New Predictors for Several ADME/Tox Properties: Aqueous Solubility, Human Oral Absorption, and Ames Genotoxicity Using Topological Descriptors.

Joseph R. Votano*, Marc Parham, ChemSilico LLC, Tewksbury, MA 01876,
Lowell H. Hall, Department of Chemistry, Eastern Nazarene College, Quincy, MA 02170, and
Lemont B. Kier, Department of Medicinal Chemistry, School of Pharmacy, Virginia Commonwealth University, Richmond, VA 23298.

ABSTRACT

In silico predictive models for aqueous solubility, human intestinal absorption (HIA), and Ames genotoxicity were developed principally using artificial neural net (ANN) analysis and topological descriptors. Approximately 10,000 compounds spread across three datasets were used in the construction of these quantitative-structure/property relationships (QSAR/QSPR) models. For aqueous solubility, 5,037 chemically diverse compounds were used to construct ANN-QSPRs for intrinsic aqueous solubility. When these robust models were applied to 938 compounds in external validation, they gave an $r^2 = 0.78$ with 84% predicted within 1 log unit for these new chemical entities (NCEs). 417 therapeutic drugs were used in the development of an ANN-QSPR to predict for percent oral absorption (%OA). For validation testing on 195 new drugs, 92% of the compounds were predicted to within 25% of their reported %OA values, which ranged from 0% to 100%. Polar surface area and logP, the octanol-water partition coefficient, were found to be important descriptors in our QSPR model. Development of an ANN-QSAR as a genotoxicity predictor for *S. Typhimurium* employed 2963 compounds including 290 therapeutic drugs. Validation results on 400 NCEs with the ANN-QSAR gave a concordance of 83% which rose to 91% when a confidence indictor was applied. With new drugs a concordance of 92% was reached, which increased to 97% when the reliably indicator was invoked.

Key words: aqueous solubility, genotoxicity, drug absorption, QSAR models, topological descriptors

Introduction

Topological descriptors by themselves or in combination with other variables have become for many, the mainstay in the development of quantitative-structure-activity/property-relationship (QSAR/QSPR) models. Their frequent use is due to their ability to assist in establishing meaningful structure-activity and structure-property relationships between compound structures and their end-points in a host of different physiochemical and biological processes. ADMET (adsorption, distribution, metabolism, excretion, and toxicity) properties is one area where they have started to be employed. Over the last several years, efforts in the pharmaceutical R&D area have stressed the need for better and wider use of *in silico* predictive tools for assessing ADMET properties of drug-like compounds in the early discovery and enhancement stages in the drug discovery process. Similarly, these same considerations apply to human and environmental research with respect to industrial chemicals.

Aqueous solubility, human intestinal absorption, and genotoxic potency are three important ADMET properties for drug efficacy. Equally, they apply to exposure of humans and the environmental fate of xenobiotics. In pharmacokinetics, aqueous solubility has a pronounced affect on the pharmacological activity of a compound in terms of its uptake, distribution, and ultimately its bioavailability. Human intestinal absorption (HIA) is the first litmus test for an orally taken therapeutic. Without moderate to high intestinal absorption, the therapeutic effect of drugs can appreciably diminish. Long or even short-term exposure to xenobiotics brings into play toxicity concerns. Here we focus specifically on genotoxicity. In this study, genotoxicity refers to Ames mutagenicity in *S. Typhiumurium*, a standard test requirement in drug submission to regulatory bodies.

Topological descriptors, by themselves or in combination with others, have been reported in QSAR models to predict aqueous solubility [1-5], genotoxicity [6,7], and in several HIA studies [8,9]. Here we present results on QSAR models for aqueous

* Author to whom correspondence should be address:
JVotano@ChemSilico.com

solubility (S), HIA, and Ames genotoxicity. All models employed large, chemically diverse datasets in their development with strict use of topological descriptors with exception in the HIA-QSAR model with use of logP and TPSA, the topological polar surface area. Comparison of our results on one ADMET property is made with that of a linear model using fragment or group descriptors.

Methods

Data. *Aqueous solubility* data for intrinsic water solubility (So), defined as solubility for an uncharged compound in water usually at 25°C, came from various sources described elsewhere [5] and the addition of 80 predominately charged compounds provided [10] from the recent literature. A database was constructed from all these sources, examined for duplicate molecular structures and experimental values. Duplications in structure entries were determined by comparing both the sum of all computed descriptor values and molecular weight values. For experimental data, if duplicates did not agree within 0.4 in log units, both were discarded; otherwise, the lower of two values was kept. 5975 compounds were used for a final dataset. Approximately 25% of the aromatic members contained nitrogen-heteroaromatics rings of which 33% were polycyclic systems. Close to 33% of the non-aromatics contained at least one ring. 938 compounds from the 5975 data set were used only in validation testing and not involved in construction of the QSAR models.

HIA, percent oral absorption (%OA), data came from five sources as listed elsewhere [11]. Compounds were carefully scrutinized to remove those that are reported to be actively transported or underwent facilitated transport. 612 therapeutic drugs were found suitable for analysis. Modeling building employed 417 compounds and 195 randomly selected drugs from initial dataset were set aside for validation testing. The compounds were segregated into two classes, those with molecular weight (MW) less than 251 and those with higher MWs. Twenty-three (23) compounds with a formal positive charge were treated as special class.

Ames mutagenicity endpoints came from numerous sources described elsewhere [7]. The dataset contained 3363 compounds clearly reported as mutagens and non-mutagens and tested in at least two strains of *S. Typhiumurium* of five (TA97, TA98, TA100, TA102, TA104, TA1535 and TA1537) commonly used. A randomly selected validation set of 400, 12% of the initial dataset, were set aside for validation testing.

Model building. Analysis of data and model building mainly focused on use of artificial neural net (ANN) analysis on datasets for aqueous solubility, human intestinal absorption, and genotoxicity. Multiple linear regression (MLR) was applied as well to the genotoxicity dataset. Receiver operator curve construction was done using Analyse-It [12]. Final descriptor selection for MLR was performed using a genetic algorithm in the QsarIS software [13]. The genetic algorithm (GA) employed r^2 optimization and qualified by the reciprocal of the Friedman's lack-of-fitness function [14]. All highly correlated variables (r^2 >0.8) were removed to eliminate redundant information. MLR analysis was accomplished with the QsarIS software. In the MLR modeling process, both forward and backward regression was used to assess any substantial changes in the statistical outcomes for the addition or removal of a descriptor. No improvements were found. Goodness of fit was determined by r^2, q^2, and the F statistic with all parameters accepted at the 95% confidence level. A 100-fold randomization of mutagenicity index values, 0 or 1, was performed with r^2 computed for each case, (standard method in QsarIS), yielding an average r^2 less than 0.02 in the MLR-GA model. The results of this randomization method indicate that the model is different from an equation based on random numbers, indicating that significant information is contained in the model.

ANN analysis was performed on a train/test dataset with a certain percentage set aside for external validation. The train/test set, designated the principal set, was split into 85% for train and 15% as a selection set for early stopping of the learning process to avoid over fitting. 90% of the train set was selected 10 times in 10 folds of data and each fold accompanied by 10% of train as a mutually exclusive withholding set where each compound (row) appeared only once in each withheld set. This multiple selection process gives a set of 10 ANN models derived from the principal set using mutually exclusive test sets. Using this approach, the non-contributory variables are pruned to give an optimal subset of significant variables. The relative importance of each eliminated variable is based on its contribution across the entire principal set by calculation of r^2 in each instance when the row (compound) appears in the withholding test set. This value is designated q^2, that is, the r^2 value for all instances when the data was withheld from the modeling process. Since q^2 is used to select the variables, it does not provide a completely reliable assessment of the predictive accuracy of the overall algorithm. This task is reserved for a validation set. The standard back propagation network is used with no more than 9 hidden neurons, using the backward elimination approach [14, 15] adapted from traditional neural network approaches. The ten-fold cross-validation algorithm is used as a consensus model in which the average value of ten neural nets gives the predicted ADMET property for a compound.

Descriptor sets. The complete set of topological descriptors (ChemSilico) contained 542 members e.g., atom-type E-State and hydrogen-E-States, molecular connectivity, bond -E-States, binary indicators, atom counts. This set of 542 descriptors was reduced to a lower number for each dataset based on the criterion that, at least, 5% of compounds in a dataset must be non-constant (usually non-zero). In the case of aqueous solubility, the reduction was to 128 descriptors and 160 respectively for the non-aromatic and aromatic datasets. The genotoxicity dataset gave 148 and HIA 141. CSLogP [16], the octanol-water partition coefficient, and TPSA, topological polar surface area [17], were included as members in the HIA descriptor set.

Formulation for aqueous solubility as function of pH

Aqueous solubility as a function of pH for monoprotic, monobasic, and ampholytes can be computed with use of S_o and pK_as of the compound. The pK_a values are computed from CSpKa [16], which uses only the 2D structure of the compound. An example formulation for logS for an acid-base ampholyte is:

$$logS = logS_o + log[1 + 10^{(pH - pK_{a1})} + 10^{(pK_{a2} - pH)}]$$

where pK_{a1} and pK_{a2} are pK_{as} for the acid and base respectively. When the pH>>pK_a exists for acids or the reverse for bases, salt precipitation takes effect for a given K_{sp} (the solubility product) at the saturation concentration where logS = logSp, a constant. A rule thumb can be used to estimate logSp [18] for acid and bases in a 0.15M in NaCl. Using a cut-off value, logSp, is essential to arrive at a meaningful logS value at extreme pH values, since the logS formulations (based on equilibrium relations) are, by their nature, not self-limited in low (<3) or high (>9) pH regions.

Results

Aqueous solubility. In development of in silico predictive models for intrinsic aqueous solubility the initial dataset of 5975 compounds was sub-divided into two classes; aromatic (3343) and non-aromatic (1674). 518 compounds were selected randomly for validation from the initial set exclusive of the 420 diverse entities provided [10] for validation testing. Results from the two ANN-QSARs models, one for aromatic moieties and other for non-aromatics, are given in Table 1.

Table 1. ANN Statistical Parameters for Aromatic and Non-Aromatic Datasets[a]

Model	Nv	N	Type	r^2	MAE	RMSE	%Cpds[b] (AE<=0.5)	%Cpds (AE<=1)
ANN	47	3363	aromatic-train	0.88	0.51	0.74	63	88
ANN	35	1674	non-aromatic-train	0.88	0.44	0.61	67	93
ANN	47	772	aromatic-validation	0.77	0.62	0.89	58	82
ANN	35	166	non-aromatic-validation	0.84	0.56	0.75	55	86
CSLogWS[c]	47	80	charged aromatic-validation	0.77	0.57	0.66	44	91
ACD v.6.0[d]	fragments	74[e]	charged aromatic-validation	0.05	1.27	1.60	31	46

[a] ANN: artificial neural net; Nv = number of variables; N: number of compounds; r2: square of correlation coefficient;
MAE: mean absolute error S(| logSocalc - logSoexp |)/N; RMSE = sqrt(S(| logSocalc - logSoexp |)^2 / N).
[b]Percentage of compounds below an AE value.
[c]CSLogWS, version 1.5 built using only topological descriptors for estimations of pKas and employs the ANN models in this study.
[d]ACD Labs, version 6.0 software;
[e]Six compounds unable to process, all parameters based on 74 compounds used.

With respect to the training sets, an r^2 of 0.88 respectively was found for both classes with slightly better fitting of So with the smaller non-aromatic set. More importantly both models calculated, on average, 65% of the compounds with an absolute error (AE) below or equal to 0.5 log units and 90% of them at an AE<1. With validation compounds the mean absolute error, MAE, was 0.62 for 772 aromatics and 0.56 for 166 non-aromatics. These numbers translate into an average of 57% of the predicted $logS_o$ values having AE equal to or below 0.5 and 84% within 1 log unit. The robustness in predictive capability of these ANN models can be seen in Fig 1, where reasonably tight clustering is apparent for 938 validation compounds around the diagonal, $log(1/S_o)_{exp}$.

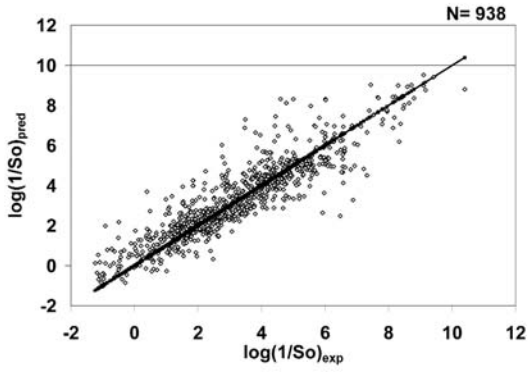

Figure 1. Correlation of log(1/So) for predicted versus observed intrinsic aqueous solubilities for 938 aromatic and non-aromatic compounds in the validation set. Predicted log(1/So) values are results from an ANN analysis using 47 and 35 variables respectively.

Table 2. Rank and Frequency of 10 Important Descriptors in ANN[a] Model

Non-Aromatic Compounds (train)			Aromatic Compounds (train)		
Variable	***$Rank_N$***	***Frequency***	***Variable***	***$Rank_N$***	***Frequency***
EPSA[b]	1.64	1460	SsssN	1.54	667
SsssN	1.54	174	SHCarOH1	1.47	371
SHBint2	1.49	598	ka2	1.41	3343
Gmax	1.48	1674	xv0	1.40	3343
x1	1.46	1674	Narom	1.36	835
Hmax	1.43	1674	dxp3	1.36	3343
SssNH	1.40	330	SHPheOH1	1.34	427
numHBa	1.39	1544	SHBint2	1.32	1128
Qv	1.34	1674	SHCsats	1.26	1783
SCarOH1	1.27	301	SPheOH1	1.21	427

[a] RankN: normalized rank determined as the ratio of the difference in RSS(sum of squares of residuals) in the presence and absence of the variable divided by the same difference for the least important variable where its value is the average across all 10 all 10 ANN models; Frequency: number compounds expressing the variable.
[b]**Descriptor definitions:**
dxp3: difference chi path 3 minimizes dependency on molecular size and encodes branching and path for three sequential bonds.
EPSA: sum of atom E-States for O, N, P, and S.
Gmax: maximum E-state of an atom in the molecule.
Hmax: maximum hydrogen atom E-state in molecule.
Ka2: 2nd order shape descriptor, which encodes the degree globularity of the molecule- more branching the smaller Ka2.
Narom: binary indictor for an N-heteroaromatic ring.
NumHBa: number of hydrogen bond acceptors in molecule.
Qv: polarity index of the molecule- the smaller its value, greater the polarity.
SCarOH1: sum of the O atom E-States for carboxylic acids.
SHCsats: sum E-States of hydrogen attached to sp3 carbons.
SHBint2: maximum product E-States for H-acceptor and donor atoms separated by two non-hydrogen bonds.
SHPheOH1: sum of hydrogen E-States for phenolic hydrogens.
SPheOH1: sum of O atom E-States of phenolic OHs in molecule.
SssNH: sum of secondary Ns atom E-States in the molecule.
SsssN: sum of all atom-E-states for all tertiary Ns in the molecule.
x1: simple chi indice-encode branching, decreases with branching.
xv0: encodes for molecular size and heteroatom contribution.

A listing in Table 2 gives the ten most important descriptors for each ANN model and their frequency. All descriptors were represented in compounds in varying amounts, anywhere from 10% to 100%. About 50% of the most important QSAR variables for the aromatic set have a (non-zero) occupancy of less than 20%; 40% of important descriptors in the QSAR for the non-aromatics with (non-zero occupancy) less than 25%. Four atom E-States, EPSA, SsssN, SssNH, and SCarOH1, were found among the most important variables for non-aromatics. Whereas with aromatics three hydrogen-E-States indices, SHCarOH1, SHPheOH1, and SHCsats, were the predominate members accompanied by non-hydrogen atom-E-States, SsssN and SPheOH1. Molecular shape, size, and branching were important structural attributes for aromatics via use of Ka2,

xv0, and dxp3 indices. These important shape and connectivity indices address, in part, the highly diverse structures present in the aromatic set. Some structures are compact, others highly extended with molecular weight ranging from 67 to 1140. We also examined charged compounds to see how well the ANN models presented here (together with pK_a models [16]), based solely on topological descriptors for estimating pK_a, perform when compared to an additivity approach employing principally molecular fragments for both So and pKa. Eighty (80) charged aromatic compounds with aqueous solubilities determined at pH 7.4 were examined. The data set composition is 31 cations, 24 anions, and 25 zwitterions at pH 7.4. They were run through CSLogWS [16] and ACD ver. 6.0 [19]. The former software contains the ANN models presented here. The results for the fragment approach based program were triple checked. As seen in Table 1, topological variables in combination with the ANN learning algorithm perform substantially better than the use of fragments. A similar finding has been reported earlier [20].

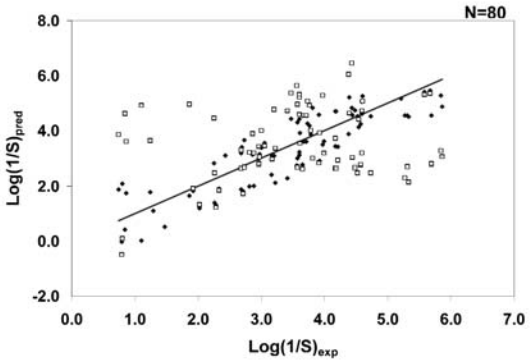

Figure 2. Aqueous solubility of charged compounds predicted by two QSARs, CSLogWS (black-filled diamonds) and ACD Labs version 6.0 (open boxes). Black diagonal is log(1/S0)exp values.

There is great deal of spread in the predicted values of log(1/S0) using fragments (open boxes) as seen in Figure 2 as compared to the predicted values (filled diamonds) from the ANN-QSARs using topological descriptors. Six representative structures from the 80 compound dataset are shown in Figure 3.

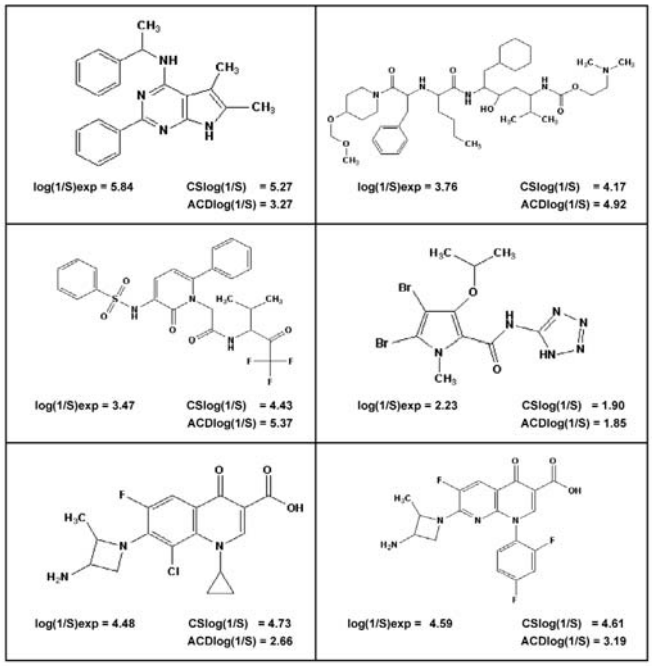

Figure 3. Six representative charged composed from the 80 compounds validate set. Attached to each structure are results from both CSlogWS and ACD Labs aqueous solubility predictor.

Human intestinal absorption

Human oral absorption is presented as the percent oral absorption (%OA), which ranges from 0% to 100% in their reported values. Since all 612 compounds studied are oral therapeutic drugs, it is not too surprising a large percentage should have high %OA values as shown in Figure 4A.

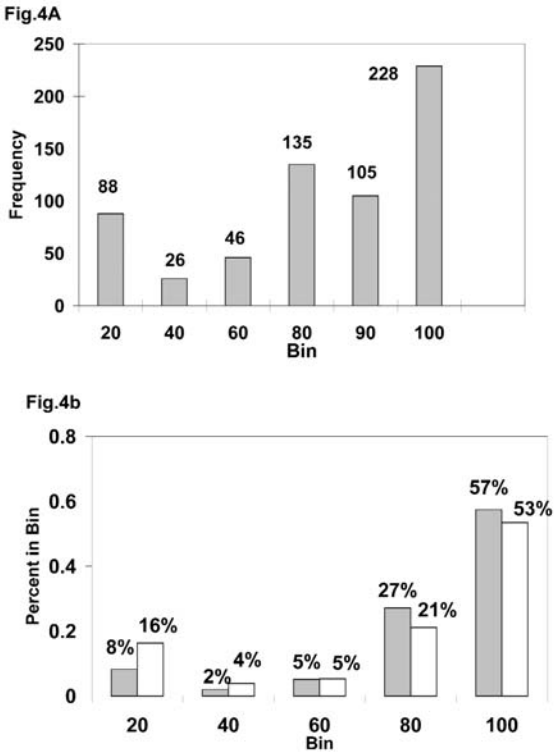

Figure 4. Upper figure, 4A, shows the overall percentages of compounds with %OA values binned in 20% intervals. Lower figure, 4B, is the final disposition of the percent of absorbers in train and validate sets over 20% intervals where empty rectangles represent train set compounds.

Fifty-four percent (54%) of the drugs had a %OA greater than 80% and 46% were below 80%. However, 25% of the compounds had OA values below 60%, which diminished the biasing in %OA in the use of a large number of compounds with high values. Compounds in the HIA dataset were selected randomly, except that a higher percent of poor absorbers were placed into the train set to somewhat balance the influence of the high percentage of good absorbers. Final percentages of train and validation members with %OA values in 20% intervals are shown in Figure 4B.

One ANN model was constructed to address low molecular weight (MW) compounds (90 compounds with MW<251) that undergo passive paracellular transport. A second ANN-QSAR used 308 compounds involved in transcellular passage with molecular weights in the range from 252 to 1450. Combined results for two ANN models are shown in Table 3 along with the 23 charged compounds modeled with MLR (but not discussed here).

Table 3. Statistical parameters for ANN models for % oral absorption[a]

N	Type	r^2	MAE	RMSE	%Fraction
417	train	0.88	8.5	11.5	100%
407	train	0.91	7.8	9.9	98%
396	train	0.92	7.3	9.1	95%
195	validation	0.71	11.2	15.9	100%
191	validation	0.77	10.5	14.5	98%
185	validation	0.80	9.8	11.8	95%

[a]N is the number of compounds; % fraction is the percent fraction of compounds remaining after removal of noisy compounds. 32 unique desciptors were used in the two ANN models.

The 417 member train sets gave an $\langle r^2 \rangle$ 0.88 and MAE = 8.5. For 195 new drug entities, their validation statistics are as follows: r^2 = 0.71, MAE and RMSE values of 11% and 16%. Human intestinal data is known to be noisy due to difficulties associated with making these determinations with a high degree of precision and accuracy. About 5% of our dataset did appear noisy for various reasons not discussed here. With removal of 5% of the worst predicted compounds in the validation set, a decrease of approximately 13% in MAE and 26% in RSME was found as shown in Table 3. Scatter in predicted %OA values can be seen in Figure 5 where dashed diagonals on either side of the principle diagonal (%QAexp) indicate 25% error.

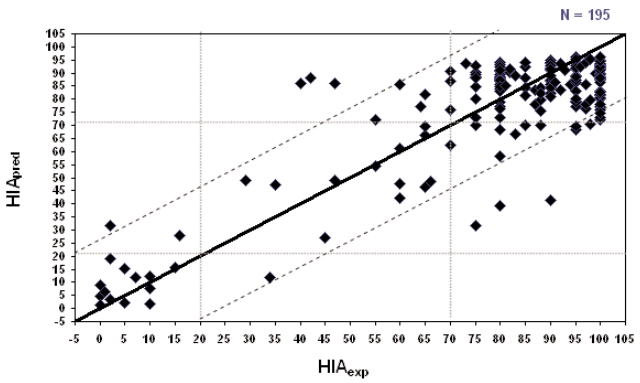

Figure 5. ANN predictive results for 195 compound validation set. Dashed lines indicate a 25% error below and above the black diagonal, the experimental values. Filled-in diamonds are the predicted end-point values for %OA.

Nonetheless, the percentage of the predictions within these error lines is 92% for the 195 drugs, which is indicative of quality of the models, in spite of unavoidable noise associated with oral absorption data. Of interest are the ten most important descriptors given in Table 4 for the 308 train compounds, which undergo passive transcellular transport.

Table 4. ANN descriptors for transcellular transport

Variable	Trend	Rank	Frequency
SHBint4[a]	+	1.55	59
ArNH1	+	1.44	52
SHBint2	-	1.33	146
CSLogP	+	1.28	308
SCarOH1	-	1.27	57
e1C3O1d	-	1.26	66
Qv	+	1.25	308
Hmax	-	1.19	308
SssO	-	1.19	147
TPSA	-	1.16	308

[a]**Descriptor definitions**:
ArNH1: sum of the N atom-E-States for secondary amines attached to aromatic ring.
CSLogP: octanol-water partition coefficient
e1C3O1d: bond atom E-State for C-O in carboxylic acid
SHBint4: sum of the products of donor and acceptor atom E-States separated by 4 skeletal bonds.
SssO: sum of oxygen atom-Estates for -O-.
TPSA: topological polar surface area.

Hydrophilicity plays a strong role in cellular permeability so it was not too surprising to find logP (CSLogP) ranked fourth in importance. Likewise, the polar surface area, PSA, of the molecule has also been shown to be important [21-24]. It ranked in the top ten descriptors in significance. Some of the remaining variables in Table 4 are the sum or product of atom E-States, polarity (Qv), and Hmax (maximum hydrogen E-State). One bond E-State, e1C3O1d, is mostly associated here with carboxylic acid >C=O groups. The most important descriptor, SHBint4, encodes internal hydrogen bonding for acceptor and donors atoms separated by four skeletal bonds.

Genotoxicity

Two different modeling techniques, ANN and MLR, were applied to the genotoxicity dataset to predict the mutagenicity index, a binary indicator: 1 for a mutagen and 0 for a non-mutagen. The outcomes from these models are continuous values from 0 to 1. A cutoff-threshold was used where a mutagenicity index equal to or greater than 0.5 was considered indicative of a mutagenic compound for Ames mutagenicity. A confidence level for a prediction by the consensus based ANN model was 0.27, which is the standard deviation (SD) of the mean value coming from the 10 models. Outcomes with SDs outside 0.27 are considered as unreliable. Special attention was given to therapeutic drugs; 329 were used with 290 in the train set and 12% in the validation set. As shown in Table 5, the ANN and MLR-GA models employed 38 and 39 topological variables to model 2963 highly diverse compounds. The ANN model delivered a somewhat better train concordance, 89%, versus 82% for the MLR-GA model.

Table 5. Statistical parameters on prediction of Ames genotoxicity by ANN and MLR-GA models[a]

Model	N	Nv	ROC	Concordance (train)	%FPos (train)	%FNeg (train)	N	ROC	Concordance (validate)	%FPos (validate)	%FNeg (validate)
ANN	2963	38	0.96	89%	2%	9%	400(319)	0.93	83(91%)	4(3%)	13(6%)
MLR-GA	2963	39	0.89	82%	9%	9%	400	0.89	81%	11%	8%
				Therapeutic Drugs					Therapeutic Drugs		
ANN	290			93%	4%	3%	39(34)		92(97%)	5(0%)	3(3%)
MLR-GA	290			69%	22%	9%	39		56%	36%	8%

[a] Dataset split into 2963 compounds for train/test and 400 for validation testing; Validate values in parenthesis are considered reliable outcomes in the ANN model; N: number of compounds; Nv: number of descriptors used in the model; ROC: area under the receiver operator characteristic curve; %FPos and %FNeg: percentages of Ames false positives and negatives.

In validation testing, both models gave very similar concordance values, 83 and 81%, for the 400 compounds accompanied by reasonably low percentages in false positives and negatives. However, small differences in the ROC values in Table 5 indicate differences in sensitivity with these two models. A ROC (receiver operator characteristic curve) value checks whether the threshold of 0.50 used here to distinguish a mutagen from non-mutagen, results in high sensitivity; a minimum number of false positives and false negatives. This threshold was suitable since the ROC values were high (>0.85) for both models. Likewise, no appreciable changes in concordance occurred until it was set too low, below 0.45, or too high, above 0.6. In the ROC analysis, the closer the area under the curve is to 1.0, the greater the predictive ability of the model for a given threshold value.

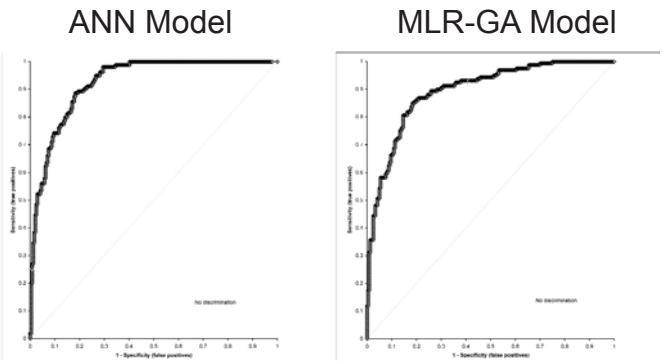

Figure 6. Receiver operator curves (ROCs) for the ANN model (left curve) and MLR-GA model (right curve). Straight-line, dotted diagonal represents outcomes when a model has no discriminatory power in predicting binary outcomes

In Figure 6 are two ROCs for these models based on their continuous output values from 0 to 1.0 and experimental mutagenicity index values, 0 or 1, for the 400 validation compounds. It is apparent that the shapes do differ. The ANN model curve is nearly L-shaped with its area closer to 1.0 than is the MLR-GA curve. Since the differences in sensitivity to detect the presence or absence of toxicophores (substructures) are reflections of these small differences in ROC values, we examined results coming from both models for therapeutic drugs.

As shown in Table 5, the difference in robustness in the concordance values between the ANN and MLR-GA QSARs is substantial. The MLR-GA gives almost random results, 59%, in its accuracy with the drug validation set; whereas the ANN QSAR yields a very high concordance and low percentages of false positives and negatives.

When the ANN-QSAR confidence level indicator, 0.27, was applied to both the validate and therapeutic drug sets, the concordance increased to 91% for the validate set with exclusion of 20% of compounds considered to have unreliable predictions. In a case of drugs, the concordance rose to 97% (33 out 34 compounds were considered reliably predicted) as well as a concomitant drop in false positives and negatives as shown in Table 5.

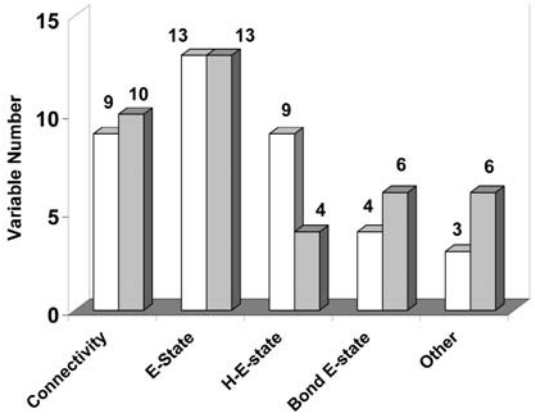

Figure 7. Topological descriptors subdivided into classes for genotoxicity models. Unfilled rectangles are the ANN-QSAR members and shaded are for the MLR-GA-QSAR. The number of variables in class and each model are shown.

The topological variables used in both of the models, 38 in ANN and 39 in MLR-GA, do differ; only 42 percent of the descriptors were common to both models. As shown in Figure 7, the number of descriptors per class was fairly constant for molecular connectivity and atom-type E-State indices found in the two models, but differed by approximately 2 fold in the number of H-atom E-States.

The latter accounted for 24% and 10% of the total variables in the ANN and MLR-GA QSARs respectively. A similar difference was observed with the class labeled "Other". It contained shape indices, polarity indices, and internal hydrogen bonding E-State values indices, which were found more frequently in the MLR-GA model.

Table 6. Important topological descriptors for the genotoxicity ANN and MLR-GA models[a]

ANN		MLR-GA	
Variable	*Frequency*[b]	*Variable*	*Frequency*
ArNH21[c]	417	e1N2N3d	146
Hmax	2963	ka2	2963
ArHNH21	245	phia	2963
Gmax	2963	e1C3O1d	252
SHBint2	665	SHother	2963
Qv	2963	SaasC	2119
eaC2C3s	1986	SaaN	405
SdsN	370	xp10	1946
Gmin	2963	e2N3O1s	284
SddsN	284	xch5	708

[a]Ranking of ANN descriptors described in Table 2. MLR ranked descriptors determined by the F-ratio employing RSS (sum of residues squared) in absence of a variable.
[b]Frequency is number of descriptor containing compounds in train set.
[c]Descriptor definitions:
ArHNH21: the larger atom H-E-state in an amine attached to an aromatic ring.
ArNH21: the larger atom E-state of N of a primary amine attached to an aromatic ring.
e1C3O1d: bond-E-State for =C-O group, e.g., carboxylic acid group.
e1N2N3d: bond E-State for -N-N= group.
e2N3O1s: bond E-State for - N=O group, e.g., nitrosamine.
eaC2C3s: bond E-State for -C=C-, e.g., two aromatic carbons, one with a substituent group attached.
phia: flexibility index, decreases with cyclicity or branching.
SaaN: sum of atom-E-States for aromatic nitrogens.
SaasC: sum of atom-E-States aromatic carbons with substituents.
SddsN: sum of atom E-States for N in nitro groups.
SdsN: sum of atom-E-States for -N= groups.
SHother:sum of H-E-States for hydrogens on atoms other than, O, N, C, or S.
xch5: chain chi 5 (5 membered ring) encodes for degree of substitution on the ring and cyclization.
xp10: path -10 subgraph for 10 member sequential bond path

In Table 6, two of the most important descriptors in MLR-GA model were Ka2 and phia, a shape and flexibility descriptor. In contrast, the bulk of ANN QSAR descriptors were either atom-type E-State or hydrogen E-State descriptors (ArNH21, Hmax, ArHNH21, Gmax, SdsN, Gmin, and SddsN) whereas bond E-States (e.g., e1N2N3d), shape, flexibility, and molecular connectivity indices made up majority for MLR-GA model. Many of these descriptors, as will be discussed, are directly connected to recognized structural alerts, toxicophores, for Ames mutagenicity.

Discussion

Various types of descriptors have been employed to model small datasets for aqueous solubility, including molecular fragments [25], melting point and logP [26] or with the latter two properties in conjunction with partial surface charges [27]. However, when confronted with the formidable task of developing predictive models with large datasets, topological indices by themselves or in conjunction with other descriptors have been used. In our approach, we chose not to use logP even though it reflects two opposing properties, the hydrophobicity and hydrophilicity of molecule as they influence So, the intrinsic aqueous solubility. Within our family of 542 topological variables, several indices mirror these opposing properties. Nonetheless, the 62 unique descriptors found for treating two different classes, aromatic and non-aromatic molecules, resulted in an 81:1 ratio of endpoints to indices in QSAR modeling.

Aqueous solubility is influenced by the degree hydrogen bonding between water molecules and the solute, the polar nature of the compound, and shape dependent intramolecular interactions among molecular substructures of the molecule, which, in turn, is a function of molecular branching, flexibility, and cyclicity. Therefore, it is surprising that the descriptor SsssN ranked first and second among the two ANN-QSARs. It was also found to be the most important variable for non-charged compounds in a 1291 compound study [2]. The E-State descriptor for a tertiary amine probably echoes the influence of the three groups attached to the nitrogen. Hydroxyl groups are an important contributor to intermolecular hydrogen bonding, which accounts for oxygen and hydrogen E-states present in four of the important descriptors via phenols or carboxylic acids, which together are found in 24% of aromatic training compounds. EPSA and Qv reflect the polar nature and polarity of molecule while shape, branching, and molecular size are encoded in Ka2, xv0, and x1 respectively. Both numHBa and numHBd, the number of hydrogen acceptors and donors, were found in both ANN models but only numHba was found in the most important variables. SHBint2, an E-State descriptor for acceptor and donor atoms separated by two skeletal bonds, was important in both ANN models. Since the formulations to compute aqueous solubility for charged compounds are well established [28], the results on the aqueous solubility for charged compounds presented in Table 1 may reflect a real difference in use of topological descriptors versus fragments with respect to predicting aqueous solubility. Fragments are treated mainly as isolated substructures without being influenced by either steric or electronic effects of other groups within the molecular context by the overall molecular shape of the molecule. In contrast, such effects are encoded in the E-State topological indices.

Predictive permeability models developed in this study followed upon the work of Norinder and Osterberg [8] on HIA and Yamashita et al. [9] with Caco-2. Both groups used topological variables. In this present work, we employed a much larger dataset. Furthermore, we used the TPSA descriptor [29]; the number of topological variables was larger than in the aforementioned studies since we introduced many new descriptors in this work. TPSA is the sum of the van der Waals areas of N, O, P, and S, which correlates very well with 3D-PSA values based on a single conformer approach. Since its frequent use in permeability modeling attest to its value, it was included in developing the ANN-QSAR for HIA.

Our interest is on transcellular transport, which is the major permeability pathway for most therapeutic drugs. The principle impediments to cellular permeability for this route are hydrogen bonding and formal charge as opposed to lipophilicity, which makes a positive contribution. The majority of the important descriptors in the ANN model encode structure features that are related to permeability. Furthermore, these descriptors show the expected trend directions as shown in Table 4. The most important descriptor found, SHBint4, makes a positive (+) contribution to the %OA. It reflects intramolecular H-bonding between an acceptor and donor.

Such an intermolecular interaction removes the donor and acceptor from the potential for intermolecular H-bonding with membrane proteins or charged lipids. On the other hand, the groups involved in the SHbint2 descriptor cannot undergo intramolecular H-bonding, since the acceptor and donor atoms are separated by only two skeletal bonds. These groups, then, contribute to H-bonding with other entities and lead to a negative trend. The index obtained from CSLogP, the calculated octanol-water partition coefficient, exhibits a positive trend as expected. The descriptor SCarOH1, contributes to intermolecular H-bonding through the carboxylic acid's hydroxyl oxygen atom that may seem a surprising finding. However, the low dielectric constant in the interior of the membrane means the pKa of an acid increases so the percent ionization is much lower giving way to a greater percentage of the unionized form.

The bond E-State descriptor, e1C3O1d, encodes for >C=O group, a polar entity in the molecule. Qv makes a positive contribution in keeping with the fact that the larger Qv becomes, the smaller is the polarity the molecule. TSPA is a negative factor to cellular uptake of compound. In our results, an increase in the polar surface area is mirrored by its negative trend. As a rule-of-thumb [21], PSA values greater than 140 A^2 are associated with poor compound permeability. However, this is not always the case. In the present data set, it is found to go as low as 100 A^2 , e.g., pentamidine and xamoterol for poor permeability. The ArNH1 descriptor, ranked number two in importance, is an E-State index for nitrogen in a secondary amine attached to an aromatic ring. It exhibits a positive trend and its role is not clear at present The negative contribution of Hmax, the largest H-E-State value in the molecule, is interpretable since Hamx is associated with the most polar hydrogen atom in the molecule. The descriptor SssO encodes electron accessibility for -O- atoms in heterocyclics or with ether linkages. The negative trend does suggest the oxygen atom may reflect weak H-bond formation with hydrogen donors among membrane members. Almost half the drugs involved in transcellular transport have single bonds to non-hydroxyl oxygens.

Genotoxicity can be defined as toxic changes emanating from effects on DNA. Several such changes are brought about by gene mutations due to covalent bond formation between DNA and a chemical to form a DNA adduct or nucleotide strand breakage, and others, indirectly, through interference in chromosomal segregation which may not be electrophilic in nature.

Many toxicophores, or structural alerts have been identified [30,31] within compounds found to be mutagenic in the S. Typhimurium histidine reversion assay, the Ames mutagenicity test. Examples of structure alerts can be a single continuous substructure, several non-adjacent ring substituent groups, alkyl halides, or polycyclic aromatics with or without one or more ring substituents. One can see why topological indices are so amendable to model genotoxicity since these structure descriptors encode whole molecule structure information as well as encoding both the topological environment of each atom and also the electronic influence of all other atoms. Correlating the most important descriptors in Table 6 to toxicophores makes interpretation of the important topological indices reasonably straight forward for many of them. We will only examine the ANN model descriptors. Many variables from the MLR-GA are also interpretable but not as many lend themselves to a 1: 1 correlation with alerts. With the ANN model, ArNH21 and ArHNH21 are among the most important descriptors in the 38 used in the consensus QSAR as shown in Table 6. These descriptors are for one or more primary amines attached to an aromatic ring. Amines, by themselves, are considered alerts [33] in this case. However, amines and nitro groups as ring substituents make up some of the most familiar and potent toxicophores. The SddSN descriptor, ranked tenth in importance, encodes for the nitrogen E-State for N in a nitro group. Hence, ArNH21, ArHNH21, and SddsN combine to signal this important family of toxicophores found by the ANN-QSAR. The SdsN descriptor is the E-State index for the nitrogen atom type, -N=, a group that is associated with structural alerts such as an azo group or nitrosamine. Those descriptors that are not directly correlated to common alerts are Hmax, Gmax, and Gmin. The latter is the E-State for the atom with the lowest E-State value, hence the most electrophilic atom. Its presence is understandable and somewhat rewarding in that it showed up as one of the most important descriptors. The meaning of Hmax is not clear at present. A related quantity was designated as a structural alert in another structure-activity relationship study employing quantum chemical parameters [32] for genotoxicity prediction for aromatics and heteroaromatics. Gmax, the highest atom E-State, is not correlated per se to an alert, but it may reflect the immediate adjacency of the atom (with the maximum E-State value) to an electrophilic center. Gmax is associated mainly with electronegative withdrawing group, e.g., halogens and nitro groups. The descriptor SHBint2, as of yet, is not interpretable other than it reflects H-acceptors and donor groups two bonds apart. Qv is a polarity index, whose value is function of the number of heteroatoms. It does correlate to logP; the larger its value, the more hydrophobic the compound. However, logP has not shown to correlate [7] with genotoxicity so Qv may simply reflect the presence H-acceptor atoms.

Conclusion

The modeling approach described here has produced very strong QSAR models for aqueous solubility, human intestinal absorption, and genotoxicity. Based on the these results, we contend that strong models of large diverse data sets for biologically related properties can be achieved by the combination of topological structure information representation, ANN and kNN modeling methods, and carefully developed training and external validation test data sets.

Statistical significance of a model, as presented here, focuses primarily on external validation tests. High quality direct statistics and cross-validation statistics are, of course, necessary. In our general methodology, heavy emphasis is placed on the use of representative external validation test sets. The strength of these models is made possible by a combination of high quality topological structure information representation, well-developed ANN modeling methods, and carefully validated data sets. Approximately 10,000 diverse compounds compose the ADMET datasets and only 108 topological structure descriptors were required. One measure of the significance of the descriptors is indicated by the diverse nature of the compounds and properties represented in the data sets.

For each property, the modeling process produced a set of descriptors that are significantly related to that property, in the statistical sense. Further examination of the nature of the topological descriptors reveals significant structure information about the relation of the property to molecular structure. The topological structure representation method (molecular connectivity, electrotopological state and kappa shape) has produced a set of descriptors that encode information about both fragment molecular structure as well their molecular context. Traditional fragment methods only count various molecular fragments. Each fragment is considered independent of its molecular environment. Based on our work, it appears that topological descriptors encode more significant information, leading to the higher quality models presented here In this investigation both training (MLR) and learning (ANN) methods were used. Based on this and other experiences, it appears that learning algorithms perform better for large, diverse data sets (N > 1000), such as those presented here. Furthermore, the ANN method permits a non-linear relation between structure and property.

References

1. Huuskonen; J., Salo, M. and Taskinen, J., Aqueous solubility prediction of drugs based on molecular topology and neural network modeling, J. Chem. Inf. Comput. Sci.,38 (1998) 450-456.

2. Huuskonen, J., Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology, J. Chem. Inf. Comput. Sci., 40 (2000) 773-777.

3. Tetko, I.V.; Tanchuk, V.Y.; Kasheva, T.N. and Villa, A.E.P., Estimation of aqueous solubility of chemical compounds using E-State indices. J. Chem. Inf. Comput. Sci., 41 (2001) 1488-1493.

4. Cheng, A. and Merz Jr, K.M., Prediction of Aqueous solubility of a diverse set of compounds using quantitative structure-property relationships, J. Med. Chem. 46 (2003) 3572-3580.

5. Votano, J.R., Parham, M., Hall, H.H., Kier, L.,B. and Hall, L.M., Prediction of aqueous solubility on large datasets using several QSPR models utilizing topological structure representation (submitted for publication, J. Chem. Info., Comput. Sci., 2004)

6. Mattioni, B.,E., Kauffman, G., W. and Jurs, P., C., Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble, J. Chem. Inf. Comput. Sci., 43 (2003) 949-963.

7. Votano, J.R., Parham, M., Hall, H.H., Kier, L.B., Oloff, S., Tropsha, A., Tonga, W. and Xie, Q., Three new QSAR consensus models for prediction of Ames mutagenicity (submitted for publication, Mutagenesis, 2004)

8.Norinder, U. and Osterberg,T., Theoretical calculation and prediction of drug transport processes using simple parameters and partial least squares projections to latent structures (PLS) statistics. Use of topological indices. 90 (2001) 1076-1085.

9. Yamashita, F., Wanchana, S. and Hashida, M. Quantitative structure/property relationship analysis of Caco-2 permeability using a genetic algorithm-based partial least squares method, J. Pharm. Sci., 91 (2002) 2230-2239.

10. Personal communication. Mario Lobell, OSI Pharmaceutical, Watlington Road, Oxford OX4 6LT, UK

11. http://www.chemsilico.com/CS_prHIA/HIAexp.html

12. Analyse-It Software, Ltd. PO Box 103, Leeds LS27 7WZ, United Kingdom; www.analyse-it.com.

13. QsarIS, v2, MDL Information Systems, San Leandro, CA; www.mdli.com.

14. Devillers, J. (Ed.) Genetic Algorithms in Molecular Modeling, Academic Press, 1996

15. Alan Miller, in "Subset Selection in Regression", 2nd Edition Chapman & Hall/CRC Press, 2002.

16. ChemSilico LLC, 48 Baldwin St., Tewksbury, MA 01876; www.chemsilico.com.

17. Ertl, P., Rohde, B. and Selzer, P., Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport, J. Med . Chem. 43 (2000) 3714-3717.

18. Avdeef, a., Berger, C.M. and Brownell, C., pH-metric solubility. 2. Correlation between the acid-base titration and the saturation shake-flask solubility-pH methods, Pharm. Res., 17 (2000) 85-89.

19. Advanced Chemistry Development Inc.90 Adelaide Street West, Suite 600, Toronto, Ontario M5H 3V9, Canada; www.acdlabs.com.

20. Lobell, M.; Sivarajah, V. In Silico Prediction of Aqueous Solubility, Human Plasma Protein Binding, and Volume Distribution From Calculated pKa and AlogP98 Values, J. Molec. Diversity, 7, (2003) 69-87.

21. Palm, K., Luthman, K. and Artursson, P., Polar molecular surface properties predict the intestinal absorption of drugs in humans, Pharm. Res. 14 (1997) 568-571

22. Clark, D. E., Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption, J. Pharm.Sci. 88, (1999) 807 - 814.

23. Stenberg, P., Norinder, U., Luthman, K. and Artursson, P., Experimental and computational screening models for the prediction of intestinal drug absorption, J. Med. Chem. 44 (2001) 1927-1937.

24. Derety, E., Feher, M. and Schmidt, J.M., Rapid prediction of human intestinal absorption, Quant. Struct.-Act. Relat. 21 (2002) 493-506.

25. Klopman, G. and Zhu, H.J., Estimation of the aqueous solubility of organic molecules by the group contribution approach, J. Chem. Inf. Comput. Sci. 41 (2001) 439-445.

26. Ran, Y. and Yalkowsky, S.H., Prediction of drug solubility by the general solubility equation (GSE), J. Chem. Inf. Comput. Sci. 41 (2001) 354-357.

27. McFarland, J.W., Avdeef, A., Berger, C.M. and Raesky, O.A., Estimating the water solubilities of crystalline compounds from their chemical structures alone, J. Chem. Inf. Comput. Sci. 41 (2001) 1355-1359.

28.Avdeef, A., High-throughput measurements of solubility profiles, In Testa, B., van de Waterbeemd, H., Folkers, G. and Guy, R. (Eds.) Lipophilicity in Drug Disposition: Practical and Computational Approaches to Molecular Properties Related to Drug Permeation, Absorption, Distribution, Metabolism and Excretion, Wiley-VCH Publisher Weinheim, Germany 1999, 305-325.

29. Ertl, P., Rohde, B. and Selzer, P., Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of transport properties, J. Med. Chem. 43 (2000) 3714-3717.

30. Ashby, J.; Tennant R.W.; Zeiger, E. and Stasiewiz, S., Classification according to chemical structure, mutagenicity to Salmonella and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. National toxicology Program. Mutat. Res. 223 (1989) 73-103.

31. Ashby, J. and Tennant, R.W., Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U. S. NTP. Mutat. Res. 257 (1991) 229-306.

32. King, R.D., Muggleton, S.H., Srinivasan, A. and Sternberg, M.J., Structure-activity relationships derived by machine language learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming, Proc. Natl. Acad. Sci., 93 (1996) 438-442.

33. Cariello, N.F., Wilson, J.D., Britt, B.H., Wedd, D.J., Burlinson, B., Gombar, V. Comparison of computer programs DEREK and TOPKAT to predict bacterial mutagenicity, Mutagenesis, 17, 2002, 321-329.