Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity

Joseph R. Votano*, Marc Parham, ChemSilico LLC, Tewksbury, MA 01876,
Lowell H. Hall, Department of Chemistry, Eastern Nazarene College, Quincy, MA 02170,
Lemont B. Kier, Department of Medicinal Chemistry, School of Pharmacy, Virginia Commonwealth University, Richmond, VA 23298,
Scott Oloff and Alexander Tropsha, The Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina 27599, and
Qian Xie and Weida Tong, Center for Toxicoinformatics, Division of Biometry and Risk Assessment, National Center for Toxicological Research, Jefferson, Arkansas 72079.

ABSTRACT

Three QSAR methods, artificial neural net (ANN), k-nearest neighbors (kNN), and Decision Forest (DF), were applied to 3,363 diverse compounds tested for their Ames genotoxicity. This group of compounds includes more than 300 therapeutic drugs. All models were developed using the same initial set of 148 topological indices: molecular connectivity chi indices, electrotopological state indices (atom-type, bond-type and single-atom E-State), as well as binary indicators. While previous studies have found logP to be one of the determining factors in genotoxicity, it was not found to be important by any modeling method employed in this study. The three models yielded an average train/test concordance value of 88%, with a low percentage of false positive and negatives. External validation testing on 400 compounds not used in the development of the QSAR models, gave an average concordance of 82%. This value increased to 92% upon removal of less reliable outcomes as determined by a reliability criterion used within each model. The ANN model showed the best performance in predicting drug compounds, yielding a 97% concordance (34/35 drugs) after the removal of less reliable predictions. The appreciable commonality found among the top ten ranked descriptors from each model is of particular interest because of the diversity in the learning algorithms and descriptor selection techniques employed in this study. 40% of the most important descriptors in any one model are found in one or two other models. Fourteen of the most important descriptors relate directly to known toxicophores involved in potent genotoxic responses in S.Typhimurium. A comparison of our validation results with those of TOPKAT, MULTICASE, and DEREK indicates clearly that the new models presented in this work perform substantially better than other commercially available models. The potential for the identification of predictions that are less reliable and show a greater probability of a false outcome is also explored.

Key words: Ames Mutagenicity predictor, QSAR models, artificial neural network, k-nearest neighbors, Decision Forest, topological structure representation

Introduction

In silico predictive models for genotoxicity fall into two principal categories: a rule based (expert system) and quantitative structure-active relationships models (QSAR). Rule based systems are built upon the earlier work of Miller and Miller (1977) and others (Ashby et al, 1988, 1989, 1991; Tennant et al., 1991) who first systematized relationships between chemical substructures and observed toxic outcomes. Expert systems are composed of structural rules derived from specific toxicological mechanisms or plausible modes of action of chemical agents in combination with pattern recognition routines to identify substructures associated with specific toxic effects (Sanderson and Earnshaw, 1991).

A great deal of effort is required in order for toxicologists to develop structural rules. Such rules, however, are favored by toxicologists because of their perceived transparency in the form of structure alerts that have a straight forward interpretation. The practice of relying on structural rules may be problematic. The designation of a substructure, or fragment, within a molecule as a toxicophore is a binary indication, and it is possible for the structure alert fragment to be present in a molecule without being a toxicophore in that molecule. The steric and electronic environment surrounding a structural alert fragment can diminish or enhance its genotoxic potency. This can render the fragment non-toxic, or create a toxic fragment that has not previously been identified. Both of these cases can lead to a false prediction. Recent work uses a phylogenetic-like tree algorithm (Bacha et al., 2002) to reduce the strict dependency of a yes/no rule for the presence/absence of a toxicophore. This method considers similar compounds with shared substructures, which can be either mutagenic or non-mutagenic. In this way, both active and inactive compounds contribute to the rule. It is possible that this technique may

---

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

serve to partially alleviate the problems that arise in the use of structure alerts. To complicate matters further, genotoxic changes can occur without covalent bond formation with DNA through the moiety's involvement in chromosomal segregation, DNA breakage, or interference in activities of non-DNA targets. In these cases, structural alert rules are absent because, to the authors knowledge, structure alerts for non-DNA targets have not yet been defined.

A QSAR model that has been designed to predict genotoxicity correlates the chemical structure, given in terms of continuous or binary molecular descriptors, to a compound's biological activity. Many different types of descriptors have been used in predictive models for Ames bacterial mutagenicity: chemical substructures (Klopman et al, 1984; Klopman et al., 1990); topological (Basak et al., 2001; TOPKAT, 2003), logP and electronic parameters (King et al., 1996) and various combinations of different classes of descriptors such as geometric, electronic, polar surface area, and topological (Mattioni et al., 2003). Most investigators agree that the validity of a QSAR model is based on its ability to predict biological activities for new chemical entities (NCEs) with a high degree of accuracy. Strong predictive ability depends on several factors: (i) the choice of descriptors used; (ii) the learning algorithm or methodology employed in model development; and (iii) the selection of the compounds used in the learning or training process. Lack of quality in any of the three factors usually translates into an unsuitable predictor for NCEs. Selection of a meaningful and statistically significant set of descriptors, those that correlate well to activity, is accomplished by using various stochastic sampling techniques: genetic algorithm (Hopfinger and Patel, 1996) or simulated annealing (Kirkpatrick et al., 1982; Sutter et al., 1995) in conjunction with a correlation method, e.g., MLR. Algorithms to build predictive genotoxic models have utilized multiple linear regression (MLR), partial least squares (PLS), hierarchical QSAR, and inductive logistic programming (ILP). These algorithms usually assume the existence of a linear relationship between biological activity and the molecular descriptors employed. However, given the quantity and variety of mechanisms involved in toxic interactions, the relationship between descriptor values and toxicity may not be strictly linear. To model non-linear effects between descriptors and toxic activities, some nonlinear QSAR methods make use of artificial neural net (ANN) analysis [using mainly back propagation (Devillers, 1996)] or k-nearest neighbor algorithms (kNN-QSAR) (Zheng, 2000). At this time as done here, we are unaware of any ANN-QSAR or kNN-QSAR for genotoxicity that has been reported for a large dataset of mutagens and non-mutagens.

Structural and chemical diversity among members of a large dataset is a key consideration in this study in the development of several new QSAR models for Ames genotoxicity. Approximately 3,400 chemically diverse compounds are used in this work. Approximately 3000 compounds served as the train/test set for model development and 400 compounds were used for external validation testing. The dataset includes more than 300 therapeutic drugs. In model development, the same set of topological descriptors (Kier and Hall, 1996, 1999) was used along with new proprietary desciptors (Chemsilico, LLC) in all models. These structure variables include descriptors that encode whole molecule structure information as well as atom level descriptors that encode the electronic character of each atom, taken along with the topological environment and electronic influence of all other atoms in the molecule. Three different QSAR models were developed by the use of three separate modeling techniques: ANN-QSAR, kNN-QSAR, and Decision Forest (DF), an advanced tree-based classification algorithm.

Materials and methods

Data sources for compounds and their selection: Data on bacterial mutagenicity for strains of S. *typhimurium* was obtained from various sources: toxicological data from National Library of Medicine's Toxicology Data Network (TOXNET, http://toxnet.nlm.nih.gov); Gold and Zeiger, 1999; RTECS (Register of Toxic Effects of Chemical Substances), and the Physicians Desk Reference (version 6.0a, 2003). Only mutagenicity data without metabolic activation by the S9 rat liver microsomal preparation for strains (TA97, TA98, TA100, TA102, TA104, TA1535 and TA1537 of S. typhimurium) was used. Compounds did not need to be tested on all seven strains to be included in the data set; a minimum count of two strains was used. A number of additional preprocessing steps were performed on the data, such as the removal of duplicate structures and removal of compounds with conflicting reported mutagenicity values. Only clearly reported bacterial mutagenicity values were used. For tracking and labeling purposes, we designated an active (mutagenic) or inactive (non-mutagenic) compound by a mutagenic index, MI: 0 for a non-mutagen and 1 for mutagen. A total of 3393 compounds were selected with a 60/40 ratio of mutagens to non-mutagens. Compound selection for the train/test and the external validation sets was random with 12% of the initial compounds set aside for external validation testing. The structures and experimental Ames genotoxicity values for the compounds in the external validation set did not contribute to either descriptor selection or model development for the three methods examined in this investigation. Chemical structures for all compounds in the dataset were entered in Mol file format using structures from TOXNET's ChemPlus database and ChemDraw version 7.0. The methods employed in this study do not require three dimensional atom coordinates or an optimized minimum energy conformation for the input of structures.

Molecular descriptors: An initial set of 542 topological descriptors indices was computed by ChemSilico software (ChemSilico LLC) and reduced to a set of 148, using the criterion that at least 5% of the descriptor values must be non-constant (non-zero in most cases) for the 2963 compounds in the train/test set. The indices include molecular connectivity chi indices, atom-type, group-type, bond-type and single atom

E-State and hydrogen E-State descriptors, kappa shape indices, several binary indicators (e.g., presence of aromatic ring), topological polarity, and others. The selected set of 148 descriptors was then further reduced in separate variable selection routines used by the three modeling algorithms in this study.

Model Development

Model developed by ANN analysis: Artificial Neural Network analysis was performed on 88% of the dataset (train/test) with 12% set aside for external validation. The train/test set, designated the principal set, was randomly split into 85% for train and 15% as a selection set for early stopping of the learning process to avoid over-fitting. The train set was selected randomly ten times, forming ten data sets using 75% of the train set, and a mutually exclusive 10% withholding set in which each compound appears only once in each withheld set. This multiple selection process produces a set of ten models derived from the principal set using ten mutually exclusive withholding sets. Using this approach, the non-contributory variables are pruned to give an optimal subset of significant variables. The initial starting set of 148 descriptors was reduced to 38 in the ANN analysis. The relative importance of each eliminated variable is based on its contribution across the entire train/withholding sets by calculation of r^2 in each instance when the row (compound) appears in the withholding set. This value is designated q^2, that is, the r^2 value for all instances in which the data was withheld from the modeling process. Since q^2 is used to select the variables, it does not provide a reliable assessment of the predictive accuracy of the overall algorithm. This task is reserved for an external validation set.

A standard back-propagation neural network was used for this study. The network contained no more than nine hidden neurons and utilized the backward elimination approach (Devillers, 1996; Miller, 2002), which has been adapted from traditional linear regression methods. The ten-fold cross-validation algorithm is used as a consensus model in which the average value of ten neural nets gives the predicted MI value in the range from 0 to 1 for the Ames bacterial mutagenicity of a compound. A threshold value of 0.5 was used in which threshold = 0.5 means the compound is Ames positive, i.e., a mutagen (MI = 1). A possible standard for the identification of predictions that are potentially less reliable has arisen from examination of the standard deviation associated with the mean value from ten ANN models that form the consensus. Predictions with a standard deviation value exceeding 0.27 were found to give a false prediction more often than those with a smaller standard deviation. For this reason, these predictions are flagged as having a lower confidence level. Ranking of descriptors with respect to their importance was 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 smallest difference for the least important variable in the train-test set, using an average RSS

values from all ten ANN models. The model produced by this ANN procedure is commercially available in the ChemSilico Predict software (ChemSilico LLC).

Model developed by Decision Forest: Decision Forest is an algorithm (Tong et al., 2003) that creates and combines multiple heterogeneous but comparable Decision Tree models to achieve an improved prediction outcome. Development of a forest consists of four steps: (1) develop a qualified tree; (2) develop the next qualified tree based on only the descriptors not used in the previous tree(s); (3) repeat steps 1 and 2 until no additional qualified trees can be developed; and (4) classify a compound's activity based on the mean value of the predicted activities from four trees. Every tree uses a distinct set of descriptors; 144 descriptors were used among all trees. Decision Forest assigns a probability value to each prediction for which chemical compounds with probability values greater than 0.5 are designated as mutagens: MI = 1; others are classified as non-mutagens. The algorithm assigns a confidence level to every prediction, which is defined as high confidence when the predicted value lies between 0.0 to 0.3 or 0.7 to 1.0 range, whereas the range from 0.30 to 0.70 is considered to be in a low confidence region. This confidence value is used as a reliability criterion for the Decision Forest method. Ranking of variables was done by determining the frequency of each descriptor in a ten-fold cross-validation procedure that was repeated 50 times, using ten train/test sets.

Model developed by kNN-QSAR: K-nearest Neighbors QSAR models are developed by a method (Zheng and Tropsha,. 2000) that is an advanced non-linear, non-parametric, variable selection technique that assigns predictions based on a test compound's similarity to training compounds with known activities. The model development starts with use of a sphere exclusion algorithm (Golbraikh and Tropsha, 2002) to split the training data of 2963 compounds into ten distinct training and test sets for model building and validation. Each training set consists of 50 to 80% of the available compounds; the remaining compounds are placed in an external test set to validate the models.

To build classification models of genotoxicity, a recently developed method was implemented (RWkNN-Classification method) that has evolved from the earlier original implementation (Zheng and Tropsha, 2000). This new approach is much faster for large datasets than the previous implementation. RWkNN-Classification represents training set compounds as points in a multidimensional descriptor space and uses the distance between points to measure similarity between compounds. The descriptor space sought by the modeling procedure is a subset of the entire descriptor space, optimized by a simulated annealing-based sampling method. The central idea of this method assumes that compounds with similar activities are positioned relative to one another such that descriptor similarity correlates with activity similarity.

3

The criterion used during descriptor space optimization is the ability of the space to achieve the above correlation. To do this, an optimal value of k is identified such that each compound in the training set can be best predicted by the k-most similar compounds (out of the remaining training set). The resulting model consists of an activity relevant descriptor subspace, a defined k, and training set compounds with known activities used to populate this descriptor subspace. An external test compound is predicted by identifying its location in the model descriptor space and then the k-most similar training set compounds are used to predict its class (toxic or not toxic) in a similarity-weighted fashion.

Since the procedure as implemented requires a pre-specified number of descriptors in the optimized sub-space, and there is no foreknowledge of what this number should be. A stochastic search was done within a specific range of values known to work well for kNN. In this study, descriptor numbers from 10 to 20 with an interval of 2 were searched, which resulted in 6 distinct values for the number of descriptors. At this point there are 60 differing sets of input parameters to search (10 training/test sets * 6 descriptor numbers). Three models were built for every set of input parameters, yielding 180 models in total. Each model was validated by its ability to predict its corresponding external test set not available during model building. Of the 180 models generated, 140 were capable of predicting a test set with greater than 75% accuracy. These 140 validated models were then used to predict the 400 blind external test compounds in a consensus fashion. The predictions made for each blind compound were collected and any compound that was not consensually predicted to belong to one specific class by at least 65% of the models was excluded; the prediction in this case was considered unreliable. The ability, or inability, to assign blind compounds to at least 65% of models was used as a reliability criterion for the kNN method. Ranking of variables was done by the frequency of occurrence of each descriptor in all 140 models.

Data Set Characterization: The structural and chemical diversity of the 2,963 compounds in the train dataset is illustrated in Table I. Compounds containing ring systems account for 80% of these compounds, 53% of which are fused-ring systems. N-heteroaromatic rings are present in 16% of the train set with 66% of these compounds being mutagens. Non-heteroaromatic rings, fused or unfused, make up 56% of the train set with only 7% of all the train compounds having either unsaturated or partially saturated ring(s). Amines attached directly to aromatic rings (single or polycyclic) are found in 30% of compounds; 26% of these amine-bearing compounds are mutagens. Compounds in the train set have an average of six rotatable bonds per compound. Compounds containing halogens and nitro groups make up 20 and 26%, respectively, of the train set. Therapeutic drugs account for approximately 10% of the 3363 compounds in the dataset. Only 1.2% of the drugs are Ames positive. Very similar percentages of compounds with these attributes are found among the 400 compounds in the external validation set, as shown in Table I

Table I. Compound attributes in train and validate sets

Attribute	Train			Validate		
	No.[a]	Ames-negative	Ames-positive	No.	Ames-negative	Ames-positive
Ring(s)	2371	807	1574	314	103	211
Fused ring(s)	1268	312	956	171	42	129
N-heteroaromatic ring(s)	462	155	307	58	19	39
Aromatic Rings(only)	1674	674	1462	342	140	202
Ar–NH₂[b]	417	77	340	54	13	41
Ar–NHR	245	78	167	32	9	23
Ar–NR₂	235	61	174	30	9	21
<Rotbonds>[c]	5.6	1127	1676	5.2	155	223
<numHBd>[d]	1.1	686	1024	1.1	92	124
<numHBa>	3.8	1108	1661	3.7	152	222
Halogen(s)[e]	590	247	343	85	34	51
Amine(s)[f]	1383	485	898	179	68	111
–N=	370	50	320	62	15	47
–NO₂	766	36	248	47	4	43
<MW>	243	244	242	238	226	245
Therapeutic Drugs	290	255	35	39	35	4
Total Compounds	2963	1161	1802	400	159	241

[a] Number of compounds with given attribute or its average. Train means compounds in both train and test sets.
[b] Ar-NH₂,-NHr, -NR₂ are amines attached directly to an aromatic ring.
[c] <Rotbonds>: average number of rotatable bonds.
[d] <numHBd, a>: average number of H-bond donors or acceptors.
[e] Halogen(s): a compound contains one or more F, Cl, Br, or I substituents.
[f] Amine(s): compound contains one or more -NH₂, -NHR, or -NR₂ groups.

Results

Statistical Information: Statistical parameters for the results from three predictive models, ANN-QSAR, DF, and kNN-QSAR, are given in Table II for the train and validate sets as well as results on therapeutic drugs. All models output a continuous value between 0 and 1 for the predicted genotoxicity index, MI, except for the kNN model which directly creates a binary output. An arbitrary threshold of 0.5 was used by the ANN and DF models: MI = 1 (mutagens) when threshold $\geq$ 0.5. To check whether this threshold value gives high sensitivity, a minimum number of false positives and false negatives, areas under receiver operator characteristic curves (ROC) were computed (Analyse-It, 2003). A typical curve is shown in Figure 1 for the ANN-QSAR train dataset. A model with no predictive ability yields the diagonal line. The DF and kNN models show almost identical curves to that in Figure 1. The closer the area under the ROC curve is to 1, the greater is the predictive ability of the model. As shown in Table II, all three models have essentially the same ROC area values. Small changes in the threshold value, $\pm$ 0.05, have no significant impact in all three models but larger changes show significant negative impact on the outcome.
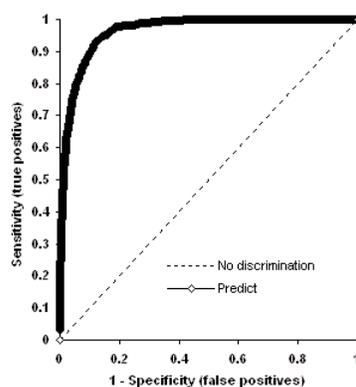

Figure 1. Receiver operator curve (ROC) for the ANN model of the 2963 training compounds. Dotted diagonal represents outcomes when a model has no discriminatory power in predicting binary outcomes.

4

Table II. Statistical parameters on prediction of Ames genotoxicity from three QSAR models, Artificial Neural Network (ANN), k Nearest Neighbors (k-NN) and the classification method Decision Forest (DF).

Model	N	M	ROC	Concordance (train)	%F Pos (train)	%F Neg (train)	N	ROC	Concordance (validate)	%F Pos (validate)	%F Neg (validate)
ANN	2963	38	0.96	89%	2%	9%	400(319)	0.93	83(91%)	4(3%)	13(6%)
DF	2963	144	0.94	90%	3%	8%	400(244)	0.91	82(93%)	7(2%)	11(5%)
KNN	2963	51	0.95	84%	7%	9%	400(313)	0.92	84(91%)	8(6%)	9(3%)
				Therapeutic Drugs					**Therapeutic Drugs**		
ANN	290	38		93%	4%	3%	39(34)		92(97%)	5(0%)	3(3%)
DF	290	144		93%	5%	2%	39(16)		85(88%)	13(12%)	3(0%)
KNN	290	51		90%	6%	4%	39(28)		69(82%)	10(7%)	21(4%)

Data set split into 2963 compounds for train/test and 400 for validation testing; Validation numbers in parenthesis are confidence outcomes from each model; N: number of compounds; M: number of descriptors used in the consensus model; ROC: area under the receiver operator characteristic curve; %F Pos and %F Neg: percentages of Ames false positives and negatives.

External Validation Information: Each of the learning algorithms investigated in this study produced a consensus based model with a solid statistical performance for external validation. This quality result is of interest, given the diverse methods involved in the generation the three models. The evaluation of these models is based on the overall concordance for both train and external validation compounds. The average values for the three models are 87%±3% and 83%±1%, respectively, for the train and external validate sets. Averages were calculated without the exclusion of outcomes estimated to be less reliable. Each of the three methods employed a reliability criterion designed to identify outcomes estimated to be less reliable. When compounds are excluded from the external validation set on the basis of the reliability criterion, the average concordance for all models increases from 83% to 92%. This is considered an important feature for commercial considerations where it is an advantage to able to inform the software user that a given prediction may be less reliable. Concomitantly, the percentage of Ames false negatives decreases appreciably from an average of 11% to 5% with a smaller decline for false positives. Application of the respective reliability criterion for the ANN and kNN models indicated that approximately 20% of the predictions should be flagged as less reliable. The reliability criterion used for the model developed by the DF method indicated that approximately 40% of its predictions should be considered as less reliable.

The proficiency in predictive ability of the three models differs significantly in their ability to predict genotoxicity for therapeutic drugs. Among the models, an average concordance of 92%±2% was found for 290 drugs in the train set; however, results on the validation set (39 drugs) varied considerably. The ANN-QSAR performs very well, yielding a 92% concordance value for all 39 drugs. The prediction rate increases to 97% (33/34 predicted correctly) after applying the reliability criterion to remove of less reliable outcomes. The classification-based DF model yields 85% and 88% respectively, the latter value arising from removal of 60% of the drugs by its reliability criterion (14 out 16 being correctly assigned as mutagen or non-mutagens after removal of less reliable predictions). The kNN-QSAR model increases from a 69% concordance to 82% after removal of 28% of the drugs considered to have less reliable outcomes.

Since the concordance for both the kNN and DF models on train set (290 drugs) is excellent, it is necessary to address the possibility that the difference in external validation performance among the models may arise from the random selection of the 39 drugs used for external validation. It is necessary to confirm that the external validation compounds are representative of the train set in terms of descriptors used by each consensus model. Three hierarchical clustering procedures (Ward's method), each composed of ten clusters (data not given), were run on the total dataset of 3363 compounds, employing the descriptor values used in each consensus model. All the drugs are well represented in 6 or 7

Table III. The 10 most important descriptors in ANN, k-NN, and Decision Forest (DF) models, trend of ANN descriptors, and frequency of variables in training compounds. Descriptors in bold were found to be important in by more than one modeling technique. Descriptor definitions are given in Table IV.

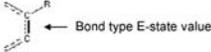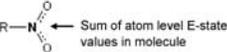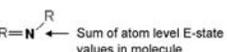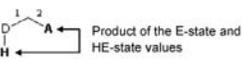|	ANN model			DF model		k-NN model	
Rank[a]	Variable	Trend[b]	Frequency[c]	Variable	Frequency	Variable	Frequency
1	ArNH21	+	417	**SdsN**	370	SaasC	417
2	Hmax	+	2963	SsssN	610	**SdsN**	370
3	ArHNH21	+	245	**Qv**	2963	ArN1	240
4	**Gmax**	varialbe	2963	**Gmax**	2963	**SddsN**	264
5	SHBint2	varialbe	665	phia	2963	e1N2N3d	146
6	**Qv**	-	2963	ka1	2963	**Gmin**	2963
7	eaC2C3s	+	1986	**e2C3O1s**	1103	SsCl	400
8	**SdsN**	+	370	SHBa	2777	**e2C3O1s**	1103
9	**Gmin**	+	2963	**Gmin**	2963	SHHBd	1710
10	**SddsN**	+	284	SHCsats	1676	e2N3O1s	284

[a] Rank for ANN model 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 averaged across all 10 ANN models. DF ranked descriptors determined by the frequency each descriptor was used in 50 times 10-fold cross validations. K-NN rank is number of times each descriptor occurred in 140 models.

[b] A positive trend indicates that the Ames bacterial mutagenicity will tend to increase as the descriptor value increases. A negative trend indicates the inverse effect. A variable trend is dependent on the range of the descriptor value or interdependent on the value of other descriptors. The trend is a mean over the 10 ANN models.

[c] Frequency is the number of compounds in the train/test sets expressing the topological descriptor.

of the clusters with no drug singletons present. The average percent of validation drugs (as a percentage of the total drugs in these clusters) was reasonably constant: 13%, 14%, and 15% for the kNN, DF, and ANN models respectively. Therefore, it is reasonable to state that the 39 drugs in validation set were well represented in the train set by all models.

Table IV. Summary definitions of all descriptors that were found to be in the top 10 in importance by the 3 modeling methods evaluated.

Index	Description	Illustration
ArHNH21	In a molecule, the largest hydrogen E-state value for a hydrogen on a primary amine attached to an aromatic ring. Calculated for the protonated form of the amine.	Largest atom level hydrogen E-state value
ArNH21	In a molecule, the largest E-state value of a primary amine attached to an aromatic ring system. Calculated for the protonated form of the amine.	Largest atom level E-state value
e2C3O1s	Sum of the bond E-state values for double bonds in carbonyl groups (>C=O).	Bond type E-state value
eaC2C3s	Sum of the bond E-state values for carbon-carbon aromatic bonds in which one carbon is substituted.	Bond type E-state value
SddsN	Sum of the atom level E-state values for all the nitro group (-NO$_2$) nitrogen atoms in the molecule.	Sum of atom level E-state values in molecule
SdsN	Sum of the atom level E-state values for all the -N= group nitrogen atoms in the molecule (e.g., azo, nitrosoamine).	Sum of atom level E-state values in molecule
SHBint2	The largest product of E-state and HE-state values from all acceptor and donor pairs separated by 2 skeletal bonds.	Product of the E-state and HE-state values
SHCsats	Sum of the hydrogen E-state values of all hydrogen atoms attached to sp3 carbons that are bonded only to other sp3 carbons.	
Phia	A kappa shape molecular flexibility descriptor that increases with homologation and decreases with increased branching or cyclicity. Larger Phia values indicate greater molecular flexibility.	
Ka1	1st order Kappa shape descriptor, encodes the degree cyclicity of the molecule.	
Q$_v$	A whole molecule polarity index that decreases in value as the polarity increases.	
SHBa	Sum of the atom level E-state values of hydrogen bond accepting atoms in the molecule.	
Hmax	The maximum hydrogen atom level E-state value in a molecule.	
Gmax	The maximum atom level E-state value in a molecule.	
Gmin	The minimum atom level E-state value in a molecule.	

Structure Descriptors: Table III lists the top ten ranked descriptors found in each consensus model along with their frequency of occurrence in the train dataset. It is of interest, given the different nature of the learning algorithms and the descriptor selection processes, that four of the top ten ranked variables in any one model are common to one or two other models. These descriptors are shown in bold-faced-type in Table III along with their frequency (See Table IV for descriptor definitions). **Gmin** and **SdsN** are found in all models. Four of the important descriptors (**Gmin**, **Hmax**, **Gmax**, and **Q**$_v$) in the ANN and DF models are what may be called global topological indices as indicated by their frequency values: there exists a computable value for any organic molecule. The frequency of different topological descriptors used by the three models is given in Figure 2. Topological descriptors were subdivided into classes: molecular connectivity chi indices, bond-type and atom-type E-State descriptors and hydrogen E-State descriptors, and others including, MW, number rotatable bonds, number hydrogen donors and acceptors. DF favored use of both atom-type E-State descriptors and molecular connectivity indices;
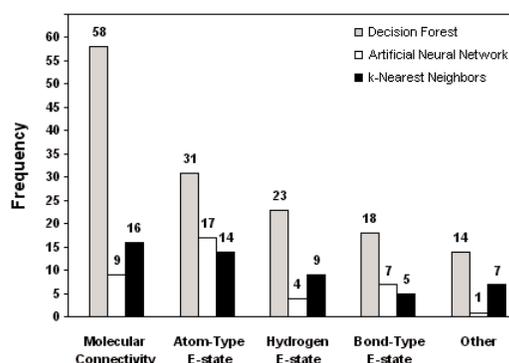
Figure 2. Frequency of topological descriptors subdivided into classes. Black, gray, and white rectangles are KNN, Decision Forest, and ANN models respectively. Other includes descriptors such as MW, flexibility, and the number of rotational bonds.

62% of all it variables were in these two classes. For the ANN model, 19 and 37 descriptors are in common with the kNN and DF models respectively from among these classes. The percentages of descriptors used by the ANN and kNN models differed among the various classes with the largest disparity being in number of atom-type E-State and hydrogen E-State indices. Members of these classes accounted for 63% of the 38 descriptors used in the ANN models as opposed to 47% of the 51 variables used by kNN-QSAR.

Discussion

Several studies on toxicity, including benzene toxicity (Hall et al., 1989a, 1989b) , phenol toxicity (Hall and Vaugh, 1997), fish toxicity (Rose, 2003), and amide herbicide toxicity (Gough and Hall, 1999) have employed only topological indices in modeling structure-activity relationships (QSARs) as is done in this present study. Because logP has been shown to be a useful parameter in some previous QSAR studies on toxicity, it was included in the 148 descriptors that were considered by each of the modeling methods. The three modeling methods employed in this study did not select logP to be included in the best model. As an individual descriptor logP showed a negligible correlation (r^2 = 0.02) to the Ames mutagenicity for the compounds in the train set. Recently, a similar lack of correlation has also been reported in another genotoxicity QSAR study (Mattioni et al., 2003).

The models of Ames genotoxicity presented in this investigation perform very well in the prediction of external validation compounds. Based on the authors experience with these and other models, the authors contend that one of the critical determining factors for this success arises from the nature of the molecular structure representation employed in the model development process. In all these models, topological structure descriptors have been used. Three aspects of this approach form the basis for the strong models. First, whole molecular descriptors (usually molecular connectivity chi and kappa shape indices) encode significant information on

the general features of the molecules. Whole molecule descriptors encode information on general structure features such as molecular size and shape, as well as specific information on skeletal variations and complexity. These structure features are expected to have a relationship to properties arising from intermolecular interactions and may also function to provide discrimination among multiple structural classes. Second, atom-type, group-type, bond-type and single-atom E-State descriptors encode information on specific molecule features such as atom and bond types associated with important functional groups. For all three genotoxicity models, some of the E-State descriptors found to be important relate directly to known structural alerts for mutagenicity (see figure 3). Each of these E-State descriptors encodes three aspects of structure: electron accessibility at that atom or bond, the presence/absence of that feature, and finally the count of that feature. We emphasize that these descriptors are not counts; they encode both the influence of the molecular context as well as the count. Because the E-State is a continuous value, a model based on E-State descriptors can correlate genotoxicity to a specific range of a descriptor value, whereas the use of fragment based structure alerts limits the model to a correlation to the presence/absence or simple count of a given fragment. As was stated in the introduction, the practice of structure alerts has been demonstrated to lead to false prediction for this reason.
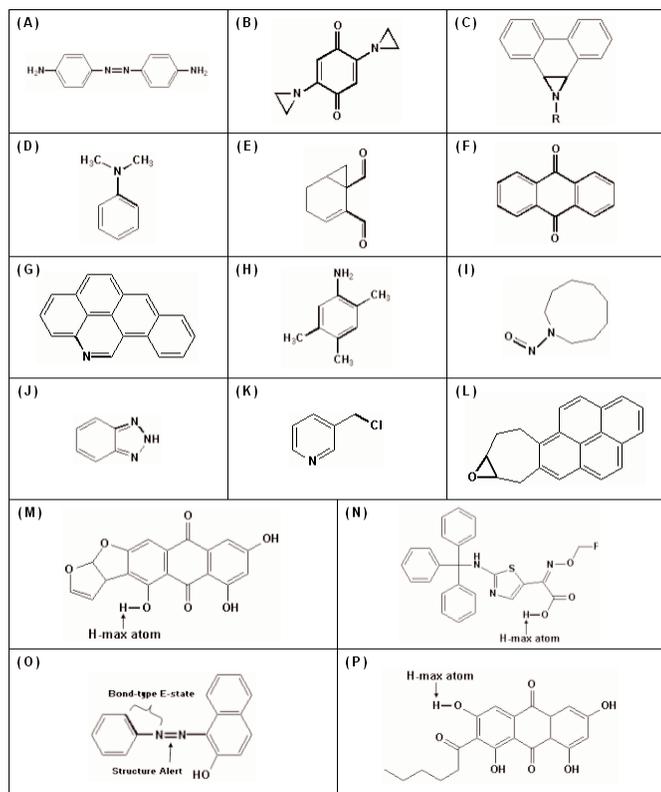
Figure 3. 2D structures are labeled A-P. Bold, wide bonds show positions within structures where descriptors indicate a structural alert for Ames mutagenicity as found among most important E-State indices. Also, positions of a bond E-State alert, eaC2C3s, and Hmax in several structures.

The topological modeling approach is said to mechanism-free, or not to be mechanism biased (Adamson and Bawden 1976, Adamson and Bush 1976, Hall, 2004). This aspect is very important because it is not necessary to assume any set of mechanistic steps in order to make the computation for these complicated biological properties. Further, it is also not necessary to make approximations in order to carry out the computations related to an assumed mechanism, a necessity when dealing with any assumed mechanism of interaction.

A second critical determining factor for the success of the models presented in this study is the use of advanced non-linear modeling and descriptor selection techniques. It is intuitive to assume that the relationship between the presence of a genotoxic sub-structure in a molecule does not relate to toxicity in a linear fashion. A compound that contains five toxicophoric centers is not likely to be 5 times more genotoxic that a compound with one such structure. Such relationships are generally asymptotic with sigmoidal being the most common example. There are many other factors that influence toxicity, such as solubility, the ability to penetrate biological membranes and the potential for defensive metabolism, each of which have a complex relation to structure that may or may not be linear. The authors believe that the ability to thoroughly explore the non-linear relation between structural characteristics and genotoxicity is essential to the development of a successful model. This includes the use of a variable selection process that can evaluate the potential of a descriptor to show a strong non-linear correlation to the activity or property under consideration.

These factors work together to provide the basis for strong modeling: use of a wide range of topological descriptors including whole molecule and atom level descriptors, the use of advanced non-linear descriptor selection and modeling techniques and the resulting development of a direct relationship between structure and activity.

Structure Interpretation

Many of the most important descriptors shown in Table II relate directly to or are associated with structural alerts as reported by other investigators (Ashby and Tennant, 1991 and King et al., 1996). For purposes of this discussion, the descriptors will be presented in groupings related to their fundamental nature. The definition of each descriptor is given in Table IV.

Atom-Type E-State Descriptors:
ArNH21 and **ArHNH21**: The descriptors ArNH21 and ArHNH21 encode electron accessibility of amino groups attached to aromatic ring systems. These substructures are prevalent among mutagens and non-mutagens, especially polycyclic aromatics (Kubo et al., 2002). According to the model, an increase in either of these descriptors leads to an increase in the predicted genotoxicity value.

7

SddsN: The SddsN index, (atom type E-State descriptor for nitrogen in nitro groups), together with either ArNH21 or ArHNH21, signals common structural alerts for compounds containing either group or the combination of one or more nitro and amino groups. Molecules with larger SddsN values tend to have larger predicted genotoxicity values.

SdsN: The SdsN descriptor (E-State for the nitrogen atom type -N=), found in all models, is associated with the azo group a structure alert (see structure A in Figure 3). Molecules with larger SdsN descriptor values tend to have larger calculated output values.

SsssN and **ArN1**: In some cases, several E-State descriptors encode similar information about an atom type involved in a QSAR, one descriptor being more general than the other. Such is the case with SsssN and ArN1. The latter is the atom-type E-State descriptor for a tertiary nitrogen atom attached to an aromatic ring whereas SsssN is the more general case, the atom-type E-State of all tertiary nitrogens in molecules. Tertiary nitrogen group alerts occur when the nitrogen is attached to either an aromatic or partially unsaturated rings (see structures B to D in Figure 3). The DF model selects SsssN whereas the kNN model employs ArN1.

Bond-Type E-State Descriptors:
Of interest is the selection of bond-type E-State descriptors, which encode the electron accessibility at the bond and along with perturbations from all other bonds in the molecule (Kier and Hall, 1999).
e2C3O1s: The bond-type, e2C3O1s, found in the kNN and DF models, encodes electron accessibility of the carbonyl bond, >C=O, which can be a structure alert or part of an alert (see structures E and F in Figure 3).

eaC2C3s: The bond type descriptor, eaC2C3s, found in the ANN model, encodes an aromatic C-C bond in which one carbon is attached by a single bond to a substituent group.

SaasC: The bond type is analogous to the atom-type E-State, SaasC, in the kNN model, which is the atom type E-State for aromatic carbons with an attached substituent atom. Neither the atom-type nor the bond-type E-State descriptor stands for an alert per se. However, their E-State values reflect the nature of structural alert(s) attached to the ring system (see structures D, H, and O in figure 3). The bond type encoded in eaC2C3s was identified (King et al., 1996) as a structural alert for aromatic or heteroaromatic nitro compounds.

e1N2N3d: Another bond type descriptor, e1N2N3d (sum of E-State values for the >N-N= group) identifies a known structural alert (see structures I and J in Figure3).

e2N3O1s: The presence of a nitro group is also reflected in the bond-type E-State, e2N3O1s, found in the DF and kNN models. In this bond-type, the bond encoded is >N=O.

Global E-State Descriptors:
Four of global descriptors, **Hmax**, **Gmax**, **Q^v**, and **Gmin**, are also found to be most important.

Hmax: Of significant interest is Hmax since it was ranked number two in the ANN-QSAR and was also found to be of moderate importance in the DF model. It is the maximum hydrogen atom E-State value in a molecule, which signifies the largest polarity on a hydrogen atom in the molecule. It also correlates with partial charge. The Hmax values range from 0 to 3.1 in the train set. Examination of the 100 lowest (0 - 0.65) and highest (2.8 - 3.1) Hmax values of compounds in the train set reveals a significant trend with respect to Ames mutagenicity. Twenty-eight percent (28%) of compounds in the low Hmax range are mutagenic, as compared to 79% of compounds with high values of Hmax. The presence of a highly polar H atom does not represent a structural alert in the typical sense of how a toxicophore is defined. Atoms encoded as Hmax tend to be associated with -OH, -NH2 and -NH- groups of acids, amines and amides and substituents attached to aromatic rings (see structures M, N, and P). Furthermore, if the molecule does not possess any of these highly polar centers, the atom with the maximum H E-State value can still be significant. It is not clear whether these polar atoms are involved mechanistically in a non-nucleophilic interaction with DNA or that Hmax reflects the influence of a nearby toxicophore. Despite the lack of a clear mechanistic implication, the study points to a statistically significant relationship between high values of Hmax and Ames genotoxic potency. The same effect has also been associated with high mutagenic potency by King utilizing electronic descriptors (King, 1996).

Gmax: The E-State descriptor Gmax is the largest E-State value in a molecule, usually associated with the most electronegative atom in the molecule, and there is a strong probability that its selection relates to structural alerts containing such moieties (structures G, K, and L in Figure 3) that are adjacent to electrophilic centers.

Gmin and **Q$_v$:** The E-State descriptors Gmin and Q$_v$, respectively, are a measure of the most electrophilic atom in the molecule and the polarity of the molecule. Mechanistically, an electrophilic center is important for covalent bond formation with nucleophilic DNA, and so it is not surprising that Gmin is found to be important in all three modeling methods. Q$_v$ is inversely proportional to the polarity and hydrophilicity of a molecule. The Q$_v$ descriptor reflects the presence of heteroatoms and polar functional groups, many of which are related to genotoxic structure alerts. Molecules with smaller values for Q$_v$ are more polar and tend to have larger predicted genotoxicity values.

Group-Type E-State Descriptors:
Three other important descriptors are all group-type E-State and hydrogen E-States, respectively.

SHBa, SHHBd: The group-type E-State descriptors SHBa and SHHBd relate to toxicophores that have hydrogen bond acceptor or donor atoms. SHHBa is the sum of E-State values for acceptors; SHHBd is the sum of hydrogen E-State values for donors.

SHCsat: The E-State descriptor SHCsats encodes E-State values for hydrogens on sp3 hybrid carbons bonded only to other sp3 carbon atoms. The electron accessibility of these sp3 hydrogens may relate in some manner to hydrophobic interactions between substrates and DNA or may have a relation to alkyl chlorides that are known toxicophores.
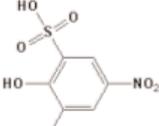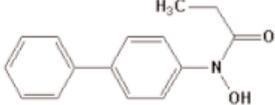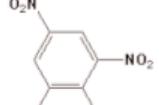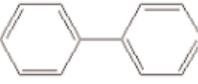
Kappa Shape Descriptors

phia and **Ka1**: The Kappa shape descriptors phia and Ka1 were found to be important in the DF consensus model. The Ka1 index encodes information about the cyclicity of a molecule where long, straight chain molecules have the highest values and the value decreases with increases in globularity and cyclicity. The phia index relates to molecular flexibility and increases in value with greater molecular flexibility. The phia index is related to the number of rotatable bonds.

With the exception of phia and Ka1, all of the important descriptors selected by the three modeling methods are directly related to the atom-type and bond-type E-State descriptors. The molecular connectivity chi indices demonstrate a more limited impact on the predicted genotoxicity value. Connectivity indices account for branching, type and composition of polycyclic aromatic ring systems, and the structural patterns of members of the train set. The low-order connectivity indices relate to fundamental parameters such as molecular surface area, molecular weight and the presence of heteroatoms. Such descriptors may function to provide discrimination among multiple structural classes. The higher order connectivity indices provide information about more complex skeletal variation and may also perform a classification function.

Column three of Table III shows the trend for the ten most important variables in the ANN model. The trend for seven of the descriptors shows that increased Ames mutagenic potency is associated with an increase in the descriptor magnitude. Two descriptors, SHBint2 and Gmax, showed a variable trend. The topological polarity descriptor Q_v shows a negative trend, as molecules with smaller values for Q_v tend to have larger predicted genotoxicity values. SHBint2 was selected for all the models with varying degrees of importance. It is the group-type E-State of a hydrogen donor and acceptor separated by two skeletal bonds (e.g., a carboxylic acid, primary or secondary amide, or urea). There is no reported relationship between a structure descriptor that encodes SHBint2 information and a structural alert for genotoxicity.

An interpretation of the affect that the number, positions, and type of ring substituents have on mutagenic potency is an indication of the sensitivity of the three models presented in this study (see illustration in Figure 4). Examination of the first set of four structures in the left column indicates that they all have two or more of the same substituents with a small number of changes in their relative ring positions. Yet, replacement of a sulfonic acid group in the non-mutagen (structure 3 in Figure 4) with a nitro group in structure 4 converts it to a mutagen, which was predicted correctly by all the models. Again, in structures 5-8 the affects of the introduction of a different substituent or an absence of one was correctly predicted by the ANN-QSAR for structures 7 and 8 based on two somewhat similar templates in the train set. The kNN-QSAR and the DF model incorrectly predicted the genotoxicity for structure 7, indicating somewhat less sensitivity in this case to affects of substituents on potency. Given the diversity in the learning techniques used among these models, some differences in predictability with specific compounds or classes are to be expected.


Figure 4. Examples of the sensitivity by ANN-QSAR to the presence and position of substituent groups in determining mutagenic potency of structures shown in left and right columns for train and validation compounds. Similar sensitivities were also found for the k-NN and DF models.

Comparison to other Methods

Two other modeling methods, multilinear and logistical regression, were examined in the course of this investigation. The results were not presented in this study because these methods produced models that proved to be rather poor predictors when applied to the subset of therapeutic drugs in training. The external validation results with drugs for these methods showed an outcome that was nearly random. It is, however, of interest to make comparisons with the validation results on genotox predictability obtained from other regressional and ruled-based models.

TOPKAT is a commercial QSAR regression model employing Kier and Hall indices. MULTICASE is a commercial QSAR regression model that uses fragments and rules, and DEREK is strictly a rule-based commercial program. A reported validation study (Pearl, 2001) used a 55/45 ratio of Ames mutagens to non-mutagens within 516 compounds obtained from publicly available sources. After elimination of indeterminates (<3%), TOPKAT, MULTICASE, and DEREK gave concordances of 70%, 81%, and 56%, respectively, with an average for all three of these models of 18% and 13% respectively for false positive and negatives. In this study, 361 external validation compounds gave concordances of 84%, 84%, 81% for ANN, kNN, and DF respectively with an average for all models of 6% and 10% for false positives and negatives as shown in Figure 5.
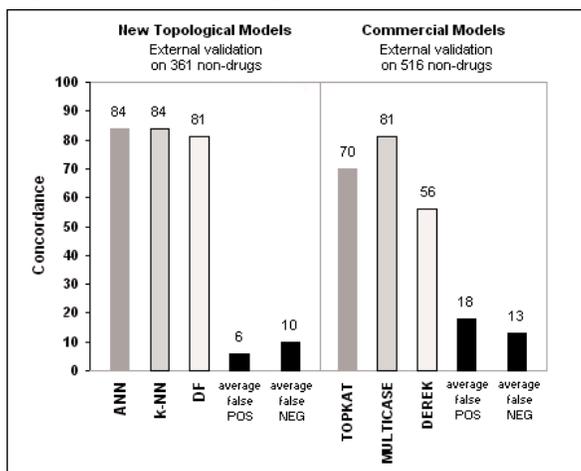
Figure 5. Comparison of the 3 models discussed in this paper with three available commercial models. The statistical comparisons are calculated for non-drug NEC's and are based predictive concordance, average false positive predictions and average false negative predictions.

When confidence criteria (all three current models) are applied, these models reveal their high quality by yielding an average of 92% in concordance that translates into a 14% to 64% improvement in accuracy over their commercial counterparts. Although there is a size difference in the validation sets, it is very doubtful this would account for such substantial differences in the predictability found between these two sets of in silico predictors. All models very likely use genotoxicity data from the same published sources. A fair comparison of the predictive capability for Ames mutagenicity with drugs could not be made because of a lack of available data for comparison. Although the commercial models had an average concordance of 67%±3% with 126 drugs in validation testing (Pearl, 2001) as well as a similar performance for two of these models in another study (Carriello et al., 2002), the 39 drugs used in the validation set of this study may be too small a number to make a meaningful comparison.

Conclusions

The important factors in the development of robust predictive models for Ames mutagenicity, as found in this study, include the use of consensus models, structure description based on topological descriptors and a large, diverse set of training compounds. The nature of the models developed by the three methods allows the potential for identification of less reliable predictions yielding excellent concordance values (91% or higher) after such outcomes have been removed. We conclude that this result is a key attribute of the ANN, kNN, and Decision Forest models. It is reported that S. typhimurium mutagenicity testing, performed by experienced laboratories using standardized protocols, leads to reproducibility between 84 to 87% (Piegorisch and Zeiger, 1991). This analysis of experimental accuracy suggests that there is an upper limit to the predictive ability of genotoxicity models that corresponds to the experimental limit of ~ 87%. The models reported here yield predictions for external validation testing that are within 3 to 5% of this upper limit. For the three models developed here, ANN, kNN and DF have approximately the same level of robustness for all compounds excepting therapeutic drugs. The ANN model shows somewhat better performance than the kNN and DF models for this class of compounds. Decision Forest performs equally well as the other QSAR models (ANN and kNN) in overall validation results before the application of its reliability criteria for high-low confidence. The DF model shows moderate performance in drug genotoxicity predictability. Furthermore, the use of the DF reliability criteria appears to cause both over and under prediction and leads to the removal of a large number of predictions deemed less reliable. This was not the case with the ANN or kNN models. Despite these limitations, the DF is a strong competitor to DEREK for a comparison of rule-based and classification-driven predictions for genotoxicity.

The modeling approach described in this study has produced very strong QSAR models for aqueous solubility and human intestinal absorption (Votano 2004 and Lobell 2004). In our overall methodology, heavy emphasis is placed on the use of representative external validation test sets. The strength of these models arises from a combination of high quality topological structure information representation, well-developed non-linear modeling methods, and data sets that have been carefully constructed and screened for anomalies. The appreciable degree of commonality among the most

important descriptors selected by the three modeling methods is also of interest, especially in light of the diverse methods (training (DF) versus learning (ANN and kNN) and modeling algorithms used in this study.

References:

Adamson, G. W. and Bawden, D. (1976) An empirical method of structure activity correlation for polysubstituted cyclic compounds using Wiswesser line notation. J. Chem. Inf. Comput. Sci., 16, 161-169.

Adamson, G. W. and Bush, J. A. (1976) Evaluation of an empirical structure-activity relationship for property prediction in a structurally diverse group of local anesthetics. J. Chem. Soc. Perkin Trans. 1, 1, 168-179.

Analyse-It (2003), Analyse-It Software, Ltd. PO Box 103, Leeds LS27 7WZ, United Kingdom

Ashby, J. and Tennant, R.W. (1988) Chemical-Structure, Salmonella Mutagenicity and Extent of Carcinogenicity as Indicators of Genotoxic Carcinogenesis Among 222 Chemicals Tested in Rodents by the United-States NCI/NTP. Mutat. Res.,. 204, 17-115.

Ashby, J.; Tennant R.W.; Zeiger, E. and Stasiewiz, S. (1989) 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, 73-103.

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

Bacha, P.A., Gruver, H.S., Den Hartog, B.K., Tamura, S.Y., Nutt, R.F. (2002) Rule extraction from a mutagenicity data set using adaptively grown phylogenetic-like trees. J. Chem. Inf. Comput. Sci., 42, 1104-1111.

Baska, S.C., Mills, D.R., Balaban, A.T., Gute, B.D. (2001) Prediction of mutagenicity of aroamyc and heteroaromatic amines from structure: A hierarchical approach. J. Chem. Inf. Comput. Sci., 41, 671-678.

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

ChemSilico LLC, 48 Baldwin Road, Tewksbury, MA 01876 (http://www.chemsilico.com/)

Devillers, J (1996) Strengths and Weaknesses of the back propagation neural network in QSAR and   QSPR studies. In Devillers, J (Ed.), Genetic Algorithms in Molecular

Modeling. Academic Press, pp. 1-24.

Golbraikh, A. and Tropsha, A. (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J. Comput. Aided. Mol. Des., 16, 357-369.

Gold, L.S. and Zeiger, E. (1996) Handbook of Carcinogenic Potency and Genotoxicity Databases, CRC Press.

Gough, J. and Hall, L.H. (1999) Modeling the toxicity of amide herbicides using the electrotopological state. J.   Environ. Tox. Chem., 18, 1069-1075.

Hall, L. H. (2004) A Structure-Information Approach to Prediction of Biological Activities and Properties. Chem & Biodiversity (in press).

Hall, L.H. and Vaughn, T.A. (1997) QSAR of phenol toxicity using electrotopological state and kappa shape indices. Med. Chem. Res., 7, 407-416.

Hall, L.H.; Maynard, E.L.; and Kier, L.B. (1989a) Structure-activity relationship studies on the toxicity of benzene derivatives: III. Predictions and extension to new substituents. J. Environ. Toxicol. Chem., 8, 431-436.

Hall, L.H.; Maynard, E.L.; and Kier, L.B. (1989b) QSAR investigation of benzene toxicity to fathead minnows using molecular connectivity. J. Environ. Toxicol. Chem., 8,783-788.

Hall; L.H. and Kier, L.B. (1995) Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information, J. Chem. Inf. Comput. Sci., 35, 1039-1045.

Hoffman, B.; Cho, S.J.; Zheng, W.; Wyrick, S.; Nichols, D.E.; Mailman, R.B.; Tropsha, A. (1999) Quantitative structure-activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. J. Med. Chem., 42, 3217-3226.

Hopfinger, A.J. and Patel, H.,C. (1996) Application of genetic algorithms to the general QSAR problem   and   to guiding molecular diversity experiments. In Devillers, J (Ed.), Genetic Algorithms in Molecular Modeling. Academic Press, pp. 131-154

Kier; L.B.and Hall, L.H. (1999) Molecular Structure Description; The Electrotopological State, Academic Press,San Diego, CA

King, R.D., Muggleton, S.H., Srinivasan, A., Sternberg, M.J. (1996) 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, 438-442.

Kirkpatrick, S.; Gelatt, C.D., Jr.; Vecchi, M.P.(1983) Optimization by Simulated Annealing. Science, 220, 671-680.

Klopman, G., Frierson, M.R., Rosenkranz, H.S. (1990) The structural basis of the mutagenicity of   chemicals   in Salmonella typhimurium: the Geno-Tox Data Base. Mutat. Res., 228, 1-50

Klopman, G., Frierson, M.R., Rosenkranz, H.S. (1984) Structural requirements for mutagenicity of environmental nitroarenes. Mutat. Res., 126, 277-238

Kubo, T., Urano, K., Utsumi, H. (2002) Mutagenicity characteristics of 255 environmental chemicals. J. Health Sci.,48, 545-554.

Lobell, M., Sivarajah, V. (2004) In Silico Prediction of Aqueous Solubility, Human Plasma Protein Binding, and Volume Distribution From Calculated pKa and AlogP98 Values. J. Molec. Diversity (in press).

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

Mersch-Sundermann,V., Rosenkranz, H.S., Klopman, G. (1992) Structural basis of polycyclic aromatic hydrocarbons. Mutagenesis, 7,211-218.

Miller, A. (2002) "Subset Selection in Regression", 2nd Edition Chapman & Hall/CRC Press

Miller, J.A.and Miller, E.C. (1977) Ultimate chemical carcinogen as reactive mutagenic electrophiles. In Origin of Human Cancers; Hiatt, H.H., Watson, J.D. and Winsten, J.A., (Eds) Cold Spring Harbor Laboratory Press: Cold   S p r i n g Harbor, NY, pp 605-627.

Pearl, G.M., Livingston-Carr, S., Durham, S.K. (2001) Integration of computational analysis as a sentinel tool in toxicological assessments, Curr.Topics Med. Chem., 1, 247-255

Piegorsch, W.W. and Zeiger, E. (1991) Measuring intra-assay agreement for Ames Salmonella assay. In Hotorn, L (Ed), Statistical Methods in Toxicology, Lecture Notes in Informatics, Vol.43, Springer, Heidelberg, pp. 35-41.

Rose, K. and Hall, L. H., E-State Modeling of Fish Toxicity Independent of 3D Structure Information, SAR QSAR Environ. Res., 14, 113-129 (2003).

Sanderson, D.M. and Earnshaw, C.G. (1991) Computer prediction of possible toxic action from chemical structure; the DEREK system. Hum. and Exp. Toxicol. 10, 261-273.

Sutter, J.M.; Dixon, S.L.; Jurs, P.C. (1995) Automated Descriptor Selection for Quantitative Structure-Activity Relationships Using Generalized Simulated Annealing. J. Chem. Inf. Comput. Sci., 35, 77-84.

Tennant, R.W. and Ashby, J.(1991) Classification according to chemical structure, mutagenicity to Salmonella   and level of carcinogenicity of a further 39 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. Mutat. Res., 257, 209-227.

Tong, W., Hong, H., Fang, H., Xie, Q., Perkins, R. (2003) Decision Forest: Combining the predictions of multiple independent decision tree models. J. Chem. Inf. Comput. Sci., 43, 525-531

TOPKAT software 2003; Accelrys, Inc., San Diego, CA 92121

Votano, J. R., Parham, M., Hall, L. H., Kier, L. B., Hall L. M. (2004) Prediction of Aqueous Solubility Based on Large Datasets Using Several QSPR Models Utilizing Topological Structure Representation. J. Chem. Inf. Comput. Sci. (in press)

Zeiger, E., Ashby, J., Bakale, G., Enslein, K., Klopman, G., Rosenkranz, H.S. (1996) Prediction of Salmonella mutagenicity. Mutagenesis, 11, 471-484.

Zheng, W. and Tropsha, A. (2000) Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. J. Chem. Inf. Comput. Sci., 40, 185-194.