Use of Statistical and Neural Net Approaches in Predicting Toxicity of Chemicals

Subhash C. Basak,*,† Gregory D. Grunwald,† Brian D. Gute,† Krishnan Balasubramanian,‡ and David Opitz§

Natural Resources Research Institute, University of Minnesota, Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811, Department of Chemistry and Biochemistry, Physical Sciences Building, Arizona State University, D-106 Tempe, Arizona 85287-1604, and Department of Computer Science, Social Science Building, University of Montana, Missoula, Montana 5981

Received September 9, 1999

Hierarchical quantitative structure−activity relationships (H-QSAR) have been developed as a new approach in constructing models for estimating physicochemical, biomedicinal, and toxicological properties of interest. This approach uses increasingly more complex molecular descriptors in a graduated approach to model building. In this study, statistical and neural network methods have been applied to the development of H-QSAR models for estimating the acute aquatic toxicity (LC_{50}) of 69 benzene derivatives to *Pimephales promelas* (fathead minnow). Topostructural, topochemical, geometrical, and quantum chemical indices were used as the four levels of the hierarchical method. It is clear from both the statistical and neural network models that topostructural indices alone cannot adequately model this set of congeneric chemicals. Not surprisingly, topochemical indices greatly increase the predictive power of both statistical and neural network models. Quantum chemical indices also add significantly to the modeling of this set of acute aquatic toxicity data.

1. INTRODUCTION

An important aspect of modern toxicology research is the prediction of toxicity of xenobiotics and environmental pollutants from their molecular structure.[1−13] The potential toxicity of a chemical is normally assessed on the basis of a wide variety of relevant physical and biological properties. Table 1 provides a partial list of such properties. Risk assessors use these kinds of toxicological indicators to estimate the potential risk posed by a given compound, using simpler properties relevant to a chemical's toxicity to make more complex assessments relevant to human and environmental health. However, the Toxic Substances Control Act (TSCA) Inventory currently includes about 80 000 chemicals, most of which do not have data for the toxicologically relevant properties mentioned in Table 1. In fact, roughly 50% of these chemicals do not have any experimental property data at all.[14] Worldwide, more than 16.7 million distinct organic and inorganic chemicals are known, as is evident from the number of entries in the Chemical Abstract Service (CAS) inventory.[15] For many of these chemicals we do not have the data necessary for risk assessment. Additionally, modern combinatorial chemistry techniques have led to the production of vast libraries of chemicals at a very rapid rate. Most of these substances have none of the test data needed for their hazard estimation.

Recently there have been efforts by the chemical industry and government agencies to develop reliable databases of properties that will be used for hazard estimation.[16] This effort, although commendable, falls short of the need; and the picture will remain so in the foreseeable future. In the area of molecular biology, innovative techniques are emerging where specially engineered cell lines can be used to detect the activity or toxicity of chemicals to the genetic system.[17−19] Effects of chemicals on the pattern of cellular proteins, analyzed by proteomics technology, are being used to detect their potential toxic effects.[20−22] Such methods are faster than the traditional in vivo test methods, and it is possible that they could be developed to the point where they will replace or significantly decrease the need for whole-animal screening methods. At present, neither the available test data nor the combination of in vitro toxicity testing methods provides adequate resources for hazard assessment.

Quantitative structure−activity/−toxicity relationship (QSAR/QSTR) models have emerged as useful tools to handle the data gap in toxicology and pharmacology.[1−13,22−26] QSAR models can be used to estimate complex properties of chemicals from simpler experimental or computed proper-

Table 1. Physicochemical and Biological Properties Relevant to the Assessment of toxicity

physicochemical	biological
molar volume	receptor binding (K_D)
boiling point	Michaelis constant (K_m)
melting point	inhibitor constant (K_i)
vapor pressure	biodegradation
aqueous solubility	bioconcentration
dissociation constant (pK_a)	alkylation profile
partition coefficient	metabolic profile
octanol−water (log P)	chronic toxicity
air−water	carcinogenicity
sediment−water	mutagenicity
reactivity (electrophile)	acute toxicity
	LD_{50}
	LC_{50}

* To whom all correspondence should be addressed. Telephone: (218) 720-4230. Fax: (218) 720-4328. E-mail: sbasak@nrri.umn.edu.
† University of Minnesota, Duluth.
‡ Arizona State University.
§ University of Montana.

10.1021/ci9901136 CCC: $19.00
Published on Web 06/22/2000

ties. In view of the fact that most chemicals in commerce and environmental pollutants have very little test data, it would be desirable if we could develop toxicologically relevant QSARs from properties that can be calculated directly from a chemical's structure. In some of our recent papers we have developed a novel hierarchical QSAR (H-QSAR) approach where four classes of theoretical molecular descriptors, viz., topostructural, topochemical, geometrical, and quantum chemical parameters, have been used sequentially in the formulation of H-QSAR models for predicting physical, biomedicinal, and toxicological properties.[1,3,6,8,23−26]

Most of our H-QSARs are based on linear statistical methods such as multiple linear regression, principal components analysis (PCA), and variable clustering. Such methods yield useful models, but they suffer from the limitation that in some cases the relationship between a molecular descriptor and toxicity may be intrinsically nonlinear. In such cases, the use of linear statistical methods may not result in the best models. Therefore, in this paper, we have carried out a comparative study of multiple regression vis-à-vis neural net methods in predicting the acute aquatic toxicity (LC_{50}) of a set of 69 benzene derivatives.

2. METHODS

2.1. Toxicity Database. The utility of this approach of generating numerous hierarchical theoretical descriptors of compounds was tested on a set of acute aquatic toxicity (LC_{50}) data for 69 benzene derivatives. The data were taken from a study by Hall et al.,[12] who collected acute aquatic toxicity data measured in fathead minnow (*Pimephales promelas*). These data were compiled from eight other literature sources and included some original work which was conducted at the U. S. Environmental Protection Agency Environmental Research Laboratory (USEPA-ERL) in Duluth, MN. This set of chemicals was composed of benzene and 68 substituted benzene derivatives. According to the authors, these benzene derivatives were tested using methodologies comparable to their own 96-h fathead minnow toxicity test system. The derivatives chosen for this study (see Table 2) have seven different substituent groups that are present in at least six of the molecules: chloro-, bromo-, nitro-, methyl-, methoxyl-, hydroxyl-, and amino-.

2.2. Calculation of Topological Indices. The complete set of topological indices (TIs) used in this study, both topostructural and topochemical, have been calculated using POLLY 2.3 and other software developed by Basak et al.[27] These indices include the Wiener index,[28] the connectivity indices developed by Randić,[29] higher order connectivity indices formulated by Kier and Hall,[30] bonding connectivity indices defined by Basak et al.,[31] a set of information theoretic indices defined on the distance matrices of simple molecular graphs,[32,33] a set of parameters derived on the neighborhood complexity of hydrogen-filled molecular graphs,[34−36] and Balaban's *J* indices.[37−39] Table 3 provides the symbols of the topological indices and brief definitions.

The set of TIs was divided into two distinct subsets: topostructural indices (TSI) and topochemical indices (TCI). TSIs are topological indices which encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of the chemical nature of the atoms involved in the bonding or factors such as hybridiza-

Table 2. Experimental and Estimated Acute Aquatic Toxicity Data for 69 Benzene Derivatives, Expressed as $-\log(LC_{50})$ for the Linear Regression Model (LR) and the Neural Network Model Using the 23 Parameters Selected by Variable Clustering

compound	expt	LR	NN
benzene	3.40	3.42	3.65
bromobenzene	3.89	3.77	3.79
chlorobenzene	3.77	3.75	3.77
phenol	3.51	3.38	3.51
toluene	3.32	3.66	3.62
1,2-dichlorobenzene	4.40	4.29	4.30
1,3-dichlorobenzene	4.30	4.37	4.12
1,4-dichlorobenzene	4.62	4.51	4.27
2-chlorophenol	4.02	3.79	3.91
3-chlorotoluene	3.84	3.88	3.79
4-chlorotoluene	4.33	3.87	3.76
1,3-dihydroxybenzene	3.04	3.43	3.53
3-hydroxyanisole	3.21	3.33	3.45
2-methylphenol	3.77	3.64	3.67
3-methylphenol	3.29	3.60	3.58
4-methylphenol	3.58	3.53	3.55
4-nitrophenol	3.36	3.61	3.76
1,4-dimethoxybenzene	3.07	3.28	3.51
1,2-dimethylbenzene	3.48	3.93	3.91
1,4-dimethylbenzene	4.21	3.87	3.68
2-nitrotoluene	3.57	3.66	3.81
3-nitrotoluene	3.63	3.53	3.71
4-nitrotoluene	3.76	3.49	3.68
1,2-dinitrobenzene	5.45	5.24	4.99
1,3-dinitrobenzene	4.38	4.18	4.19
1,4-dinitrobenzene	5.22	4.94	4.85
2-methyl-3-nitroaniline	3.48	3.79	3.88
2-methyl-4-nitroaniline	3.24	3.51	3.75
2-methyl-5-nitroaniline	3.35	3.68	3.86
2-methyl-6-nitroaniline	3.80	3.84	3.79
3-methyl-6-nitroaniline	3.80	3.78	3.62
4-methyl-2-nitroaniline	3.79	3.80	3.66
4-hydroxy-3-nitroaniline	3.65	3.61	3.58
4-methyl-3-nitroaniline	3.77	3.73	3.72
1,2,3-trichlorobenzene	4.89	4.89	5.04
1,2,4-trichlorobenzene	5.00	5.04	4.83
1,3,5-trichlorobenzene	4.74	5.11	4.78
2,4-dichlorophenol	4.30	4.33	4.47
3,4-dichlorotoluene	4.74	4.26	4.28
2,4-dichlorotoluene	4.54	4.36	4.44
4-chloro-3-methylphenol	4.27	3.87	4.07
2,4-dimethylphenol	3.86	3.76	3.72
2,6-dimethylphenol	3.75	3.80	3.84
3,4-dimethylphenol	3.90	3.80	3.79
2,4-dinitrophenol	4.04	4.14	4.01
1,2,4-trimethylbenzene	4.21	4.09	3.87
2,3-dinitrotoluene	5.01	5.20	5.28
2,4-dinitrotoluene	3.75	4.10	4.33
2,5-dinitrotoluene	5.15	4.84	4.72
2,6-dinitrotoluene	3.99	4.41	4.63
3,4-dinitrotoluene	5.08	5.11	5.09
3,5-dinitrotoluene	3.91	4.05	4.16
1,3,5-trinitrobenzene	5.29	5.37	5.32
2-methyl-3,5-dinitroaniline	4.12	4.13	4.23
2-methyl-3,6-dinitroaniline	5.34	4.80	4.54
3-methyl-2,4-dinitroaniline	4.26	4.28	4.20
5-methyl-2,4-dinitroaniline	4.92	4.14	4.02
4-methyl-2,6-dinitroaniline	4.21	4.67	4.58
5-methyl-2,6-dinitroaniline	4.18	4.80	4.78
4-methyl-3,5-dinitroaniline	4.46	4.34	4.43
2,4,6-tribromophenol	4.70	4.89	5.47
1,2,3,4-tetrachlorobenzene	5.43	5.62	5.56
1,2,4,5-tetrachlorobenzene	5.85	5.80	5.61
2,4,6-trichlorophenol	4.33	4.79	4.96
2-methyl-4,6-dinitrophenol	5.00	4.21	4.16
2,3,6-trinitrotoluene	6.37	6.36	5.81
2,4,6-trinitrotoluene	4.88	5.16	5.42
2,3,4,5-tetrachlorophenol	5.72	5.36	5.58
2,3,4,5,6-pentachlorophenol	6.06	6.03	5.83

Table 3. Symbols, Definitions, and Classifications of Topological, Geometrical, and Quantum Chemical Parameters

	Topostructural
$\underline{I}_D^W$	information index for the magnitudes of distances between all possible pairs of vertexes of a graph
$\overline{I}_D^W$	mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	degree complexity
H^V	graph vertex complexity
$\underline{H}^D$	graph distance complexity
$\overline{IC}$	information content of the distance matrix partitioned by frequency of occurrences of distance h
O	order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	a Zagreb group parameter = sum of square of degree over all vertexes
M_2	a Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertexes
${}^h\chi$	path connectivity index of order $h = 0-6$
${}^h\chi_C$	cluster connectivity index of order $h = 3, 5$
${}^h\chi_{Ch}$	chain connectivity index of order $h = 6$
${}^h\chi_{PC}$	path-cluster connectivity index of order $h = 4-6$
P_h	no. of paths of length $h = 0-10$
J	Balaban's J index based on distance
	Topochemical
I_{ORB}	information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertexes
IC_r	mean information content or complexity of a graph based on the rth ($r = 0-6$) order neighborhood of vertexes in a hydrogen-filled graph
SIC_r	structural information content for r^{th} ($r = 0-6$) order neighborhood of vertexes in a hydrogen-filled graph
CIC_r	complementary information content for rth ($r = 0-6$) order neighborhood of vertexes in a hydrogen-filled graph
${}^h\chi^b$	bond path connectivity index of order $h = 0-6$
${}^h\chi^b{}_C$	bond cluster connectivity index of order $h = 3, 5$
${}^h\chi^b{}_{Ch}$	bond chain connectivity index of order $h = 6$
${}^h\chi^b{}_{PC}$	bond path-cluster connectivity index of order $h = 4-6$
${}^h\chi^v$	valence path connectivity index of order $h = 0-6$
${}^h\chi^v{}_C$	valence cluster connectivity index of order $h = 3, 5$
${}^h\chi^v{}_{Ch}$	valence chain connectivity index of order $h = 6$
${}^h\chi^v{}_{PC}$	valence path-cluster connectivity index of order $h = 4-6$
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
	Geometrical
V_W	van der Waals volume
${}^{3D}W$	3D Wiener no. for the hydrogen-suppressed geometric distance matrix
${}^{3D}W_H$	3D Wiener no. for the hydrogen-filled geometric distance matrix
	Quantum Chemical
E_{HOMO}	energy of the highest occupied molecular orbital
E_{HOMO1}	energy of the second highest occupied molecular orbital
E_{LUMO}	energy of the lowest unoccupied molecular orbital
E_{LUMO1}	energy of the second lowest unoccupied molecular orbital
ΔH_f	heat of formation
μ	dipole moment

tion states of atoms and number of core/valence electrons in individual atoms. TCIs are parameters that quantify information regarding the topology (connectivity of atoms), as well as specific chemical properties of the atoms and bonds comprising a molecule. TCIs are derived from weighted molecular graphs where each vertex (atom) is properly weighted with relevant chemical/physical properties. Table 3 shows the division of the topological indices into topostructural and topochemical indices.

2.3. Calculation of Geometrical Indices. The geometrical indices include the three-dimensional (3D) Wiener numbers for hydrogen-filled and hydrogen-suppressed molecular structures and van der Waals volume. van der Waals volume, V_W, was calculated using SYBYL 6.4 from Tripos Associates, Inc.[40] The 3D Wiener numbers were calculated by SYBYL using an SPL (Sybyl Programming Language) program developed in our laboratory. Calculation of the 3D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3D coordinates for the atoms were determined using CONCORD 3.2.1.[41] The symbols and definitions of the geometrical indices are included in Table 3.

2.4. Quantum Chemical Parameters. Quantum chemical parameters were calculated using the Austin Model version one (AM1) semiempirical Hamiltonian. These parameters were calculated using MOPAC 6.00 in the SYBYL interface.[42] Brief definitions and symbols for the quantum chemical parameters used in this study are included in Table 3.

2.5. Statistical Analysis and Hierarchical QSAR. Initially, all topological indices were transformed by the natural logarithm of the index plus one. This was done to scale the indices, since some may be several orders of magnitude greater than others, while other indices may equal zero. The geometric indices were transformed by the natural logarithm of the index for consistency; the addition of one was unnecessary.

The set of 86 topological indices was then partitioned into the two distinct sets: topostructural indices (35) and topochemical indices (51). The sets of topostructural and topochemical indices were then divided into subsets, or clusters, based on the correlation matrix using the SAS variable clustering procedure (VARCLUS)[43] to further reduce the number of independent variables for use in model construction. This procedure divides the set of indices into

disjoint clusters, such that each cluster is essentially unidimensional.

From each cluster, the index most correlated with the cluster was selected for modeling, as well as any indices that were poorly correlated with their cluster ($R^2 < 0.70$). These indices were then used in the modeling of the acute aquatic toxicity of benzene derivatives in fathead minnow. The variable clustering and selection of indices was performed independently for both the topostructural and topochemical indices. This procedure resulted in a set of five topostructural indices and a set of nine topochemical indices.

Reducing the number of independent variables is critical when attempting to model small data sets using linear statistical methods. The smaller the data set, the greater the chance of spurious error when using a large number of independent variables (descriptors). A study by Topliss and Edwards[44] has shown that for a set with about 70 dependent variables (observations), no more than 40 independent variables may be used while keeping the probability of chance correlations below 1%. This number is dependent on the actual correlation achieved in the modeling process; higher correlation results in a better chance of using more variables with the same limited probability of chance correlations. In this study we are well below the cutoff of 40 independent variables. In fact, the total number of descriptors which will be used for model construction and estimation is 23, well within the bounds of the Topliss and Edwards criteria.[44]

Regression modeling was accomplished using the SAS procedure REG[43] on four distinct sets of indices. These sets were constructed as part of a hierarchical approach to QSAR model development. The hierarchy begins with the simplest parameters, the TSIs. After using the TSIs to model the activity, the next level of parameters are added. To the indices included in the best TSI model, we add all of the TCIs and proceed to model the activity using these parameters. Likewise, the indices included in the best model from this procedure are combined with the indices from the next complexity level, the geometrical indices, and modeling is conducted once again. Finally, the best model utilizing TSIs, TCIs, and geometrical indices is combined with the quantum chemical parameters to develop the final model in the hierarchy.

Additionally, the entire set of 95 descriptors (topostructural, topochemical, geometrical, and quantum chemical) was subjected to the variable clustering procedure and a reduced set of independent variables was used in constructing a QSAR model. This varies from the other approach in that the indices were clustered as one set, rather than as four distinct sets, and resulted in a somewhat different set of variables. This was done to determine if there is any advantage in final model predictive power between model development based on the H-QSAR approach versus the "kitchen sink" approach, i.e., using the entire descriptor set in order to find the "best" model.

2.6. Neural Network Methods. Using neural networks, we studied two classes of approaches for modeling toxicity: (1) giving all the descriptors to a learning algorithm (neural network in this case) and (2) reducing the feature set before giving the (reduced) feature set to a learning algorithm. Results for our approaches are from leave-one-out experiments (i.e., 69 training/test set partitions). Leave-one-out works by leaving one data point out of the training set and giving the remaining instances (68 in this case) to the learning algorithms for training. This process is repeated 69 times so that each example is a part of the test set once and only once. Leave-one-out tests *generalization* accuracy of a learner, whereas training set accuracy tests only the learner's ability to memorize. Generalization error from the test set is the true test of accuracy and is what we report here.

First we trained neural networks using all 95 parameters: 35 TSI, 51 TCI, 3 geometrical, and 6 quantum chemical parameters. The networks contained 15 hidden units and were trained for 1000 epochs. Each input parameter was normalized to a value between 0 and 1 before training. Additional parameter settings for the neural networks included a learning rate of 0.05, a momentum term of 0.1, and weights initialized randomly between -0.25 and $+0.25$.

For our next experiment, we used a smaller set of 23 independent variables divided further into the four levels of the hierarchy. The 23 independent variables included the 5 topostructural and 9 topochemical parameters provided by the variable clustering technique (see section 3.1 for a list of the indices) combined with the 3 geometrical and 6 quantum chemical parameters described in Table 3. The parameter settings for these networks were the same as the settings for the other neural network experiment mentioned above.

3. RESULTS

3.1. Results of Statistical Regression Procedures. The variable clustering of the topostructural indices resulted in the retention of five indices: M_1, $\overline{\mathrm{IC}}$, O, P_8, P_9. All-subsets regression resulted in the selection of a four-parameter model to estimate $-\log(\mathrm{LC}_{50})$ with an explained variance (R^2) of 45.3% and a standard error (s) of 0.58. While this is an unsatisfactory model, the indices were retained and combined with the topochemical indices in the second step of model development. The second step combined the 4 indices used in the first tier model with the 9 topochemical indices selected in the variable clustering procedure: SIC_0, SIC_1, SIC_4, CIC_0, ${}^2\chi^{\mathrm{b}}$, ${}^5\chi^{\mathrm{b}}{}_{\mathrm{C}}$, ${}^5\chi^{\mathrm{v}}{}_{\mathrm{C}}$, ${}^6\chi^{\mathrm{v}}{}_{\mathrm{PC}}$, J^{X}. Again, all-subsets regression was conducted resulting in a four-parameter model with an explained variance (R^2) of 78.3% and a standard error (s) of 0.36. The 4 indices from the second tier model were combined with the three geometric parameters: ${}^{3\mathrm{D}}W_{\mathrm{H}}$, ${}^{3\mathrm{D}}W$, V_{W}. This resulted in a four-parameter model that replaced the topochemical index CIC_0 with the geometric parameter ${}^{3\mathrm{D}}W_{\mathrm{H}}$. This model had an explained variance (R^2) of 79.2% and a standard error (s) of 0.36. The final step in the hierarchical method combined the four parameters from the third tier model with the semiempirical quantum chemical parameters: E_{HOMO}, E_{HOMO1}, E_{LUMO}, E_{LUMO1}, ΔH_f, μ. This set of 10 indices led to a seven-parameter model with an explained variance (R^2) of 86.3% and a standard error (s) of 0.30. This model retained all indices from the third model and added three of the AM1 quantum chemical parameters. Our final model, using indices selected from a variable clustering of the entire set of 95 indices resulted in a seven-parameter model including three topostructural indices (${}^0\chi$, P_9, $\overline{\mathrm{IC}}$), one topochemical index (${}^5\chi^{\mathrm{v}}$), one geometrical index (${}^{3\mathrm{D}}W_{\mathrm{H}}$), and two quantum chemical descriptors (ΔH_f, μ). This model had an explained variance (R^2) of 86.1% and a standard error (s) of 0.30.

Table 4. Relative Effectiveness of Statistical and Neural Network Methods in Estimating the Acute Aquatic Toxicity of 69 Benzene Derivatives

	neural networks		linear regression	
model	R_c^2	s	R_c^2	s
TSI	0.299	0.63	0.366	0.629
+ TCI	0.619	0.47	0.754	0.392
+ 3D	0.656	0.44	0.763	0.384
+ QC	0.770	0.36	0.825	0.339
all 95 indices	0.758	0.37	0.827	0.337

Leave-one-out analysis was conducted on all models for purposes of comparison with the results from the neural networks. The resulting values for cross-validated R^2 (R_c^2) and standard error (s) are reported in Table 4.

3.2. Results of the Neural Network Procedures. The first approach incorporating all 95 parameters, obtained a test-set correlation coefficient between predicted toxicity and measured toxicity (explained variance) of $R^2 = 0.868$ and a standard error of 0.29. The second approach utilizes the hierarchical method of grouping descriptors resulted in four models, one for each level of the hierarchy. The results from the leave-one-out analysis of these four models, as well as those for the linear statistical models are summarized in Table 4. Table 2 presents the experimental acute aquatic toxicity ($-\log[LC_{50}]$) values for the 69 benzene derivatives as well as the values estimated by the best statistical model and the best neural network model, both of which resulted from the fourth H-QSAR model.

4. DISCUSSION

The results show that both statistical and neural network models give acceptable estimates for the toxicity of the 69 benzene derivatives studied in this paper. As can be clearly seen from the comparative results in Table 4, there are two points in the hierarchical approach in which there are significant improvements in modeling the data. The addition of the topochemical indices increases the variance explained in both the statistical and neural network models by 30−40% with a consequent drop in the standard error of the calculations as well. Addition of the quantum chemical parameters also creates a significant increase in the efficacy of both models, a 6.2% increase in the variance explained for the statistical model and an 11.4% increase for the neural network model.

It is interesting to note that the neural network model using the subset of 23 inputs selected in part by the VARCLUS procedure gave slightly better results as compared to the network developed using all 95 input variables. This could be the result of filtering out redundant, or nearly redundant, parameters from the set of independent variables.

Further work on the relative utility of statistical vis-à-vis neural network methods is necessary to determine which types of models are best suited to the estimation of chemical toxicity.

ACKNOWLEDGMENT

This paper is contribution no. 270 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by Grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force, Grant IRI-9734419 from the National Science Foundation, and a MONTS grant from the University of Montana.

REFERENCES AND NOTES

(1) Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach: Reading, U.K., 1999; pp 675−696.
(2) Basak, S. C. In *Practical Applications of Quantitative Structure−Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1990; pp 83−103.
(3) Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *QSAR in Environmental Sciences*, Vol. 7; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1998; pp 245−261.
(4) Basak, S. C.; Gute, B. D. In *Discrete Mathematical Chemistry*; Hansen, P., Paradis, N., Eds.; DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 51; American Mathematical Society: Providence, RI, 2000; pp 9−24.
(5) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Assessment of the mutagenicity of chemicals from theoretical structural parameters: A hierarchical approach. *SAR QSAR Environ. Res.* **1999**, *10*, 117−129.
(6) Gute, B. D.; Grunwald, G. D.; Basak, S. C. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. *SAR QSAR Environ. Res.* **1999**, *10*, 1−15.
(7) Basak, S. C.; Gute, B. D. Characterization of molecular structures using topological indices. *SAR QSAR Environ. Res.* **1997**, *7*, 1−21.
(8) Gute, B. D.; Basak, S. C. Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117−131.
(9) Mushrush, G. W.; Basak, S. C.; Slone, J. E.; Beal, E. J.; Basu, S.; Stalick, W. M.; Hardy, D. R. Computational study of the environmental fate of selected aircraft deicing compounds. *J. Environ. Sci. Health* **1997**, *A32* (8), 2201−2211.
(10) Basak, S. C.; Grunwald, G. D. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: A similarity based study. *Chemosphere* **1995**, *31*, 2529−2546.
(11) Basak, S. C.; Grunwald, G. D. In *Proceeding of the XVI International Cancer Congress*; Rao, R. S., Deo, M. G., Sanghui, L. D., Eds.; Monduzzi: Bologna, Italy, 1995; p 413.
(12) Hall, L.; Kier, L.; Phipps, G. Structure-activity relationship studies on the toxicities of benzene derivatives: I. An additivity model. *Environ. Toxicol. Chem.* **1984**, *3*, 355−365.
(13) Gombar, V. K.; Enslein, K.; Blake, B. W. Assessment of developmental toxicity potential of chemicals by quantitative structure-toxicity relationship models. *Chemosphere* **1995**, *31*, 2499−2510.
(14) Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of action and the assessment of chemical hazards in the presence of limited data: Use of structure-activity relationships (SAR) under TSCA, Section 5. *Environ. Health. Perspect.* **1990**, *87*, 183−197.
(15) CAS. The latest CAS registry number and substance count. http://www.cas.org/cgi-bin/regreport.pl, 2000.
(16) Johnson, J. Pact triggers tests: Thousands of chemicals may be tested under toxicity screening program. *Chem. Eng. News* **1998**, *76*, 19−20.
(17) Chen, J. J.; Wu, R.; Yang, P. C.; Huang, J. Y.; Sher, Y. P.; Han, M. H.; Kao, W. C.; Lee, P. J.; Chiu, T. F.; Chang, F.; Chu, Y. W.; Wu, C. W.; Peck, K. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* **1998**, *51*, 313−324.
(18) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**, *270*, 467−470.
(19) De Risi, J.; Penland, L.; Brown, P. O.; Bittner, M. L.; Meltzer, P. S.; Ray, M.; Chen, Y.; Su, Y. A.; Trent, J. M. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **1996**, *14*, 457−460.
(20) Witzmann, F. A.; Fultz, C. D.; Grant, R. A.; Wright, L. S.; Kornguth, S. E.; Siegel, F. L. Differential expression of cytosolic proteins in the rat kidney cortex and medulla: Preliminary proteomics. *Electrophoresis* **1998**, *19*, 2491−2497.
(21) Anderson, N. L.; Esquer-Blasco, R.; Richardson, F.; Foxworthy, P.; Eacho, P. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharm.* **1996**, *137*, 75−89.
(22) Lake, B. G.; Lewis, D. F. V.; Gray, T. J. B.; Beamand, J. A. Structure-activity relationships for induction of peroxysomal enzyme activities in primary rat hepatocyte cultures. *Toxicol. in Vitro* **1993**, *7*, 605−614.
(23) Basak, S. C.; Gute, B. D.; Grunwald, G. D.; Opitz, D. W.; Balasubramanian, K. In *Predictive Toxicology of Chemicals: Experiences and*

Impact of AI Tools-Papers from the 1999 AAAI Symposium; AAAI Press: Menlo Park, CA, 1999; pp 108−111.
(24) Basak, S. C.; Gute, B. D.; Ghatak, S. Prediction of complement−inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 255−260.
(25) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651−655.
(26) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054−1060.
(27) Basak, S.; Harriss, D.; Magnuson, V. *POLLY 2.3*; University of Minnesota: Duluth, MN, 1988.
(28) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(29) Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(30) Kier, L.; Hall, L. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Hertfordshire, U.K., 1986.
(31) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* **1988**, *19*, 17−44.
(32) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. Discrimination of isomeric structures using information theoretic topological indices. *J. Comput. Chem.* **1984**, *5*, 581−588.
(33) Bonchev, D.; Trinajstić, N. Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.
(34) Basak, S. C.; Roy, A. B.; Ghosh, J. J. In *Proceedings of the Second International Conference on Mathematical Modelling*, Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri − Rolla: rolla, MO, 1980; p 851.
(35) Roy, A. B.; Basak, S. C.: Harriss, D. K.; Magnuson, V. R. In *Mathematical Modelling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Lapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York, 1984; p 745.
(36) Basak, S. C.; Magnuson, V. R. Molecular topology and narcosis. *Arzneim.-Forsch./Drug Res.* **1983**, *33*, 501−503.
(37) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(38) Balaban, A. T. Topological indices based on topological distances in molecular graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206.
(39) Balaban, A. T. Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)* **1986**, *21*, 115−122.
(40) *SYBYL Version 6.4.*; Tripos Associates, Inc.: St. Louis, MO, 1998.
(41) *CONCORD Version 3.2.1.*; Tripos Associates, Inc.: St. Louis, MO, 1998.
(42) Stewart, J. J. P. *MOPAC 6.00*, QCPE #455; Frank J. Seiler Research Laboratory, U.S. Air Force Academy: Colorado Springs, CO, 1990.
(43) *SAS/STAT User's Guide*, 6.03 ed.; SAS Institute Inc.: Cary, NC, 1988; Chapters 28 and 34, pp 773−875, 949−965.

CI9901136