Prediction of Compounds with Specific Pharmacodynamic, Pharmacokinetic or Toxicological Property by Statistical Learning Methods

C.W. Yap[1], Y. Xue[2], H. Li[1], Z.R. Li[1,2], C.Y. Ung[1], L.Y. Han[1], C.J. Zheng[1], Z.W. Cao[3] and Y.Z. Chen*[1,3]

[1]*Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543*

[2] *College of Chemistry, Sichuan University, Chengdu, 610064, P. R. China*

[3] *Shanghai Center for Bioinformation Technology, Shanghai, 201203, P. R. China*

Abstract: Computational methods for predicting compounds of specific pharmacodynamic, pharmacokinetic, or toxicological property are useful for facilitating drug discovery and drug safety evaluation. The quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) methods are the most successfully used statistical learning methods for predicting compounds of specific property. More recently, other statistical learning methods such as neural networks and support vector machines have been explored for predicting compounds of higher structural diversity than those covered by QSAR and QSPR. These methods have shown promising potential in a number of studies. This article is intended to review the strategies, current progresses and underlying difficulties in using statistical learning methods for predicting compounds of specific property. It also evaluates algorithms commonly used for representing structural and physicochemical properties of compounds.

Keywords: Statistical learning methods, pharmacodynamic, pharmacokinetic, toxicology, QSAR, QSPR, molecular descriptors, structural diversity.

INTRODUCTION

Modern drug discovery efforts have primarily been based on the search and optimization of compounds that possess specific pharmacodynamic and pharmacokinetic properties, and on the test of their potential toxicological and side effects [1-3]. Methods for predicting these properties, particularly in the early design stages, are useful for facilitating drug development and drug safety evaluation [1, 4, 5]. As part of an effort for accelerating and reducing the cost of drug discovery processes, computational methods have been explored for predicting compounds that possess specific pharmacodynamic, pharmacokinetic or toxicological property [6-9]. In particular, statistical learning methods have shown promising potential for performing these tasks by statistically analyzing the structural and physicochemical features of the compounds known to possess a particular property to derive explicit or hidden statistical models or rules for predicting the activity or property of new compounds [8, 10, 11].

Quantitative structure activity relationship (QSAR) and quantitative structure property relationship (QSPR) are the first explored statistical learning methods that have found successful applications in predicting activities of compounds of specific property [6, 7]. More recently, other statistical learning methods such as neural networks (NN) and support vector machines (SVM) have been explored for the prediction of classes of compounds of more diverse ranges of structures than those covered by QSAR and QSPR methods [8, 10, 11]. These recently explored statistical learning methods classify compounds into two classes, one possessing a particular property and the other without that property, regardless of whether or not their structural and physicochemical properties obey a QSAR- or QSPR-like analytical relationship. They are therefore expected to be applicable to compounds of more diverse structural ranges than those covered by QSAR and QSPR methods.

This article reviews the strategies, current progresses and underlying difficulties in the application of these statistical learning methods. QSAR and QSPR methods have been extensively reviewed elsewhere [6, 7] and they are thus not described here. Proper representation of the structural and physicochemical features of compounds is a key to the successful application of statistical learning methods. A large number of molecular descriptors have been derived to quantitatively represent different structural and physicochemical properties [12-15]. These molecular descriptors and how they are selected and used in the statistical learning methods are also discussed.

STRUCTURAL DIVERSITY OF COMPOUNDS

Structural diversity of the compounds in a dataset can be determined by the diversity index (DI) which is the average value of the similarity between all the pairs of compounds in the dataset [16]:

$$DI = \frac{\sum_{i=1}^{N} \sum_{i=1, i=j}^{N} \mathrm{sim}(i, j)}{N(N-1)}$$

*Address correspondence to this author at the Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543; Tel: 65-6874-6877; Fax: 65-6774-6756; E-mail: csccyz@nus.edu.sg

1389-5575/06 $50.00+.00

where *sim(i,j)* is a measure of the similarity between compounds *i* and *j* and N is the number of compounds in the dataset. The closer the DI is to 0, the more diverse is the dataset. A common similarity measure is the Tanimoto coefficient [17-19]:

$$\mathrm{sim}(i,j) = \frac{\sum_{d=1}^{i} X_{di} X_{dj}}{\sum_{d=1}^{i} (X_{di})^2 + \sum_{d=1}^{i} (X_{dj})^2 - \sum_{d=1}^{i} X_{di} X_{dj}}$$

where *l* is the number of descriptors computed for the compounds in the dataset. Representativity of validation set is measured by the mean Tanimoto coefficient between compounds in the validation set and those in the training set.

Table **1** gives the DI of the compound datasets studied by statistical learning methods. It is found that the DI value of some of the datasets is very small, as low as 0.388, which is at the level of those of highly diverse datasets. For comparison, the DI values of datasets containing congeneric compounds are typically greater than 0.733, and those of the compounds used in QSAR and QSPR are typically in the range of 0.692 to 0.919. This suggests that statistical learning methods are useful for studying many different pharmacokinetic and toxicological properties which intrinsically involve compounds of highly diverse structures.

MOLECULAR DESCRIPTORS

Molecular descriptors are used for representing structural and physicochemical properties of compounds based on their 1D, 2D or 3D structure. There are a number of computer programs for deriving molecular descriptors, which include DRAGON [12], Molconn-Z [13], JOELib [14], and Xue descriptor set [15]. Over 1400 molecular descriptors can be derived from these methods, which range from constitutional

Table 1. Diversity Index (DI) of Datasets of Compounds Used in Statistical Learning Methods

	Dataset	Number of compounds	Diversity index
Statistical learning datasets	Blood-brain barrier penetrating agents[a]	158	0.388
	Genotoxicity[b]	850	0.434
	Torsade de pointes causing agents[c]	344	0.456
	P-glycoprotein substrates[d]	189	0.500
	CYP3A4, CYP2D6, CYP2C9 substrates/inhibitors[e]	692	0.525
	Human intestinal absorbing agents[f]	218	0.543
	Total clearance[g]	503	0.562
	Human serum albumin binders[a]	93	0.585
	Milk-plasma ratio[a]	121	0.596
Highly diverse datasets	Satellite structures[h]	8	0.231
	FDA approved drugs	1121	0.495
	NCI Diversity set[i]	1804	0.544
Congeneric datasets	Penicillins	59	0.733
	Cephalosporins	73	0.772
	Fluoroquinolones	39	0.865
QSAR, QSPR datasets	Estrogen receptor ligands[j]	1009	0.692
	Dihydrofolate reductase (DHFR) inhibitors[j]	756	0.726
	Benzodiazepine receptor ligands[j]	405	0.739
	Cyclooxygenase 2 (COX2) inhibitors[j]	467	0.919

[a] [37]
[b] [38]
[c] [35]
[d] [36]
[e] [41]
[f] [15]
[g] Yap, C. W.; Li, Z. R.; Chen, Y. Z. Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. *J. Mol. Graph. Mod.* **2005,** *24*, 383.
[h] [63]
[i] [64]
[j] [65]

descriptors to more complex 2D and 3D descriptors representing different geometric, connectivity, and physicochemical properties.

The commonly used descriptors can be divided into 18 classes, which include constitutional descriptors such as molecular weight, geometrical descriptors such as volume and surface areas, topological descriptors such as the number of rings and rotatable bonds, RDF descriptors representing interatomic distances in the entire molecule and other useful information such as bond distances, ring types, planar and non-planar systems, atom types and molecular weight [20], molecular walk counts [21], 3D-MoRSE descriptors describing features such as molecular weight, van der Waals volume, electronegativities and polarizabilities [22], BCUT descriptors representing connectivity information and atomic properties relevant to intermolecular interaction [23], WHIM descriptors describing size, shape, symmetry, atom distribution and polarizability of a molecule [24], Galvez topological charge indices and charge descriptors [25], GETAWAY descriptors [26], 2D autocorrelations, functional groups, atom-centred descriptors, aromaticity indices [27], Randic molecular profiles [28], electrotopological state descriptors [29], linear solvation energy relationship descriptors [30], and other empirical and molecular properties.

Not all of these descriptors are needed for representing features of a particular class of compounds. Features useful for compounds of a particular property can be selected either by intuition as those used in QSAR and QSPR studies, or by using feature selection methods. The commonly used

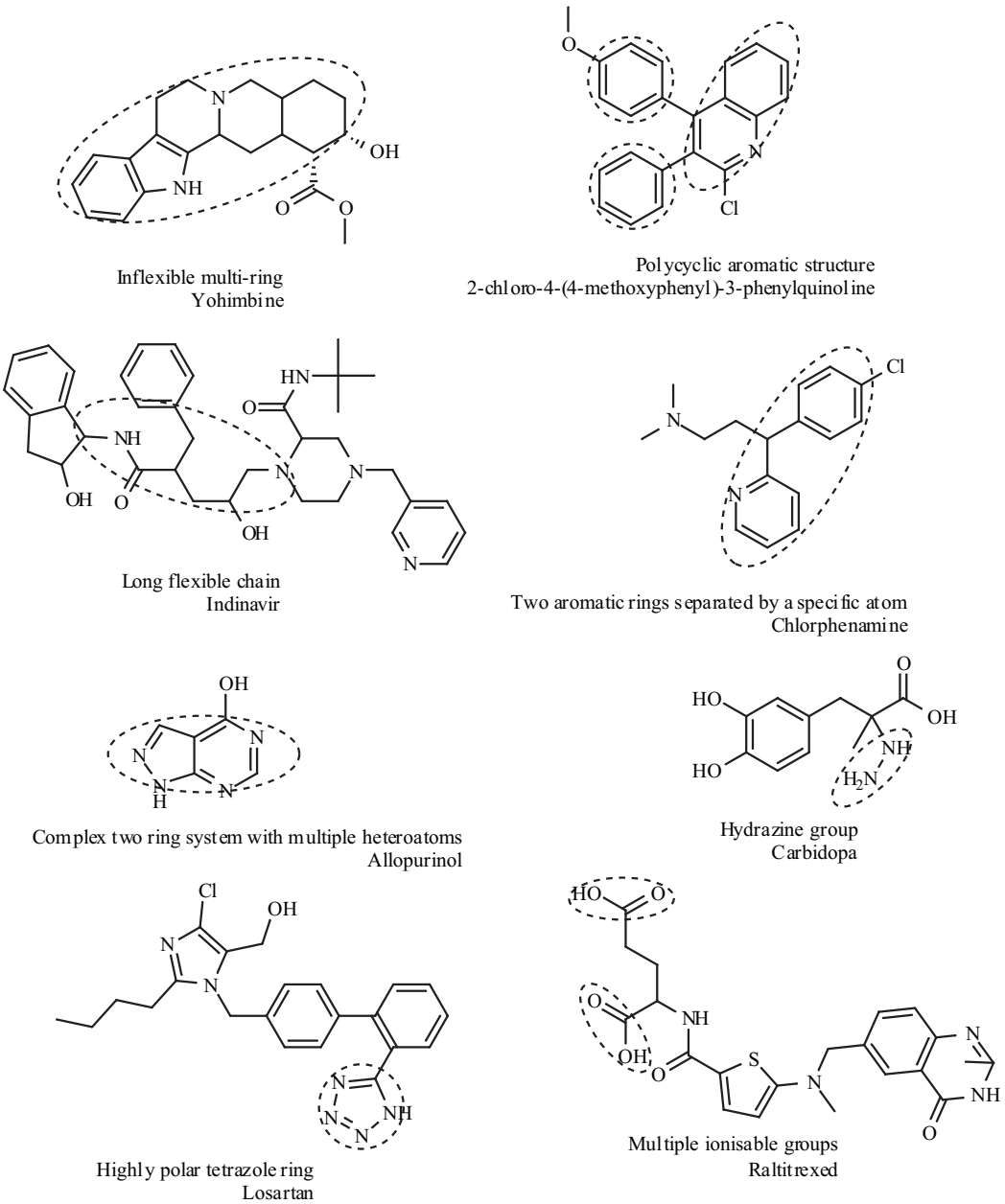

Fig. (1). Examples of compounds not-well-represented by the currently available molecular descriptors. The not-well-represented part of the structure is indicated by a dashed line.

feature selection methods include genetic algorithm-based approach [31], recursive feature eliminations [32], and simulated annealing-based approach [33]. Some of these methods have gained popularity due to their effectiveness for discovering informative features in the analysis of drug activity [32, 34] and pharmacokinetic and toxicological properties [15, 35-38].

However, in many cases, it is difficult to uniquely select an optimum set of descriptors due to the high redundancy and overlapping of many descriptors [39]. Separate sets of descriptors containing different members of redundant descriptor classes have been found to give similar prediction accuracies [40]. The interpretation of the prediction results in these cases should be more appropriately conducted at the descriptor class level where redundant and overlapping descriptors are grouped into classes [41].

It has been found that some compounds cannot be adequately represented by the currently available molecular descriptors [36, 38, 41]. (Fig. **1**) gives examples of such compounds. These include compounds containing inflexible multi-rings, highly polar tetrazole rings, aromatic rings separated by a specific atom, complex two ring system with multiple heteroatoms, polycyclic aromatic structures, long flexible chains, hydrazine group, and multiple ionisable groups. Therefore, there is a need for deriving new descriptors to adequately represent features of these and other compounds.

COMMONLY USED STATISTICAL LEARNING METHODS

Logistic Regression (LR)

LR [42] is based on the assumption that a logistic relationship exists between the probability of class membership and one or more descriptors. The probability $Y = \frac{1}{1+e^{-\beta 0 + \beta 1\chi 1 + \beta 2\chi 2 + \kappa + \beta\kappa\chi\kappa}}$, where $\boldsymbol{x}$={x1, …, xk} is a feature vector, xk is a descriptor, $\beta 0$ is the regression model constant, $\beta 1$ to βk is the coefficients corresponding to the descriptors *X1* to *Xk*. $Y > 0.5$ or $Y < 0.5$ indicates that the vector $\boldsymbol{x}$ belongs to the positive or negative class respectively.

Linear Discriminant Analysis (LDA)

LDA [43] separates two classes of vectors by constructing a hyperplane defined by the following linear discriminant function: $L = \sum_{i}^{k} w_i x_i$, where L is the resultant classification score and w_i is the weight associated with the corresponding descriptor x_i. A positive or negative L value indicates that a vector $\boldsymbol{x}$ belongs to the positive or negative class respectively.

k Nearest Neighbor (kNN)

In kNN, the Euclidean distance between an unclassified vector $\boldsymbol{x}$ and each individual vector $\boldsymbol{x}_i$ in the training set is measured [44, 45]. A total of k number of vectors nearest to the unclassified vector $\boldsymbol{x}$ are used to determine the class of that unclassified vector. The class of the majority of the k nearest neighbors is chosen as the predicted class of the unclassified vector $\boldsymbol{x}$.

C4.5 Decision Tree (C4.5 DT)

C4.5 DT is a branch-test-based classifier [46]. A branch of the decision tree corresponds to a group of classes and a leaf represents a specific class. A decision node specifies a test on a single attribute value, with one branch and its subsequent classes as possible outcomes. C4.5 decision tree uses recursive partitioning to examine every attribute of the data and rank them according to their ability to partition the remaining data, thereby constructing a decision tree. A vector $\boldsymbol{x}$ is classified by starting at the root of the tree and moving through the tree until a leaf is encountered. At each non-leaf decision node, a test is conducted to move into a branch. Upon reaching the destination leaf, the class of the vector $\boldsymbol{x}$ is predicted to be that of the leaf.

Probabilistic Neural Network (PNN)

PNN is a form of neural network that uses Bayes optimal decision rule for classification [47]. Traditional neural networks such as feed-forward back-propagation neural network rely on multiple parameters and network architectures to be optimized. In contrast, PNN only has a single adjustable parameter, a smoothing factor σ for the radial basis function in the Parzen's nonparameteric estimator. Thus the training process of PNN is usually orders of magnitude faster than those of the traditional neural networks.

Support Vector Machine (SVM)

Linear SVM constructs a hyperplane separating two different classes of feature vectors with a maximum margin [48]. This hyperplane is constructed by finding a vector **w** and a parameter b that minimizes which $\|w\|^2$satisfies the following conditions: $w \cdot x_i + b \geq +1$, for $y_i = +1$ (positive class) and $w \cdot x_i + b \leq -1$, for $y_i = -1$ (negative class). Here $\mathbf{x}_i$ is a feature vector, y_i is the group index, **w** is a vector normal to the hyperplane, $|b| / \|w\|^2$ is the perpendicular distance from the hyperplane to the origin and $\|w\|^2$ is the Euclidean norm of **w**. Nonlinear SVM projects feature vectors into a high dimensional feature space by using a kernel function such as $K(xi, xj) = e^{-[x_i - x_j]^2/2a^2}$. The linear SVM procedure is then applied to the feature vectors in this feature space. After the determination of **w** and b, a given vector **x** can be classified by using $sign[(w \cdot x) + b]$, a positive or negative value indicates that the vector **x** belongs to the positive or negative class respectively.

PREDICTION PERFORMANCE

Classification Methods

Classification-based statistical learning methods are intended for determining whether or not a compound belongs to a compound class whose members possess a common property. These methods are capable of classification of a diverse range of compounds but they are not intended for providing the activity of these compounds.

(Table **2**) summarises the performance of the commonly-used classification methods for predicting compounds of various pharmacodynamic, pharmacokinetic and toxicological properties. The performance of these methods has been measured by the positive prediction accuracy P_p for compounds that possess specific property and the negative prediction accuracy P_n for compounds without that property. Moreover, an overall accuracy P=(TP+TN)/N, where TP and TN is the true positive and true negative respectively and N is the number of compounds in the dataset, can also be used to indicate the overall prediction performance. The number of compounds in many of the studies listed in (Table **2**) is in the range of hundreds or even thousands of compounds, which is significantly higher than the tens of compounds typically used in QSAR and QSPR studies [49, 50].

The computed P_p values are in the range of 73% ~ 100%, with the majority concentrated in the range of 80%~97%. The computed P_n values are distributed in the range of 46% ~ 98%, with the majority concentrated in the range of 70%~97%. These results suggest that the classification methods surveyed here have certain level of capability for distinguishing between compounds of particular property and those without that property. In these studies, the negative accuracies P_ns appear to be somewhat lower than the positive accuracies P_ps. One likely reason for the lower P_ns is the inadequate representation of the negative compounds that are known to not have a particular property. The number of the negative compounds in the published studies is typically in the range of a few hundred or less, which is unlikely to be sufficient to fully represent the vast chemical space of millions of compounds in the chemical databases.

Regression Methods

Regression-based statistical learning methods are intended for providing some estimate about the activity value in addition to the determination of whether or not a compound possesses a specific property. (Table **3**) summarises the performance of several regression methods for predicting compounds of various pharmacodynamic, pharmacokinetic and toxicological properties. The performance of these studies is primarily measured by the r^2 value, which measures the explained variance between the

Table 2. Performance of Classification-Based Statistical Learning Methods for Predicting Compounds of Specific Pharmacodynamic, Pharmacokinetic or Toxicological Property. The Relevant Literature References are Given in Supplementary Materials

Property	Method	Molecular descriptors	Number of compounds in training set	Validation method [a]	Reported prediction accuracy		
					Positive accuracy P_p	Negative accuracy P_n	Overall accuracy P
HIA	LDA	TOPS-MODE	82	Validation set (127)	95.5	76.5	92.9
	C-SAR	Simple physicochemical parameters	977	Training set (977)	97.0	81.7	95.7
	PNN	Log P, MR, TOP	76	Validation set (10)	100.0	50.0	80.0
	SVM	Simple molecular properties, molecular connectivity and shape, E-state, Q-C, GEO	196	5 fold CV (196)	90.0	80.7	86.7
Bioavailability	ORMUCS	Log P, structural	232	Validation set (40)	-	-	60.0
	Adaptive fuzzy partition	CON, information, TOP, E-state, physicochemical, ELE	352	Validation set (75)	-	-	64.0
P-gp substrate	SVM	Simple molecular properties, molecular connectivity and shape, E-state, Q-C, GEO	142	Validation set (25)	84.2	66.7	80.0
BBB penetration	MLR	Daylight, thermodynamic, spatial, structural, TOP, charge	48	Validation set (150)	81.0	95.8	88.0
	Discrimination function analysis	TOP, substructures, GEO, Q-C	28	LOO (28)	100.0	91.7	96.4
	PLS	Log P, PSA, E-state	58	Validation set (181)	85.7	46.7	66.3
	PLS-DA	ADME screen, geometry, topology, VAMP electronic parameters, VAMP energy parameters, Sybyl surface areas	1696	Validation set (82)	90.0	92.0	91.0

(Table 2). contd.....

Property	Method	Molecular descriptors	Number of compounds in training set	Validation method [a]	Reported prediction accuracy		
					Positive accuracy P_p	Negative accuracy P_n	Overall accuracy P
	SUBSTRUCT	Substructures	8678	10 fold CV (8678)	83.3	71.2	76.3
	Bayesian neural network	CON, log P, ISIS fingerprint	>73000	Validation set (84)	94.7	73.9	83.3
	PCA	VolSurf	110	Validation set (120)	90.9	64.8	71.7
	SVM	Structural	172	Validation set (304)	78.9	60.4	76.0
		VolSurf	238	Validation set (238)	91.8	68.5	86.6
		MW, lipophilicity, H-bond	274	Validation set (50)	82.7	80.2	81.5
CYP3A4 inhibitor	PLS	CATS, TOP, ELE, count, structural, atom types	311	Validation set 1 (50) Validation set 2 (10)	93.1 100.0	85.7 66.7	90.0 90.0
	ANN	Unity fingerprint	218	Validation set (72)	91.7	88.9	90.3
	Consensus SVM	DRAGON	602	Validation set (100)	92.0	97.3	96.0
CYP2D6 inhibitor	Consensus recursive partitioning	TOP, E-state, physicochemical, fragment keys, 1D similarity scores	100	Validation set (51)	100	76.0	80.0
	Consensus SVM	DRAGON	602	Validation set (100)	90.0	95.0	94.0
CYP2C9 inhibitor	Consensus SVM	DRAGON	602	Validation set (100)	88.9	96.3	95.0
CYP2D6 substrate	Consensus SVM	DRAGON	602	Validation set (100)	98.2	90.9	95.0
CYP3A4 substrate	Consensus SVM	DRAGON	602	Validation set (100)	96.6	94.4	95.0
CYP2C9 substrate	Consensus SVM	DRAGON	602	Validation set (100)	85.7	98.8	97.0
Genotoxic	KNN	TOP, GEO, ELE, PSA	120	Validation set (20)	66.7	92.9	85.0
	Consensus KNN	TOP, GEO, ELE, Q-C, CPSA, H-bond, nitrogen-specific	334	3 fold CV (334)	69.3	74.1	72.2
	Consensus model (KNN, LDA, PNN)	TOP, GEO, ELE, CPSA, H-bond	227	3 fold CV (227)	73.8	84.3	81.2
	SVM	Simple molecular properties, molecular connectivity and shape, E-state, Q-C, GEO	577	Validation set (123)	77.8	92.7	89.4
Torsade de pointes causing agent	SVM	LSER	271	Validation set (78)	97.4	84.6	91.0

Abbreviations: **HIA** – human intestinal absorption; **P-gp** – p-glycoprotein; **BBB** – blood-brain barrier; **LDA** – linear discriminant analysis; **C-SAR** - classification structure-activity relations; **PNN** – probabilistic neural network; **SVM** – support vector machine; **ORMUCS** – ordered multicategorical classification method using the simplex technique; **MLR** – multiple linear regression; **PLS** – partial least squares; **PLS-DA** – partial least squares-discriminant analysis; **PCA** – principal component analysis; **ANN** – artificial neural network; **KNN** – k nearest neighbors; **TOPS-MODE** – topological substructural molecular design; **MR** – molar refractivity; **TOP** – topological; **E-state** – electrotopological state indices; **Q-C** – quantum-chemical; **GEO** – geometrical; **CON** – constitutional; **ELE** – electronic; **PSA** – polar surface area; **ADME** – absorption, distribution, metabolism, elimination; **MW** – molecular weight; **H-bond** – hydrogen bonding capabilities; **CPSA** – charged polar surface area; **LSER** – linear solvation energy relationship; **CV** – cross validation

[a] – number in parenthesis denotes the number of compounds used for model

Table 3. Performance of Regression-Based Statistical Learning Methods for Predicting Compounds of Specific Pharmacodynamic, Pharmacokinetic or Toxicological Property. The Relevant Literature References are Given in Supplementary Materials

Property	Activity	Method	Molecular descriptors	Validation method[a]	Reported prediction statistics
HIA	%FA	MLR	LSER	Training set (38) Validation set (131)	r^2=0.82, q^2=0.77, SE=15, F=53 RMSE=14, MAE=11
			Physicochemical, structural fragment	Training set (417) Validation set (50)	r^2=0.79, SE=12.34, F=38.83 r^2=0.79, SE=12.32
		Sigmoidal	PSA	Training set (20)	r^2=0.94, RMSE=9.2%
		PLS	Log P, molecular size, H-bond, counts	Training set (16) Validation set (63)	r^2=0.55, q^2=0.45 RMSE=28.6
			Atom type	Training set (169)	r^2=0.921, q^2=0.787
		ANN	TOP, ELE, GEO, CPSA, H-bond	Training set (67) Validation set (10)	RMSE=0.4, MAE=6.7 RMSE=16.0, MAE=11.0
			CON, TOP, chemical, GEO, Q-C	Training set (67) Validation set (10)	RMSE=0.590 r^2=0.802, RMSE=0.425
			TOP	Training set (396) Validation set (185)	r^2=0.92, RMSE=9.1, MAE=7.3 r^2=0.80, RMSE=11.8, MAE=9.8
		GRNN	Log P, MR, TOP	Training set (67) Validation set (10)	RMSE=6.5 RMSE=22.8
	FA	CART	Structural	Training set (899) Validation set 1 (362) Validation set 2 (67) Validation set 3 (90) Validation set 4 (37)	AAE=0.120 AAE=0.169 AAE=0.170 AAE=0.200 AAE=0.140
	logit(%FA)	PLS	MolSurf	Training set (13) Validation set (7)	r^2=0.903, q^2=0.685, RMSE=0.523 RMSE=0.488
			TOP	Training set (13) Validation set (7)	r^2=0.903, q^2=0.818, RMSE=0.523 RMSE=0.413
		SVR	Log P, MR, E-state	Training set Validation set	RMSE=0.445, MAE=0.404 RMSE=0.372, MAE=0.290
Bioavailability	%F	Regression	Substructure counts	Training set (591) 2000 runs of 80/20 splits (591)	r^2=0.71, q^2=0.63, RMSE=17.92 r^2=0.58, RMSE=20.40
		MLR	Bulk properties, solubility parameters, Q-C, CON, TOP	Training set (159) Validation set (10)	r^2=0.352, q^2=0.254 r^2=0.72
		ANN	CON, TOP, chemical, GEO, Q-C, bulk properties, solubility parameters	Training set (137) Validation set (15)	r^2=0.736, RMSE=19.21 r^2=0.680, RMSE=20.47
		CODES neural network	CODES	Training set (28)	q^2=0.90
P-gp inhibitor	log(1/EC_{50})	PLS	SIBAR	Training set (100)	r^2=0.731, q^2=0.661
BBB penetration	logBB	MLR	MW, log P	Training set (20)	r^2=0.691, SE=0.439, F=40.23
			LSER	Training set (57)	r^2=0.907, SE=0.197, F=99.2
			Solvation energy	Training set (55)	r^2=0.672, SE=0.41, F=108.3
			MW, log P	Training set (33)	r^2=0.897, SE=0.126, F=131.1
			H-bond	Training set (20)	r^2=0.723, SE=0.0012, F=46.93
			PSA	Training set (45)	r^2=0.841, F=229
			PSA, log P	Training set (55) Validation set 1 (5) Validation set 2 (5)	r^2=0.787, SE=0.354, F=95.8 MAE=0.14 MAE=0.24

(Table 3) contd.....

Property	Activity	Method	Molecular descriptors	Validation method[a]	Reported prediction statistics
			PSA	Training set (45)	r^2=0.95
			Solvation free energy	Training set (55) Validation set 1 (7) Validation set 2 (5) Validation set 3 (25)	r^2=0.72, SE=0.37 MAE=0.16 MAE=0.14 MAE=0.37
			MW, molecular lipoaffinity	Training set (55) Validation set (11)	r^2=0.790, q^2=0.763, SE=0.35, F=97.7 r^2=0.838, SE=0.30
			LSER	Training set 1 (148) 2 fold CV (148) 5 runs of 80/20 splits (148)	r^2=0.745, q^2=0.711, SE=0.343, F=69 r^2=0.718, SE=0.381 r^2=0.733, SE=0.356
			Hydrogen bonding, molecular volume, solvent-accessible surface area	Training set (76)	r^2=0.94, SE=0.173, F=311.307
			Spatial, structural, thermodynamic	Training set (59) Validation set (12) Validation set (21)	r^2=0.757, q^2=0.701, SE=0.408, F=42.135 RMSE=0.29 RMSE=0.50
			E-state	Training set (102) Validation set (20) 5 fold CV (102)	r^2=0.66, q^2=0.62, SE=0.45, F=62.4 RMSE=0.38, MAE=0.32 RMSE=0.47, MAE=0.38
			Solute aqueous dissolution and solvation, solute-membrane interaction, general intramolecular solute	Training set (56) Validation set (7)	r^2=0.845, q^2=0.795 RMSE=0.449, MAE=0.398
			Daylight, thermodynamic, spatial, structural, TOP, charge	Training set (48) Validation set (17)	r^2=0.837, q^2=0.786, MAE=0.26, SE=0.19 r^2=0.68, MAE=0.41
			Hydrophobicity, hydrophilicity, molecular bulkiness	Training set (78) Validation set 1 (13) Validation set 2 (22)	r^2=0.767, q^2=0.736, SE=0.364, F=81.5 r^2=0.88, RMSE=0.26, MAE=0.16 r^2=0.61, RMSE=0.48, MAE=0.39
			4D molecular similarity measures	Training set (104) Validation set (46)	r^2=0.69, q^2=0.64 r^2=0.56
			Physicochemical, GEO, structural, TOP	Training set (88) Validation set 1 (13) Validation set 2 (15)	r^2=0.864, q^2=0.847, SE=0.392, F=60.98 RMSE=0.558, MAE=0.407 RMSE=0.533, MAE=0.437
		Least-median-of-squares regression		Training set (86)	r^2=0.89, RMSE=0.31
		PCR	Log P, H-bond, PSA	Training set (61) Validation set 1 (14) Validation set 2 (25)	r^2=0.730, q^2=0.688, RMSE=0.424 r^2=0.576, RMSE=0.628 r^2=0.616, RMSE=0.789
			Atomic contributions to van der Waals surface area, log P, MR, partial charge	Training set (75)	r^2=0.83, q^2=0.73, RMSE=0.32
		PLS	MolSurf	Training set (28) Validation set 1 (28) Validation set 2 (6)	r^2=0.862, q^2=0.782, RMSE=0.288 RMSE=0.353 RMSE=0.473
			TOP, molecular volume, MW, CON, H-bond	Training set (58) Validation set 1 (12) Validation set 2 (22)	r^2=0.850, q^2=0.752, SE=0.318, F=102 RMSE=0.235 RMSE=0.408
			TOP	Training set (28) Validation set (30)	r^2=0.751, q^2=0.696, RMSE=0.368 RMSE=0.375
			Log P, MW, MR, molar volume, H-bond	Training set (19) Validation set (37)	r^2=0.905, q^2=0.791, RMSE=0.287 RMSE=0.338

(Table 3) contd.....

Property	Activity	Method	Molecular descriptors	Validation method[a]	Reported prediction statistics
			VolSurf	Training set (79)	r^2=0.78, q^2=0.65
			Log P, PSA, E-state	Training set (58) Validation set (39)	r^2=0.846, RMSE=0.308, MAE=0.232 r^2=0.617, RMSE=0.413, MAE=0.499
			Atom type	Training set (57) Validation set (13)	r^2=0.910, RMSE=0.502 RMSE=0.326
		CODES neural network	CODES	Training set (36)	q^2=0.88
		Bayesian neural net	Property-based, TOP indices, CIMI, atomic charges	Training set (106)	r^2=0.76, q^2=0.65, SE=0.54
		GRNN	DRAGON	Validation set (30)	r^2=0.701, RMSE=0.361
		SVR	Log P, MR, E-state	Training set Validation set	RMSE=0.242, MAE=0.200 RMSE=0.439, MAE=0.298
HSA binding	log Khsa	MLR	E-state	Training set (84) 10% CV (84) Validation set (10)	r^2=0.77, q^2=0.70, SE=0.29, F=43 r^2=0.68 r^2=0.74, RMSE=0.32, MAE=0.31
			ELE, TOP, information-content, spatial, structural, thermodynamic	Training set (84) Validation set (10)	r^2=0.78, q^2=0.73 r^2=0.88
		GRNN	DRAGON	Validation set (18)	r^2=0.851, RMSE=0.202
		SVR	CON, TOP, GEO, electrostatic, Q-C	Training set (84) Validation set (10)	r^2=0.94, RMSE=0.124 r^2=0.89, RMSE=0.222
Protein binding	log((1-fu)/fu)	MLR	Log P	Training set (226) Validation set (94)	r^2=0.68, MAE=0.45 r^2=0.51, MAE=0.53
	%fb	Nonlinear regression	Log P	Training set 1 (84) Training set 2 (44) Validation set (23)	r^2=0.803, MAE=0.104 r^2=0.786, MAE=0.055 r^2=0.830
	fb	ANN	Atom and functional group counts, connectivity index differences, connectivity index quotients, charge indices, vertex counts, ramifications, Wiener number, MW, Log P	Validation set (6)	r^2=0.745
Milk-plasma ratio	M/P	ANN	CON, TOP, molecular connectivity, GEO, Q-C, physicochemical, liquid properties	Training set (123)	r^2=0.61, RMSE=0.781
		GRNN	DRAGON	Validation set (20)	r^2=0.677, RMSE=0.454
Total clearance	CL	KNN	TOP, physical properties, partial charge, pharmacophore feature, potential energy	Training set (32) Validation set (6)	q^2=0.77 r^2=0.94
		ANN	Atom and functional group counts, connectivity index differences, connectivity index quotients, charge indices, vertex counts, ramifications, Wiener number, MW, Log P	Validation set (6)	r^2=0.731
		GRNN	Lipophilicity, ionization, molecular size, H-bond	Training set (23)	r^2=0.775, q^2=0.731

Abbreviations: **HSA** – human serum albumin; **FA** – fraction absorbed; **F** – bioavailability; **BB** – ratio of concentration of drug in brain to concentration of drug in blood; **Khsa** – binding affinity of drug to human serum albumin; **fu** – fraction of drug unbound in plasma; **fb** – fraction of drug bound in plasma; **M/P** – ratio of concentration of drug in milk to concentration of drug in plasma; **CL** – total clearance; **GRNN** – general regression neural network; **CART** – classification regression tree; **SVR** – support vector regression; **PCR** – principal component regression; **SIBAR** – similarity based structure activity relationship; **CIMI** – chemically intuitive molecular index; **3DMoRSE** – 3D molecule representation of structures based on electron diffraction; **ATS** – Moreau-Broto autocorrelation; **GETAWAY** - geometry, topology, and atom-weights assembly; **RDF** – radial distribution function; **WHIM** – weighted holistic invariant molecular descriptors

[a] – number in parenthesis denotes the number of compounds used for model validation.

computed activities and experimentally estimated activities. Moreover, q^2 values, RMSE values and average-fold errors for an independent validation set are also frequently computed to further evaluate the predictive capability of these studies. The computed r^2 values range from 0.68 to 0.94 [51, 52], which is compared to the range of 0.51 to 0.88 in the conventional QSAR and QSPR studies [53, 54]. These suggest that regression-based methods are useful for predicting the activity values of compounds of particular property at accuracy levels comparable to conventional QSAR and QSPR methods.

CONCLUSIONS AND PERSPECTIVES

Both classification- and regression-based statistical learning methods consistently show promising capability for predicting compounds of diverse ranges of structures and of a wide variety of pharmacodynamic, pharmacokinetic, and toxicological properties. Classification-based methods are useful for the prediction of classes of compounds with few or no quantitative activity data. Regression-based methods can be used for quantitative prediction of the activity levels if the activity data are available for a sufficient number of compounds possessing the same property. Regression methods have the capacity for estimating the contribution of specific structural and physicochemical features of the compounds to a particular property [55]. This capacity may be explored for probing the mechanism of action for a specific group of compounds that possess a particular property.

In general, a sufficiently diverse set of positive compounds (known to have a property) and negative compounds (known to not have a property) is needed for training a statistical learning system. Thus statistical learning methods are not applicable for compounds with little or no knowledge about their particular pharmacodynamic, pharmacokinetic or toxicological property. Mining of the compounds known to have a particular property and those do not have that property from the literature [56] and other sources [57, 58] is a key to more extensive exploration of statistical learning methods. Databases such as PDSP K_i database [59], KiBank [60], PubChem [61], and CLiBE [62] that provide compound property and activity data are useful resources for serving this purpose, and more such databases are desired.

ACKNOWLEDGEMENT

This work was supported in part by grants from Singapore ARF R-151-000-031-112, Shanghai Commission for Science and Technology (04DZ19850), and the '973' National Key Basic Research Program of China (2004CB720103).

REFERENCES

[1] Drews, J. *Science* **2000,** *287*, 1960.
[2] Park, B. K.; Kitteringham, N. R.; Powell, H.; Pirmohamed, M. *Toxicology* **2000,** *153*, 39.
[3] Caldwell, J.; Gardner, I.; Swales, N. *Toxicol. Pathol.* **1995,** *23*, 102.
[4] White, R. E. *Annu. Rev. Pharmacol. Toxicol*. **2000,** *40*, 133.
[5] Ekins, S.; Ring, B. J.; Grace, J.; McRobie-Belle, D. J.; Wrighton, S. A. *J. Pharmacol. Toxicol. Methods* **2000,** *44*, 313.
[6] Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A. *Bioorg. Med. Chem.* **2004,** *12*, 3391.
[7] Katritzky, A. R.; Karelson, M.; Lobanov, V. *Pure Appl. Chem.* **1997,** *69*, 245.
[8] Manallack, D. T.; Livingstone, D. J. *Eur. J. Med. Chem.* **1999,** *34*, 195.
[9] van de Waterbeemd, H.; Gifford, E. *Nat. Rev. Drug Discov.* **2003,** *2*, 192.
[10] Trotter, M. W. B.; Holden, S. B. *QSAR Comb. Sci.* **2003,** *22*, 533.
[11] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2001,** *26*, 5.
[12] Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *Dragon*, Version 5.3; 2005.
[13] Hall, L. H.; Kellogg, G. E.; Haney, D. N. *Molconn-Z*, Version 4.05+; eduSoft, LC: 2002.
[14] Wegner, J. K. *JOELib/JOELib2*, 2005.
[15] Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 1630.
[16] Perez, J. J. *Chem. Soc. Rev.* **2005,** *34*, 143.
[17] Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998,** *38*, 983.
[18] Molnar, L.; Keseru, G. M. *Bioorg. Med. Chem. Lett.* **2002,** *12*, 419.
[19] Potter, T.; Matter, H. *J. Med. Chem.* **1998,** *41*, 478.
[20] Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectrosc.* **1999,** *19*, 151.
[21] Rücker, G.; Rücker, C. *J. Chem. Inf. Comput. Sci.* **1993,** *33*, 683.
[22] Schuur, J. H.; Setzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996,** *36*, 334.
[23] Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.* **1999,** *39*, 28.
[24] Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. *J. Comput. Aided Mol. Des.* **1997,** *11*, 79.
[25] Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1994,** *34*, 520.
[26] Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002,** *42*, 682.
[27] Randic, M. *Tetrahedron* **1975,** *31*, 1477.
[28] Randic, M. *New J. Chem.* **1995,** *19*, 781.
[29] Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*. Academic Press: San Diego, **1999**.
[30] Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **1999,** *39*, 835.
[31] Lucasius, C. B.; Kateman, G. *Chemom. Intell. Lab. Sys.* **1993,** *19*, 1.
[32] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. *Mach. Learn.* **2002,** *46*, 389.
[33] Sutter, J. M.; H., K. J. *Microchem. J.* **1993,** *47*, 60.
[34] Yu, H.; Yang, J.; Wang, W.; Han, J. *Proceeding of the IEEE Computer Society Bioinformatics Conference (CSB)*, **2003**, 220.
[35] Yap, C. W.; Cai, C. Z.; Xue, Y.; Chen, Y. Z. *Toxicol. Sci.* **2004,** *79*, 170.
[36] Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 1497.
[37] Yap, C. W.; Chen, Y. Z. *J. Pharm. Sci.* **2005,** *94*, 153.
[38] Li, H.; Xue, Y.; Ung, C. Y.; Yap, C. W.; Li, Z. R.; Chen, Y. Z. *Chem. Res. Toxicol.* **2005,** *18*, 1071.
[39] Gramatica, P.; Pilutti, P.; Papa, E. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 1794.
[40] Izrailev, S.; Agrafiotis, D. K. *J. Mol. Graph. Mod.* **2004,** *22*, 275.
[41] Yap, C. W.; Chen, Y. Z. *J. Chem. Inf. Model.* **2005,** *45*, 982.
[42] Hosmer, D. W.; Lemeshow, S. *Applied logistic regression*. 2nd ed.; John Wiley & Sons: New York, **2000**; p 373.
[43] Huberty, C. J., *Applied Discriminant Analysis*. John Wiley & Sons: New York, **1994**.
[44] Johnson, R. A.; Wichern, D. W. *Applied multivariate statistical analysis*. Prentice Hall: Englewood Cliffs, NJ, **1982**.
[45] Fix, E.; Hodges, J. L. *Discriminatory analysis: Non-parametric discrimination: Consistency properties*; 4; USAF School of Aviation Medicine, Randolph Field: Texas, **1951**; pp. 261.
[46] Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann: San Mateo, Calif, **1993**.
[47] Specht, D. F. *Neural Netw.* **1990,** *3*, 109.

[48] Vapnik, V. N., *The Nature of* statistical learning theory. Springer: New York, **1995**.
[49] Grover, M.; Singh, B.; Bakshi, M.; Singh, S. *Pharm. Sci. Technol. Today* **2000,** *3*, 50.
[50] Kubinyi, H. *Drug Discov. Today* **1997,** *2*, 538.
[51] Turner, J. V.; Maddalena, D. J.; Agatonovic-Kustrin, S. *Pharm. Res.* **2004,** *21*, 68.
[52] Ng, C.; Xiao, Y. D.; Putnam, W.; Lum, B.; Tropsha, A. *J. Pharm. Sci.* **2004,** *93*, 2535.
[53] Hou, T. J.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **2003,** *43*, 2137.
[54] Lobell, M.; Sivarajah, V. *Mol. Divers.* **2003,** *7*, 69.
[55] Stanton, D. T. *J. Chem. Inf. Comput. Sci.* **2004,** *43*, 1423.
[56] PubMed. http://www.pubmed.gov
[57] MICROMEDEX *MICROMEDEX*, MICROMEDEX: Greenwood Village, Colorado, Edition expires 12/2003.
[58] Bethesda, *AHFS drug information*. American Society of Health-System Pharmacists, Inc: **2001**.
[59] Roth, B. L.; Kroeze, W. K.; Patel, S.; Lopez, E. *The Neuroscientist* **2000,** *6*, 252.
[60] Zhang, J.-W.; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata, K. *Comput. Biol. Chem.* **2004,** *28*, 401.
[61] PubChem. http://pubchem.ncbi.nlm.nih.gov
[62] Chen, X.; Ji, Z. L.; Zhi, D. G.; Chen, Y. Z. *Comput. Chem.* **2002,** *26*, 661.
[63] Oprea, T. I.; Gottfries, J. *J. Comb. Chem.* **2001,** *3*, 157.
[64] NCI/NIH Developmental Therapeutics Program. http://dtp.nci.nih.gov/index.html (5 July 2005),
[65] Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. *J. Chem. Inf. Comput. Sci.* **2003,** *43*, 1906.

Received: July 15, 2005 Revised: September 21, 2005 Accepted: September 22, 2005