev Biochem,* 2002, 71, 537-92 .

[5] Cox, D. W.; Moore, S. D. Copper transporting P-type ATPases and human disease, J *Bioenerg Biomembr,* 2002, 34,333-8.

[6] Dunbar, L. A.; Caplan, M. J. The cell biology of ion pumps: sorting and regulation, *Eur J Cell Biol,* 2000, 79, 557-63.

[7] Oesterhelt, D.; Tittor, J. Two pumps, one principle: light-driven ion transport in halobacteria, *Trends Biochem Sci,* 1989, 14,57-61.

[8] Baginsky, M. L.; Huennckcns, F. M. Electron transport function of a heat-stable protein and a flavoprotein in the oxidative decarboxylation of glycine by Peptococcus glycinophilus, *Biochem Biophys Res Commun,* 1966, 23,600-5.

[9] Busch, W.; Saier, M. H., Jr. The transporter classification (TC) system, 2002, *Crit Rev Biochem Mol Biol,* 2002,37,287-337 .

[10] Busch, W.; Saier, M. H., Jr. The IUBMB-endorsed transporter classification system, *Mol Biotechnol,* 2004, *27,* 253-62.

[11] Saier, M. H., Jr.; Tran, C. V.; Bambote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information, *Nucleic Acids Res,* 2006, 34, Dl81 -6.

[12] Saier, M. H., Jr.; Yen, M. R.; Noto, K.; Tamang, D.G.; Elkan, C. The Transporter Classification Database: recent advances, *Nucleic Acids Res,* 2009,37, D274-8.

[13] Ren, Q.; Kang, K. H.; Paulsen, I. T. TransportDB: a relational database of cellular membrane transport systems, *Nucleic Acids Res,* 2004, 32, D284-8.

[14] Ren, Q.; Paulsen, I. T. Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes, *PLoS Comput Biol,* 2005, 1, e27.

[15] Ren, Q.; Chen, K.; Paulsen, I. T. TransportDB: a comprehensive data base resource for cytoplasmic membrane transport systems and outer membrane channels, *Nucleic Acids Res* , 2007, 35, D27 4-9.

[16] Ren, Q.; Paulsen, I. T. Large-scale comparative genomic analyses of cytoplasmic membrane transport systems in prokaryotes, *J Mol Microbiol Biotechnol,* 2007, 12, 165-79.

[17] Kidambi, S.; Yarmush, R. S.; Novik, E.; Chao, P.; Yarmush, M. L.; Nahmias, Y. Oxygen-mediated enhancement of primary hepatocyte metabolism, functional polarization, gene expression, and drug clearance, *Proc Natl Acad Sci USA,* 2009, 106, 15714-9.

[18] 2012 FDA DDI draft guidance. http: //www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatorylnfomlation/Guidances/ucm292362.pdf (Accessed 28/ 12/2012).

[19] 2012 EMA guideline on investigation of drug interactions. http ://www.ema.europa.eu/docs/enGB/documentlibrary/Scientificguideline/2012/07/WC500129606.pdf (Accessed 28/12/2012).

[20] The European Drug lnitiative for Channels and Transporters EU project, *Mol Membr Biol (in press),* 2012 .

[21] Chu, X.; Bleasby, K.; Evers, R. Species differences in drug transporters and implications for translating preclinical findings to humans, *Expert Opin Drug Metab Toxicol*, 2012.

[22] Shawahna, R.; Decleves, X.; Scherrmann, J. M. Hurdles with Using *In vitro* Models to Predict Human Blood-brain Barrier Drug Permeability: A Special Focus on Transporters and Metabolizing Enzymes, *Curr Drug Metab* , 2012 .

[23] Aronica, E.; Sisodiya, S. M.; Gorter, J. A. Cerebral expression of drug transporters in epilepsy, *Adv Drug Deliv Rev,* 2012, 64, 919-29.

[24] Estudante, M.; Morais, J. G.; Soveral, G.; Benet, L. Z. Intestinal drug transporters: An overview, *Adv Drug Deliv Rev,* 2012.

[25] Feng, B.; El-Kattan, A. F.; Radi, Z. A. Renal transporters in drug disposition, drug-drug interactions, and nephrotoxicity, *Curr Protoe Toxicol,* 2012 , Chapter 23, Unit 23 3 1-15.

[26] Morrissey, K. M.; Stocker, S. L.; Wittwer, M. B.; Xu, L.; Giacomini, K. M. Renal Transporters in Drug Development, *Annu Rev Pharmacol Toxicol,* 2012.

[27] Iqbal, M.; Audette, M. C.; Petropoulos, S.; Gibb, W.; Matthews, S. G. Placental drug transporters and their role in fetal protection, *Placenta,* 2012,33, 137-42.

[28] Mruk, D. D.; Cheng, C. Y. Drug transporters in spermatogenesis: A re-evaluation of recent data on P-glycoprotein, *Spermatogenesis,* 2012 , 2, 70-72.

[29] Potschka, H. Role of CNS efflux drug transporters in antiepileptic drug delivery : overcoming CNS efflux drug transport, *Adv Drug Deliv Rev,* 2012, 64, 943-52.

[30] Giacomini, K. M.; Huang, S. M.; Tweedie, D. J.; Benet, L. Z.; Brouwer, K. L.; Chu, X.; Dahlin, A.; Evers, R.; Fischer, V. ; Hillgren, K. M.; Hollinaster, K. A.; lshikawa, T.; Keppler, D.; Kim, R. B.; Lee, C. A.; Niemi, M.; Polli, J. W.; Sugiyama, Y. ; Swaan, P. W.; Ware , J. A.; Wright, S. H.; Yee, S. W. ; Zamek-Gliszczynski, M. J.; Zhang, L. Membrane transporters in drug development, *Nat Rev Drug Discov,* 2010, 9, 215-36.

[31] Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR,* Gordon and Breach, The Netherlands 1999.

[32] Agüero-Chapin, G.; Gonzalez-Diaz, H.; Molina, R.; Varona Santos, J.; Uriarte, E.; Gonzalez-Diaz, Y. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L, *FEBS letters,* 2006, 580, 723-730.

[33] Randié, M.; Balaban, A. T. On A Four-Dimensional Representation of DNA Primary Sequences, *Journal of Chemical lnformation and Modeling,* 2003,43, 532-539.

[34] Munteanu, C. R.; Femandez-Blanco, E.; Seoane, J. A.; Izquierdo Novo, P.; Rodriguez-Femandez, J. A.; Prieto-Gonzalez, J. M.; Rabunal, J.R.; Pazos, A. Drug discovery and design for complex diseases through QSAR computational methods, *Current pharmaceutical design,* 2010,16,2640-2655.

[35] Bielinska-Wa-z, D.; Nowak, W.; Wa-z, P.; Nandyc, A.; Clark, T. Distributionmoments of 2D-graphs as descriptors of DNA sequences, *Chemical Physics Letters,* 2007,443,408-413 .

[36] Frank, I. H. W. a. E. Data Mining: Practical machine learning tools and techniques, Kaufmann: San Francisco 2005.

[37] Puzyn, T.; Leszczynski, J.; Cronin, M. T. D. eds. *Recent Advances in QSAR Studies: Methods and applications,* Springer 2010.

[38] Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation, *Mol. lnf,* 2010, 29,476-488.

[39] Lien, E. J.; Tong, G. L.; Srulevitch, D. B.; Dias, C. QSAR of narcotic analgetic agents, *NIDA Res Monogr,* 1978, 186-96.

[40] Niculescu-Duvaz, I.; Stihi, G.; Craescu, T.; Simon, Z. Potential anticancer agents. XXIII. 2. Quantitative structure--activity relationships (QSAR) in aromatic nitrogen mustards area, *Neoplasma,* 1980,17,271-8.

[41] Niculescu-Duvaz, I.; Craescu, T. Potential anticancer agents. XXIII. 2. Quantitative structure-activity relationships (QSAR) in the "classical" antifolates arca, *Neplasma,* 1982, *29,* 53-63.

[42] Roy, K. K.; Singh, S.; Saxena, A. K. lntegration-mediated prediction enrichment of quantitative model for Hsp90 inhibitors as anticancer agents: 3D-QSAR study, *Mol Divers,* 2010.

[43] Dearden, J. C.; Gregg, C. N.; Nicholson, R. M. QSAR study of analgesic and anti-inflammatory potencies of commercially available non-steroidal anti-inflammatory drugs (NSAIDs), *Prog Clin Biol Res,* 1989,291,353-6.

[44] Verma, R. P.; Hansch, C. QSAR modeling of taxane analogues against colon cancer, *Eur J Med Chem,* 2010, 45, 1470-7.

[45] Prado-Prado, F. J.; Garcia-Mera, X.; Gonzalez-Diaz, H. Multi target spectral moment QSAR versus ANN for antiparasitic drugs against differcnt parasitc specics, *Bioorg Med Chem,* 2010, 18, 2225-31.

[46] Prado-Prado, F. J.; Uriarte, E.; Borges, F.; Gonzalez-Diaz, H. Multi-target spectral moments for QSAR and Complex Networks study of antibacterial drugs, *European Journal of Medicinal Chemistry,* 2009, 44, 4516-21.

[47] Solomon, K. A.; Sundararajan, S.; Abirami, V. QSAR studies on N-aryl derivative activity towards Alzheimer's disease, *Molecules* ,2009,14, 1448-55.

[48] Vilar, S.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding lattice networks, *J Comput Chem,* 2008, 29, 2613-22.

[49] Balaban, A. T. Chemical graphs. XXXJV. Five new topological indices for the branching of tree-like graphs, *Theor. Chim. Acta,* 1979, 53, 355-375.

[50] Balaban, A. T. Topological indices based on topological distances in molecular graphs, *Pure Appl. Chem.,* 1983, 55, 199-206.

[51] Balaban, A. T.; Balaban, T. S. New vertex invariants and topological indices of chemical graphs based on information on distances, *J. Math. Chem.,* 1991, 8, 383-397.

[52] Ivanciuc, O. QSAR comparative study of Wiener descriptors for weighted molecular graphs, *J Chem lnf Comput Sci,* 2000, 40, 141 2-22.

[53] Gutman, I.; Rucker, C.; Rucker, G. On walks in molecular graphs, *J Chem Inf Comput Sci,* 2001, 41 ,739-45.

[54] Randic, M. Novel shape descriptors for molecular graphs, *J Chem Inf Comput Sci,* 2001,41,607- 13.

[55] Li, C.; Tang, N.; Wang, J. Directed graphs of DNA sequences and their numerical charactelization, *J Theor Biol,* 2005.

[56] Bornholdt, S.; Schuster, H. G. *Handbook of Graphs and Complex Networks: From the Genome to the Internet,* WILEY-VCH GmbH & CO. KGa.: Wheinheim 2003.

[57] Zhang, X. F.; Dai, D. Q.; Ou-Yang, L.; Wu, M. Y. Exploring overlapping functional units with valious structure in protein interaction networks, *PLoS One,* 2012, 7, e43092.

[58] Riera-Fernandez, P.; Munteanu, C. R.; Escobar, M.; Prado-Prado, F.; Martin-Romalde, R.; Pereira, D.; Villalba, K.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks, J *Theor Biol,* 2012,293, 174-88.

[59] Vázquez, J. M.; Aguiar, V.; Seoane, J. A.; Freire, A.; Serantes, J. A.; Dorado, J.; Pazos, A.; Mumeanu, C. R. Star Graphs of Protein Sequences and Proteome Mass Spectra in Cancer Prediction, *Current Proteomics,* 2009, 6, 27 5-288.

[60] Perez-Bello, A.; Munteanu, C. R.; Ubeira, F. M.; De Magalhaes, A. L.; Uriarte, E.; Gonzalez-Diaz, H. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices, *J Theor Biol,* 2009, 256, 458-66.

[61] Munteanu, C. R.; Magalhaes, A. L.; Uriarte, E.; Gonzalez-Diaz, H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices, *J Theor Biol,* 2009, 257, 303-11.

[62] Munteanu, C. R.; Gonzalez-Diaz, H.; Borges, F.; de Magalhaes, A. L. Natural/random protein classification models based on star network topological indices, *J Theor Biol,* 2008, 254, 775-83.

[63] Fernandez-Bianco, E.; Aguiar-Pulido, V.; Munteanu, C. R.; Dorado, J. Random Forest classification based on star graph topological indices for antioxidant proteins, *J Theor Biol,* 2012, 317, 331-7.

[64] Bhardwaj, N.; Langlois, R. E.; Zhao, G.; Lu, H. Kernel-based machine learning protocol for predicting DNA-binding proteins, *Nucleic Acids Res,* 2005, 33, 6486-93.

[65] Lanckriet, G. R.; Deng, M.; Cristianini, N.; Jordan, M. I.; Noble, W. S. Kernel-based data fusion and its application to protein function prediction in yeast, *Pac Symp Biocomput,* 2004, 300-11 .

[66] Lee, H.; Tu, Z.; Deng, M.; Sun, F.; Chen, T. Diffusion kernel-based logistic regression models for protein function prediction, *Omics,* 2006, 10, 40-55.

[67] Langlois, R. E.; Diec, A.; Dai, Y.; Lu, H. Kernel based approach for protein fold prediction from sequence, *Conf Proc IEEE Eng Med Biol Soc,* 2004, 4, 2885-8.

[68] Maji, P.; Das, C. Efficient design of bio-basis function to predict protein functional sites using kernel-based classifiers, *IEEE Trans Nanobioscience,* 2010, 9, 242-9.

[69] Wang, Y. C.; Zhang, C. H.; Deng, N. Y.; Wang, Y. Kernel-based data fusion improves the drug-protein interaction prediction, *Comput Biol Chem,* 2011 , 35, 353-62.

[70] Munteanu, C. R.; Magalhaes, A. L.; Uriarte, E.; González-Díaz, H. Multi-Target QPDR Classification Model for Human Breast and Colon Cancer-Related Proteins using Star Graph Topological Indices, *J Theor Biol,* 2009, 257, 303-311.

[71] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. ; Witten, I.A. The WEKA Data Mining Software: An Update, *SIGKDD Explorations,* 2009, 11.

[72] McLachlan, G. J.; Do, K.-A.; Ambroise, C. *Analyzing microarray gene expression data* , Wiley 2004.

[73] Ferri, C.; Hernandez-Orallo, J.; Modroiu, R. An experimental comparison of performance measures for classification, *Pattern Recogn Lett.,* 2009, 30, 27-38.

[74] Witten, I.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems), Morgan Kaufmann 2005.

[75] Jin, H. Using AUC and Accuracy in Evaluating Learning Algorithms, *IEEE Transactions on Knowledge and Data Engineering,* 2005, 17, 299-310.

[76] Swets, J. A. Signal detection theory and ROC analysis in psychology and diagnostics : collected papers, Lawrence Erlbaum Associates: Mahwah, NJ 1996.

[77] Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; Lin, C.-J. LIBLINEAR: A Library for Large Linear Classification, *J. Mach. Learn. Res.,* 2008, 9, 1871-1874.

[78] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G. ; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Boume, P. E. The Protein Data Bank *Nucleic Acids Research,* 2000, 28, 235-242.

[79] Wang, G.; R. L. Dunbrack, J. PISCES: a protein sequence culling server, *Bioinformatics,* 2003, 19, 1589-1591.

[80] Shindyalov, I. N.; Boume, P. E. Protein Structure Alignment by Incremental Combinatorial Extension of the Optimum Path, *Protein Engineering,* 1998, 11, 739-747.

[81] Harary, F. *Graph Theory:* Reading,MA 1969.

[82] Randic, M.; Zupan, J.; Vikic-Topic, D. On representation of proteins by star- like graphs, *J Mol Graph Model,* 2007, 290-305.

[83] Rappin, N.; Dunn, R. *wxPython in Action,* Manning Publications Co.: Greenwich, CT 2006.

[84] Koutsofios, E.; North, S. C. *Drawing Graphs with dot,* AT&T Bell Laboratories, Murray Hill: NJ, USA 1993.

[85] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley-VCH 2002.

[86] Vapnik, V. N.; Nauka, English Translation Springer Verlang, 1982, 1979.

[87] Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition, *Data Min. Knowl. Discov.* , 1998, 2, 121-167.

[88] Moulin, L. S.; da Silva, A. P. A.; El-Sharkawi, M. A.; Marks, R. J., II. Support vector machines for transient stability analysis of large-scale power systems, *Power Systems, IEEE Transactions on,* 2004, 19, 818-825.

[89] Smola, A. J. In *Departmenr of Computer Science;* Technical University: Berlin, Germany, 1998; Vol. *Ph.D. dissertation.*

[90] Gunn, S. Support Vector Machines for Classification and Regression, *ISJS Technical Report,* 1998.

[91] Cortes, C.; Vapnik, V. In *Machine Learning,* 1995; Vol. *20,* pp. 273-297.

[92] Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique, *J. Artif. Int. Res.,* 2002, 16, 321-3 57.

[93] Chang, C. C.; Lin, C. J. LIBSVM: A Library for support vector machines, *ACM Transactions on lntelligent Systems and Technology,* 2011, 2.

[94] Guyon, I.; Weston, J.; Barnhill , S. ; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines, *Mach. Learn.,* 2002, 46, 389-422.

[95] Duan, K. B.; Rajapakse, J. C.; Wang, H.; Azuaje, F. Mu1tiple SVM-RFE for gene selection in cancer classification with expression data, *IEEE Trans Nanobioscience,* 2005, 4, 228-34.

[96] Ambroise, C.; McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data, *Proc Natl Acad Sci U* S *A,* 2002, 99, 6562-6.

[97] Development Core, T., 2005.