SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 17, No. 6, pp. 1056–1064, 1998
Printed in the USA
0730-7268/98 $6.00 + .00

A COMPARATIVE STUDY OF MOLECULAR SIMILARITY, STATISTICAL, AND NEURAL METHODS FOR PREDICTING TOXIC MODES OF ACTION

SUBHASH C. BASAK,*† GREGORY D. GRUNWALD,† GEORGE E. HOST,† GERALD J. NIEMI,† and STEVEN P. BRADBURY‡

†Natural Resources Research Institute, University of Minnesota, Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811, USA
‡National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, 6201 Congdon Boulevard, Duluth, Minnesota 55804

(*Received* 20 *May* 1997; *Accepted* 14 *October* 1997)

Abstract—Quantitative structure–activity relationship (QSAR) models are routinely used in predicting toxicologic and ecotoxicologic effects of untested chemicals. One critical factor in QSAR-based risk assessment is the proper assignment of a chemical to a mode of action and associated QSAR. In this paper, we used molecular similarity, neural networks, and discriminant analysis methods to predict acute toxic modes of action for a set of 283 chemicals. The majority of these molecules had been previously determined through toxicodynamic studies in fish to be narcotics (two classes), electrophiles/proelectrophiles, uncouplers of oxidative phosphorylation, acetylcholinesterase inhibitors, and neurotoxicants. Nonempirical parameters, such as topological indices and atom pairs, were used as structural descriptors for the development of similarity-based, statistical, and neural network models. Rates of correct classification ranged from 65 to 95% for these 283 chemicals.

Keywords—Toxic mode prediction Topological indices Molecular similarity Neural network Discriminant function analysis

INTRODUCTION

An important goal of research in toxicology is the prediction of toxic potential of chemicals, ranging from acute toxicity to complex endpoints associated with chronic exposures. Short-term tests, structural criteria, and experimental physicochemical properties have been used by experts in assessing hazards of chemical species. In assessing the carcinogenic potential of chemicals, Arcos [1] used a combination of structural criteria, functional criteria, and guilt by association criteria. Arcos believed that, in many instances, structural criteria were not sufficient to predict the carcinogenic potential of new structural classes of chemicals. Tennant et al. [2] used results of short-term tests and a list of structural characteristics or structural alerts in predicting carcinogenesis caused by chemicals in the 2-year rodent bioassay. Bahler and Bristol [3] applied induction-based methods to generate rules that can predict carcinogenicity of chemicals to rodents. Such rules contain both structural criteria and results of short-term tests of the chemicals.

In many practical situations of predictive toxicology, the combination of structural and functional criteria in estimating toxic potential of chemicals is impractical due to the lack of relevant toxicologic data. More than 15 million distinct chemical entities have been registered with the Chemical Abstract Service (CAS) and the list is growing by nearly 775,000 per year. About 1,000 of these chemicals enter into societal use every year [1]. Few of these chemicals have the empirical data needed for risk assessment. In the United States, the Toxic Substances Control Act (TSCA) inventory has about 74,000 entries and the list is growing by nearly 3,000 per year. Of the approximately 3,000 chemicals submitted yearly to the United States Environmental Protection Agency (U.S. EPA) for the premanufacture notification (PMN) process, more than 50% have no experimental data at all, less than 15% have empirical mutagenicity data, and about 6% have experimental ecotoxicologic and environmental fate data [4]. Also, limited data are available for many of the more than 700 chemicals found on the Superfund list of hazardous substances [4].

Under such circumstances, the U.S. EPA has taken a two-fold strategy in the hazard estimation of chemicals: application of chemical class-specific quantitative structure–activity relationships (QSARs), and use of analogs (similar chemicals). Currently, the U.S. EPA uses specific QSARs for more than 40 different chemical classes to predict toxicity of new industrial chemicals [4,5]. Such QSAR models are useful when the chemical can be unambiguously assigned to a class for which there is a good QSAR model.

In environmental toxicology, especially in aquatic toxicology, such first-generation QSARs have emerged as scientifically credible tools for predicting acute and, in some instances, subchronic toxicity of chemicals when little or no empirical data are available [4]. The success of such QSARs is dependent upon the availability of well-defined and quantifiable toxicity endpoints such as the 96-h median lethal concentration (LC50) values for fathead minnows (*Pimephales promelas*). Although the accuracy of predicting toxicity endpoints continues to improve, major uncertainty exists in the selection of appropriate QSARs for estimating hazardous potential of chemicals. Thus, proper application and success of these predictive toxicology techniques requires that we can first assign a molecule to its appropriate chemical class and then apply the class-specific QSAR in the realistic estimation of its hazardous potential. Currently, this fundamental process in the use of predictive

* To whom correspondence may be addressed (sbasak@sage.nrri.umn.edu).

toxicologic tools represents a major hurdle in ecological risk assessment. Inappropriate applications of QSARs can lead to 10- to 1,000-fold errors in toxic potency estimation [6].

Analogs of new chemicals are routinely used by regulatory agencies such as the U.S. EPA in hazard assessment [4]. Traditionally, the selection of analogs and assignment of a structural class to a chemical is based on the assumption that similar compounds or compounds from the same structural class should behave in a toxicologically similar manner [7]. Within a specified toxicologic context, a chemical C is an analog of (or similar to) another chemical D if C and D resemble each other in one or more relevant aspects, for example, structurally, stereoelectronically, or physicochemically.

In this paper, we have attempted to predict the acute mode of toxic action of a set of 283 chemicals tested with the fathead minnow [8] using parameters that can be calculated directly from the chemical's structure. Such parameters include numerical graph invariants or topological indices and substructural parameters such as atom pairs.

Recent studies have used structural descriptors in quantitative molecular similarity analysis (QMSA) methods, neural networks, and discriminant analysis. Quantitative molecular similarity analysis techniques using topological indices and atom pairs have been used to define spaces for analog selection [9–12] and have been shown to compare well with physicochemical property spaces [12]. Quantitative molecular similarity analysis methods have been used for estimation of physicochemical [13,14] and toxicologic [13,15] properties as well. Recent studies have also utilized neural network methods for estimating properties from structural descriptors [16–18]. Discriminant analysis is a statistical technique commonly used for classification situations and has been used to model diverse physicochemical and toxicologic properties [19,20].

We have carried out a comparative study of molecular similarity, neural network, and discriminant analysis methods in assigning modes of action to chemicals. These results are presented in this paper along with a critical evaluation of the effectiveness of different methods in predicting modes of action of chemicals from their theoretical structural parameters.

METHODS

Data set description

Chemicals selected for analysis were from the 617-chemical Mid-Continent Ecology Division–Duluth fathead minnow database, which was recently evaluated for mode of toxic action [8]. Two hundred eighty-three of these chemicals were selected, representing eight modes of action for which higher confidence was associated with final mode classification. Specifically, chemicals with A and B levels of confidence in Russom et al. [8] were used. The higher confidence was related to the amount and level of evidence used in determining the mode of action, with concurrent information from joint toxic action studies, physiologic and behavioral response data, information in peer-reviewed literature, and toxicity over time as well as toxicity to the fathead minnow serving as a basis for this determination [8].

The mode of action classes represented included narcosis I (baseline narcosis), narcosis II (polar narcosis), mixed narcosis I/II, oxidative phosphorylation uncoupling, acetylcholinesterase (AChE) inhibition, and electrophile/proelectrophile reactivity. An additional class of chemicals exhibited central nervous system (CNS) responses. Among these chemicals were insecticides associated with a distinct mechanism of action that elicited a similar mode of action at the organism level. These chemicals were assigned as neurotoxicants. An additional group of CNS chemicals, neurodepressants, as well as respiratory blockers were combined to form a final class for the purpose of these analyses. Neither the neurodepressants nor the respiratory blockers occurred in sufficient numbers to form a group with sufficient sample size to discriminate these chemicals from the other classes. For the exercise reported here, the objective was to determine whether topological characteristics could consistently establish that the combined group of neurodepressants and respiratory blocker chemicals are dissimilar to the more specifically defined narcotics, oxidative phosphorylation uncouplers, AChE inhibitors, and electrophile/proelectrophiles and the neurotoxic CNS agents.

The data set was divided into a training set consisting of 220 chemicals and a test set of 63 chemicals. The test set was used for validation of models developed with the training set of chemicals and never used for model development. The test set was determined by ordering the chemicals by CAS registry number within each mode of action and selecting every third compound. Because of the mixed mode of action for the narcosis I/II group, these chemicals were used as part of the test set of chemicals only. Appendix 1 lists the modes of action and total sample sizes in the training and test sets. During the classification phase of the test set, if these chemicals were classified as either narcosis I or narcosis II, the classification was considered to be correct. See below for classification methods.

Topological indices

Most numerical graph invariants used for the analyses of mode of action were calculated by the computer program POLLY [21]. The remaining topological indices were calculated with programs developed by the authors. Appendix 2 provides a listing of topological indices used in this paper.

Connectivity indices of different types were calculated following the methods of Randić [22] and Kier and Hall [23]. Information theoretic indices defined on hydrogen-suppressed and hydrogen-filled graphs were calculated using the methods of Basak et al. [24], Basak and Magnuson [25], Roy et al. [26], Raychaudhury et al. [27], and Bonchev and Trinajstić [28]. The Wiener number [29], Zagreb group indices [30], and J indices [31–33] were calculated by POLLY. Topological indices based on information on distances within the chemical graph were calculated using the methods of Balaban and Balaban [34]. Finally, a set of local graph invariants (LOVI) were calculated using the approach of Filip et al. [35]. Atom pairs were calculated by APPROBE [36] following the method of Carhart et al. [37]. POLLY and APPROBE are software developed at the University of Minnesota, written in ANSI C, and are available in MSDOS and UNIX versions. Structural input for both of these programs is the SMILES line notation [38].

Index selection using principal components analysis

The large number of topological indices, and the fact that many of them are highly correlated, confounds the development of predictive models. We reduced the number of topological indices to be used in the analysis to about one-third the number of compounds used in the analysis. Principal components analysis was used to identify topological indices that were important contributors to the overall variance of the data set as determined by the correlation of each index with each

of the principal components (PCs). The number of indices chosen from a given PC axis was proportional to the variance explained by that axis. Indices selected by this approach were then used in the development of predictive models by neural network and discriminant function analyses.

Measurement of similarity and K-*nearest neighbor estimation*

In this approach, intermolecular similarity of the compounds was determined with an associative coefficient using atom pairs. Similarity between chemicals *i* and *j* was defined as

$$S_{ij} = 2C/(T_i+T_j)$$

where C is the number of atom pairs common to molecules *i* and *j*. T_i and T_j are the total number of atom pairs in *i* and *j*, respectively [37].

The five nearest neighbors (i.e., $K = 5$) were used to predict the mode of action of a probe chemical. Five was chosen because of the small size of some of the mode of action groups and several previous analyses using this method have shown that predictability often degenerates with larger K [10,12,13,15]. A chemical was classified to the mode of action that was represented more than $K/2$ times (i.e., ≥ 3) within the five neighbors. For any instance where no mode of action was represented more than $K/2$ times, the chemical was classified as unknown.

Neural networks

Neural networks are a class of computer models that are particularly effective at dealing with noisy or sparse data sets. These models can be used for prediction, classification, or optimization, and are relatively independent of data types or distributions [39]. Neural networks have recently received attention in the fields of applied chemistry and QSARs [16–18]. In this study, the Learning Vector Quantization (LVQ) classification network was used [40]. The architecture for an LVQ network consists of an input layer, a hidden layer, and an output layer. The input layer contains one node for each topological index. The output layer has one node for each mode of action. It is in the hidden layer that learning and classification occur. Specific architectures used in these analyses are described in the results section below. The NeuralWare Professional II neural network package [41] was used for the development of the LVQ networks.

Discriminant analysis

Linear models, utilizing stepwise discriminant analysis, were developed in addition to the neural network and similarity models. In stepwise discriminant analysis, at each step, the variable that adds the most discriminatory power to the model is selected for inclusion. If, at any step, a variable fails to meet the criteria for inclusion in the model and contributes the least to the discriminatory power of the model, it is removed. The result is a linear discriminant function composed of a subset of variables that best reveals differences among classes. The STEPDISC procedure of SAS [42] was used to perform the stepwise discriminant analyses. The DISCRIM procedure of SAS was used to provide detailed classification and cross-validation information from the stepwise discriminant analyses.

Tiered analyses

In preliminary studies involving attempts to classify chemicals into the eight mode of action classes used in this study, a large number of the narcosis I, narcosis II, and electrophilic/proelectrophilic chemicals were misclassified. The remaining groups, oxidative phosphorylation uncouplers, AChE inhibitors, neurotoxicants, and the combined class of neurodepressants and respiratory blockers, tended to differentiate fairly well. Overall classification rates were 65%, 82%, and 68% for the atom pair similarity, LVQ network, and discriminant analysis procedures, respectively. Although results with the training set were encouraging for the LVQ network, use of this model on the test set of chemicals resulted in only 61% correct classification. Because of the difficulty in distinguishing the narcotics and electrophilic/proelectrophilic reactives, a tiered approach to analyzing these data was developed.

The tiered approach aggregated various modes of action into broader categories. In the first analysis, narcosis I, narcosis II, mixed narcosis I/II, and electrophilic/proelectrophilic reactive chemicals were grouped into one category, and classified against the other four classes. A second analysis was performed to discriminate between narcosis I, narcosis II, and electrophilic/proelectrophilic reactive compounds only.

Evaluation of classification

The accuracy of the classification procedures was assessed both overall (i.e., percent of all chemicals that were classified correctly) and with a confusion matrix. Derived from the field of remote sensing, a confusion matrix allows a finer assessment of classification accuracy; specifically, it allows one to decompose how a given chemical was misclassified [43]. The two important outputs of the classification matrix are the producer's accuracy, defined as the percent of cases in which a chemical with a particular empirically defined mode of action were actually estimated to be that mode of action, and user's accuracy, defined as the percent of cases estimated to be of a particular mode of action empirically defined to be that mode of action.

RESULTS

Index selection

Table 1 presents a summary of the principal components analysis of 151 topological indices. The eigenvalues of each PC, the proportion of variation explained by the PC, and the cumulative variation explained are given in Table 1. Only PCs with eigenvalues greater than 1.0 were retained. From the original list of 151 topological indices, 60 were retained for developing the neural networks and discriminant analysis models. The indices were selected proportionally with respect to variance explained by a PC. As an example, PC_1 explained 59.9% of the variance within the set of topological indices. Therefore, 59.9% of 60, or 36 variables were selected from PC_1. Table 1 lists the number of topological indices retained from each PC. The criteria used to select which index to keep was the correlation of each of the indices with the PCs. The indices with the highest correlation with a PC were retained. The final column of Table 1 presents the topological indices retained from each PC.

K-*Nearest neighbor estimation*

For the tier I analysis, in which the narcosis I, narcosis II, and electrophile/proelectrophile reactive groups are combined,

Table 1. Summary of principal components analysis of 151 topological indices for 220 training compounds and the 60 topological indices selected from principal components (PCs) proportionally based upon variance explained

PC	Eigenvalue	Proportion of variance explained	Cumulative variance explained	No. of variables retained	Variables retained
1	90.4	59.9	59.9	36	$AN1_3$–$AN1_5$ ANN_1–ANN_5 ANS_1 ANV_3 ANV_4 AZN_1–AZN_3 AZN_5 AZS_1 AZS_2 AZV_1–AZV_3 DN^21_4 DN^2N_3 DN^2N_4 DN^2S_1 P_0 P_1 *W* M_1 M_2 I_D^W I^D ^0X–^3X $^0X^b$
2	14.9	9.9	69.7	6	SIC_3 SIC_4 SIC_5 SIC_6 CIC_4 CIC_5
3	11.8	7.8	77.5	5	SIC_0 SIC_1 DN^2N_1 DN^2N_5 DN^21_3
4	8.0	5.3	82.9	4	J^B J^X J^Y $^6X^b_{Ch}$
5	3.7	2.5	85.3	2	$^5X^v_{Ch}$ $^5X^b_{Ch}$
6	3.5	2.3	87.6	1	ANZ_1
7	2.8	1.8	89.4	1	AZN_4
8	2.5	1.6	91.1	1	ASV_1
9	2.0	1.3	92.4	1	$^5X^v_C$
10	1.7	1.2	93.5	1	ASV_5
11	1.5	1.0	94.6	1	$^4X^v_C$
12	1.2	0.8	95.3	1	DSN_5

the *K*-nearest neighbor ($K = 5$) method resulted in 90% of the training set of chemicals and 95% of the test chemicals being classified correctly (Table 2A and B). For the tier II analysis, in which discrimination is attempted between the narcosis I, narcosis II, and electrophile/proelectrophile reactive groups, 75% of the training chemicals and 72% of the test chemicals were classified correctly (Table 2C and D).

Producer's accuracy for correctly classifying the combined group of narcotics and electrophiles/proelectrophiles was very high, being 98% correct for both the training set and test set of chemicals (Table 2A and B). Producer's accuracy for the neurodepressants/respiratory blockers was relatively high at 83% (i.e., five out of six correct) for the training set of chemicals. For the test set of chemicals, producer's accuracy dropped to 50% (or one out of two) correct. For the remaining three groups, producer's accuracy was low for the training set of chemicals: 60% for uncouplers of oxidative phosphorylation, 50% for AChE inhibitors, and 0% for neurotoxicants. Producer's accuracy improved for these three groups when the test chemicals were examined: 100% for uncouplers, 100% for AChE inhibitors, and 50% for neurotoxicants.

The user's accuracy numbers reflect that almost all misclassifications were nonnarcotic chemicals being classified as narcotic chemicals. In particular, no chemical was misclassified as an AChE inhibitor or as a respiratory blocker/neurodepressant. The only narcotic/electrophilic chemicals misclassified were the training set chemicals 3-hydroxy-2-nitropyridine (a narcosis I chemical) and 4-amino-2-nitrophenol (a narcosis II chemical), and the test chemical hexachloro-1,3-butadiene (an electrophile/proelectrophile). Each of these three chemicals was classified as an uncoupler. User's accuracy for the neurotoxicants was low. No training set chemical was classified as an neurotoxicant, and only 50%, one of two chemicals, of the test set chemicals classified as an neurotoxicant was empirically defined to be an neurotoxicant.

For the training set of chemicals, methanol rhodamine b and sodium azide were not classified. Each of these chemicals is composed of only two nonhydrogen atoms. The definition of an atom pair implies that these chemicals are structurally unique and have no neighbors. Therefore, no estimation was possible.

For the tier 2 analyses (Table 2C and D), producer's accuracy shows the difficulty similarity analysis had in classifying electrophilic/proelectrophilic chemicals. Only 9 of the 35 electrophilic chemicals (26%) in the training set were correctly classified (Table 2C) and 3 of the 9 electrophilic chemicals (33%) in the test set were correctly classified (Table 2D). In most cases, the error was classifying the electrophilic/ proelectrophilic chemical as a narcosis I chemical. Producer's accuracies for narcosis I and narcosis II chemicals were both greater than 80% in the training set of chemicals (Table 2C). These numbers dropped to 77% for narcosis I chemicals and to 67% for narcosis II chemicals when examining the test chemicals. All nine mixed narcosis I/II mode of action chemicals were correctly classified as narcotics (Table 2D).

Learning Vector Quantization neural networks

In the tier I analysis, an LVQ network was developed to discriminate among uncouplers, AChE inhibitors, neurotoxicants, neurodepressants/respiratory blockers, and a combined group containing narcosis I, narcosis II, mixed narcosis I/II (test only), and electrophile/proelectrophile reactive chemicals. The architecture was 60-5-5 (input–Kohonen–output) network. This network correctly classified 93% of the training set chemicals and 92% of the test set of chemicals (Table 3A and B). For tier II, a 60-6-3 architecture was used. Eighty-three percent of the training set of chemicals were correctly classified and 74% of the test set of chemicals were correctly classified (Table 3C and D).

Producer's accuracy for separating the combined group of narcosis I, narcosis II, and electrophile/proelectrophile reactive chemicals from other modes of action was quite high; 98% were correctly classified in both the training and test data sets (Table 3A and B). The success of classifying uncouplers was also high, 80% and 100% in the training and test data sets, respectively. The success of classifying training set AChE inhibitors, neurotoxicants, and respiratory blockers/neurodepressants ranged from 57 to 67%; these were most commonly

Table 2. Error matrices for atom pair *K*-nearest neighbor classification ($K = 5$). Modes of action: narcotics/electrophiles (NE), uncouplers, acetylcholinesterase inhibitors (AChE), neurotoxicants (NT), and respiratory blockers/neurodepressants (RB/ND)

Estimated mode of action	Observed mode of action: NE	Uncoupler	AChE	NT	RB/ND	Row total	User's accuracy (%)
A. Tier I training data set							
NE	180	4	7	7		198	91
Uncoupler	2	6				8	75
AChE			7			7	100
NT						0	0
RB/ND					5	5	100
No estimate	1				1	2	
Column total	183	10	14	7	6		
Producer's accuracy (%)	98	60	50	0	83		Overall: 90
B. Tier I test data set							
NE	53			1		54	98
Uncoupler	1	2				3	67
AChE			3			3	100
NT				1	1	2	50
RB/ND					1	1	100
Column total	54	2	3	2	2		
Producer's accuracy (%)	98	100	100	50	50		Overall: 95

Estimated mode of action	Observed mode of action: Narcosis I	Narcosis II	Electrophile/ proelectrophile	Row total	User's accuracy (%)
C. Tier II training data set					
Narcosis I	107	2	21	130	82
Narcosis II	10	21	2	33	64
Electrophile/proelectrophile	1	1	9	11	82
No estimate	5	1	3	9	
Column total	123	25	35		
Producer's accuracy (%)	87	84	26		Overall: 75

Estimated mode of action	Observed mode of action: Narcosis I	Narcosis II	Narcosis I/II	Electrophile/ proelectrophile	Row total	User's accuracy (%)
D. Tier II test data set						
Narcosis I	23	1	5	6	35	80
Narcosis II	6	4	3		13	54
Electrophile/proelectrophile	1	1		3	5	60
Narcosis I/II[a]			1		1	100
Column total	30	6	9	9		
Producer's accuracy (%)	77	67	100	33		Overall: 72

[a] Two nearest neighbors were narcosis I and two nearest neighbors were narcosis II.

misclassified into the narcotic/electrophilic category. Sixty-seven percent of the AchE inhibitors and 50% of the neurotoxicants and zero respiratory blockers were correctly classified within the test set of chemicals. For the test set, the success of the tier I analysis was also reflected in the user's accuracy: only 5% of other categories were incorrectly classified as narcotics/electrophiles.

In the tier II analysis, in which we attempt to discriminate among narcosis I, narcosis II, and electrophile/proelectrophile reactives, producer's accuracy ranged from 66 to 88% in the training data set, and 56 to 83% in the test data set. Most of the confusion arose in classifying narcosis I and electrophile/proelectrophile reactive chemicals (Table 3C and D). In the training data set, for example, 11 of 35 electrophiles (31%) were classified as narcosis I chemicals; only 1 was classified as a narcosis II chemical. The success of correctly classifying narcosis II chemicals was relatively high: 84% and 83% in the training and test data sets, respectively.

Discriminant analyses

Stepwise discriminant analysis for tier I resulted in a model that classified 93% of the training chemicals correctly (Table 4A). For the test set, 87% of the chemicals were correctly classified (Table 4B). For tier II, 80% of the training set of chemicals were correctly classified and 76% of the test set of chemicals were correctly classified (Table 4C and D).

For the combined group of narcosis I, narcosis II, and electrophile/proelectrophile reactive chemicals, producer's accuracy was quite high, 97% and 98%, for the training and test chemicals, respectively (Table 4A and B). The combined group of respiratory blockers/neurodepressants had 100% producer's accuracy for the training chemicals, but this dropped to 0% for the test chemicals. Uncouplers had a producer's accuracy of 80% for the training chemicals, but this dropped to 50% for the test chemicals. The neurotoxicants proved difficult to classify, with a classification rate of only 43% for the training

Table 3. Error matrices for Learning Vector Quantization network classification. Modes of action: narcotics/electrophiles (NE), uncouplers, acetylcholinesterase inhibitors (AChE), neurotoxicants (NT), and respiratory blockers/neurodepressants (RB/ND)

Estimated mode of action	Observed mode of action					Row total	User's accuracy (%)
	NE	Uncoupler	AChE	NT	RB/ND		
			A. Tier I training data set				
NE	179	1	4	2	2	188	95
Uncoupler	1	8				9	89
AChE	2	1	9			12	75
NT			1	4		5	80
RB/ND	1			1	4	6	67
Column total	183	10	14	7	6		
Producer's accuracy (%)	98	80	64	57	67		Overall: 93
			B. Tier I test data set				
NE	53		1			54	98
Uncoupler	1	2			1	4	50
AChE			2	1		3	67
NT				1	1	2	50
RB/ND					0	0	0
Column total	54	2	3	2	2		
Producer's accuracy (%)	98	100	67	50	0		Overall: 92

Estimated mode of action	Observed mode of action			Row total	User's accuracy (%)
	Narcosis I	Narcosis II	Electrophile/ proelectrophile		
	C. Tier II training data set				
Narcosis I	108	3	11	122	89
Narcosis II	3	21	1	25	84
Electrophile/proelectrophile	12	1	23	36	64
Column total	123	25	35		
Producer's accuracy (%)	88	84	66		Overall: 83

Estimated mode of action	Observed mode of action				Row total	User's accuracy (%)
	Narcosis I	Narcosis II	Narcosis I/II	Electrophile/ proelectrophile		
	D. Tier II test data set					
Narcosis I	23		3	4	30	77
Narcosis II	2	5	4		11	82
Electrophile/proelectrophile	5	1	2	5	13	38
Column total	30	6	9	9		
Producer's accuracy (%)	77	83	78	56		Overall: 74

set chemicals and 0% for the test set chemicals. The results for the respiratory blockers/neurodepressants were inconsistent; 100% of the training chemicals were correctly classified, but no test chemicals were correctly classified.

For the tier II analysis, the producer's accuracy was very low for the electrophilic/proelectrophilic chemicals with only 49% of the training chemicals and 33% of the test chemicals being correctly classified. For the training set, 46% of the electrophilic/proelectrophilic chemicals were classified as narcosis I chemicals and for the test set, 67% of these chemicals were classified as narcosis I chemicals.

DISCUSSION

The goal of this paper was to investigate whether nonempirical structural parameters can be used to predict the mode of toxic action of chemicals so that the QSAR of the chemical class with correct mode of action can be used in the estimation of their potential toxicity. Most of the industrial chemicals in the existing TSCA inventory and also those that are being submitted via the PMN process have no or very little physicochemical or toxicity data. Therefore, classification schemes based on experimental data would be of limited value for the prediction of modes of action of industrial chemicals. In this paper, we have used algorithmically derived parameters, namely, topological indices and atom pairs, to predict the modes of action of chemicals. Such parameters can be calculated for any molecule directly from its structure without the input of any experimental data.

Initially, we attempted to classify the 283 chemicals into eight mode of action types using three methods: similarity, neural networks (LVQ), and discriminant analysis. Neural networks are of interest because, not only are they free of the assumptions that constrain multivariate statistical analysis, they are very efficient at extracting information from data matrices with very low sample to variable ratios. Neural models are developed by ''training'' a network based on actual data. In the training process, the predictive and response data are repeatedly presented to the network, and the network coefficients are adjusted with each iteration. This learning process continues until a specified minimum error level is reached. Neural models are thus highly empirical, but are well adapted to dealing with highly variable and nonnormal data that are characteristic of toxicologic data sets.

The initial success in predicting mode of action of the eight

Table 4. Error matrices for stepwise discriminant analysis classification. Modes of action: narcotics/electrophiles (NE), uncouplers, acetylcholinesterase inhibitors (AChE), neurotoxicants (NT), and respiratory blockers/neurodepressants (RB/ND)

Estimated mode of action	Observed mode of action: NE	Uncoupler	AChE	NT	RB/ND	Row total	User's accuracy (%)
A. Tier I training data set							
NE	178	2	4	2		186	96
Uncoupler		8		1		9	89
AChE	2		10			12	83
NT				3		3	100
RB/ND	3			1	6	10	60
Column total	183	10	14	7	6		
Producer's accuracy (%)	97	80	71	43	100		Overall: 93
B. Tier I test data set							
NE	53	1	2		1	57	93
Uncoupler	1	1				2	50
AChE			1			1	100
NT					1	1	0
RB/ND				2		2	0
Column total	54	2	3	2	2		
Producer's accuracy (%)	98	50	33	0	0		Overall: 87

Estimated mode of action	Observed mode of action: Narcosis I	Narcosis II	Electrophile/ proelectrophile	Row total	User's accuracy (%)
C. Tier II training data set					
Narcosis I	112	7	16	135	83
Narcosis II	6	17	2	25	68
Electrophile/proelectrophile	5	1	17	23	74
Column total	123	25	35		
Producer's accuracy (%)	91	68	49		Overall: 80

Estimated mode of action	Observed mode of action: Narcosis I	Narcosis II	Narcosis I/II	Electrophile/ proelectrophile	Row total	User's accuracy (%)
D. Tier II test data set						
Narcosis I	25	2	5	6	38	79
Narcosis II	3	4	4		11	73
Electrophile/proelectrophile	2			3	5	60
Column total	30	6	9	9		
Producer's accuracy (%)	83	67	100	33		Overall: 76

classes using the three modeling methods was moderate, ranging from 82% to 65% of the chemicals being classified correctly. Further analysis of results showed that the main problem was created by three groups of chemicals: narcosis I, narcosis II, and electrophile/proelectrophile reactive chemicals. Many chemicals of these three groups were classified incorrectly. The most common errors reported by Russom et al. [8], who developed a rule-based expert system based upon two-dimensional substructures for the same database, were also associated with narcotics and electrophiles/proelectrophiles. One reason for this failure could be that the parameters used in this study are derived from simple and weighted graph models of molecules. Such parameters might not encode information regarding some critical stereoelectronic aspects of molecular structure that differentiate the electrophilic/proelectrophilic chemicals from narcotics. Alternatively, it is possible that a chemical of the electrophilic/proelectrophilic class itself is not responsible for the toxicity; one or more of its metabolites may be the ultimate electrophilic/proelectrophilic toxicant (see Russom et al. [8] for a more detailed discussion of these issues and associated examples).

In the first phase of the two-tier classification scheme, molecular similarity, LVQ neural networks, and discriminant analysis were used to predict five modes of action from the structure of the chemicals, namely, the combined group of narcotics and electrophiles/proelectrophiles, uncouplers, AChE inhibitors, neurotoxicants, and the group of respiratory blockers/neurodepressants (Tables 2 to 4). All three methods gave good results for training and test sets, with the success ranging from 95% for the *K*-nearest neighbor method to 87% for the discriminant analysis technique. This consistency of results obtained using topological descriptors in different classification methods indicates that the graph theoretic parameters used here contain sufficient structural information to be capable of predicting modes of action of diverse chemical species. An earlier study by Basak and Grunwald [12] showed that structure spaces derived from topological indices and atom pairs compare reasonably well with the structure space constructed from experimental physicochemical properties. Combinations of topological indices have also been used in predicting entry of chemicals through the blood–brain barrier [19] and mutagenicity of chemicals [15].

Results of the second phase of the two-tier analysis show that all three methods had moderate success in separating nar-

cosis I, narcosis II, and electrophile/proelectrophile reactive chemicals from one another (Tables 2 to 4). The principal source of confusion was the misclassification of large percentages of electrophilic/proelectrophilic chemicals as narcosis I chemicals. As noted earlier, it is possible that the structure space does not have sufficient discriminatory power to separate narcosis I, narcosis II, and electrophile/proelectrophile reactive chemicals from one another or that chemicals of the electrophilic/proelectrophilic group are structurally quite similar to narcotic molecules, whereas their metabolites are responsible for their particular mode of action.

Of the three methods, only the LVQ network could achieve a greater than 50% correct classification of the electrophile/proelectrophile reactive group during the tier II analysis. The *K*-nearest neighbor and discriminant analysis methods tended to classify the electrophile/proelectrophile chemicals as narcosis I, perhaps because the sample size of the narcosis I group is dominating the analyses. The LVQ network was not as susceptible to this tendency, relative to the other two methods.

In conclusion, the results of molecular similarity, neural network, and discriminant analysis methods show that nonempirical graph theoretic parameters used in this paper contain sufficient structural information to identify the mode of action for a set of 283 chemicals reasonably well. It is expected that once the mode of action of a new chemical entity is predicted, the QSAR for the particular mode of action can be used to estimate the toxic potential of the chemical more effectively. Such a two-tier approach would be much more accurate as compared to the first-generation QSAR models in which no attention was given to the mode of action of the chemical in selecting the particular QSAR equation to be used for hazard estimation.

Acknowledgement—Research reported in this paper was supported in part by cooperative agreement CR 819621 from the U.S. EPA, Structure–Activity Relationship Consortium of the University of Minnesota, and Exxon Biomedical Sciences. Mention of models or modeling approaches does not constitute endorsement on the part of the U.S. EPA. This is contribution 199 of the Center for Water and the Environment of the Natural Resources Research Institute.

REFERENCES

1. Arcos JC. 1987. Structure–activity relationships: Criteria for predicting carcinogenic activity of chemical compounds. *Environ Sci Technol* 21:743–745.
2. Tennant RW, Spalding J, Stasiewicz S, Ashby J. 1990. Prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 44 chemicals by National Toxicology Program. *Mutagenesis* 5:3–14.
3. Bahler D, Bristol DW. 1993. The induction of rules for predicting chemical carcinogenesis in rodents. In Hunter L, Shavlik J, Searls D, eds, *Intelligent Systems for Molecular Biology*. AAAI/MIT, Menlo Park, CA, USA pp 29–37.
4. Auer CM, Nabholz JV, Baetcke KP. 1990. 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* 87:183–197.
5. Clements RG, Johnson DW, Lipnick RL, Nabholz JV, Newsome LD. 1988. Estimating toxicity of industrial chemicals to aquatic organisms using structure–activity relationships. EPA 560-6-88-001. U.S. Environmental Protection Agency, Washington, DC.
6. Bradbury SP. 1994. Predicting modes of toxic action from chemical structure: An overview. *SAR QSAR Environ Res* 2:89–104.
7. Basak SC, Magnuson VR, Niemi GJ, Regal RR. 1988. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl Math* 19:17–44.
8. Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA. 1997. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ Toxicol Chem* 16:948–967.
9. Johnson MA, Basak SC, Maggiora G. 1988. A characterization of molecular similarity methods for property prediction. *Math Comput Model* 11:630–634.
10. Basak SC, Bertelsen S, Grunwald GD. 1994. Application of graph theoretical parameters in quantifying molecular similarity and structure–activity studies. *J Chem Inf Comput Sci* 34:270–276.
11. Lajiness MS. 1990. Molecular similarity-based methods for selecting compounds for screening. In Rouvray DH, ed, *Computational Chemical Graph Theory*. Nova, New York, NY, USA, pp 299–316.
12. Basak SC, Grunwald GD. 1998. Use of topological space and property space in selecting structural analogs. *Math Model Sci Comput* (in press).
13. Basak SC, Grunwald GD. 1995. Tolerance space and molecular similarity. *SAR QSAR Environ Res* 4:265–277.
14. Basak SC, Gute BD, Grunwald GD. 1996. Estimation of normal boiling points of haloalkanes using molecular similarity. *Croat Chem Acta* 69:1159–1173.
15. Basak SC, Grunwald GD. 1995. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: A similarity based study. *Chemosphere* 31:2529–2546.
16. Balaban AT, Basak SC, Colburn T, Grunwald GD. 1994. Correlation between structure and normal boiling points of haloalkanes C1–C4 using neural networks. *J Chem Inf Comput Sci* 34:1118–1121.
17. Brinn M, Walsh PT, Payne MP, Bott B. 1993. Neural network classification of mutagens using structural fragment data. *SAR QSAR Environ Res* 1:169–211.
18. Domine D, Devillers J, Chastrette M, Karcher W. 1993. Estimating pesticide field half-lives from a backpropagation neural network. *SAR QSAR Environ Res* 1:211–219.
19. Basak SC, Gute BD, Drewes LR. 1996. Predicting blood–brain transport of drugs: A computational approach. *Pharmacol Res* 13:775–778.
20. Gombar VK, Enslein K, Blake BW. 1995. Assessment of developmental toxicity potential of chemicals by quantitative structure–toxicity relationship models. *Chemosphere* 31:2499–2510.
21. University of Minnesota. 1988. *POLLY 2.3*. Duluth, MN, USA.
22. Randić M. 1975. On characterization of molecular branching. *J Am Chem Soc* 97:6609–6615.
23. Kier LB, Hall LH. 1986. *Molecular Connectivity in Structure–Activity Analysis*. Research Studies, Letchworth, Hertfordshire, UK.
24. Basak SC, Roy AB, Ghosh JJ. 1980. Study of the structure–function relationship of pharmacological and toxicological agents using information theory. *Proceedings*, 2nd International Conference on Mathematical Modelling. University of Missouri–Rolla, St. Louis, MO, USA, July 11–13, 1979, pp 851–856.
25. Basak SC, Magnuson VR. 1983. Molecular topology and narcosis: A quantitative structure–activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim Forsch* 33:501–503.
26. Roy AB, Basak SC, Harriss DK, Magnuson VR. 1984. Neighborhood complexities and symmetry of chemical graphs and their biological applications. In Avula XJR, Kalman RE, Lipais AI, Rodin EY, eds, *Mathematical Modelling in Science and Technology*. Pergamon, New York, NY, USA, pp 745–750.
27. Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC. 1984. Discrimination of isomeric structures using information theoretic topological indices. *J Comp Chem* 5:581–588.
28. Bonchev D, Trinajstić N. 1977. Information theory, distance matrix and molecular branching. *J Chem Phys* 67:4517–4533.
29. Wiener H. 1947. Structural determination of paraffin boiling point. *J Am Chem Soc* 69:17–20.
30. Balaban AT, Motoc I, Bonchev D, Mekenyan O. 1983. Topological indices for structure–activity correlations. In Charton M, Motoc I, eds, *Topics in Current Chemistry*, Vol 114—Steric Effects in Drug Design. Springer-Verlag, New York, NY, USA, pp 21–55.
31. Balaban AT. 1982. Highly discriminating distance-based topological index. *Chem Phys Lett* 89:399–404.
32. Balaban AT. 1983. Topological indices based on topological distances in molecular graphs. *Pure Appl Chem* 55:199–206.
33. Balaban AT. 1986. Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *MATCH* 21:115–122.
34. Balaban AT, Balaban TS. 1991. New vertex invariants and topological indices of chemical graphs based on information on distances. *J Math Chem* 8:383–397.

35. Filip PA, Balaban TS, Balaban AT. 1987. A new approach for devising local graph invariants: Derived topological indices with low degeneracy and good correlation ability. *J Math Chem* 1:61–83.
36. University of Minnesota. 1994. *APPROBE.* Duluth, MN, USA.
37. Carhart RE, Smith DH, Venkataraghavan R. 1985. Atom pairs as molecular features in structure–activity studies: Definitions and applications. *J Chem Inf Comput Sci* 25:64–73.
38. Weininger D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28:31–36.
39. Wasserman PD. 1989. *Neural Computing: Theory and Practice.* Van Nostrand Reinhold, New York, NY, USA.
40. Kohonen T, Barna G, Chrisley R. 1988. Statistical pattern recognition with neural networks: Benchmark studies. *Proceedings,* Annual IEEE International Conference on Neural Networks, ICNN-88, San Diego, CA, USA, pp I61–I68.
41. NeuralWare, Inc. 1991. *Neuralworks Professional II.* Pittsburgh, PA, USA.
42. SAS Institute. 1989. *SAS 6.08.* Cary, NC, USA.
43. Lillesand TM, Kiefer RW. 1994. *Remote Sensing and Image Interpretation.* John Wiley & Sons, New York, NY, USA.

APPENDIX 1

Summary of sample sizes used in training and test sets for each of the eight modes of action used in this study

Mode of action	Training set sample size	Test set sample size	Total sample size
Narcosis I	123	30	153
Narcosis II	25	6	31
Mixed narcosis I/II	0	9	9
Oxidative phosphorylation uncoupling	10	2	12
Acetylcholinesterase inhibition	14	3	17
Electrophilic/proelectrophilic reactivity	35	9	44
Neurotoxicity: central nervous system seizure/stimulant mechanism	7	2	9
Respiratory inhibition/ neurodepressant mechanism	6	2	8
Totals	220	63	283

APPENDIX 2

Symbols for topological indices and hydrogen bonding parameter, and their definitions

Index symbol	Definition	Reference
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph	[28]
$\bar{I}_D^W$	Mean information index for the magnitude of distance	[28]
W	Wiener index = half-sum of the off-diagonal	[29]
I^D	Degree complexity	[27]
H^V	Graph vertex complexity	[27]
H^D	Graph distance complexity	[27]

APPENDIX 2

Continued

Index symbol	Definition	Reference
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h	[28]
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph	—
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices	[24]
O_{ORB}	Maximum neighborhood order for the hydrogen-suppressed graph	—
M_1	A Zagreb group parameter = sum of square of degree over all vertices	[30]
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices	[30]
IC_r	Mean information content or complexity of a graph based on the r^{th} (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph	[24,26]
SIC_r	Structural information content for r^{th} (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph	[24,26]
CIC_r	Complementary information content for r^{th} (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph	[25,26]
^hX	Path connectivity index of order h = 0–6	[22,23]
^hX_C	Cluster connectivity index of order h = 3–6	[23]
$^hX_{Ch}$	Chain connectivity index of order h = 3–6	[23]
$^hX_{PC}$	Path-cluster connectivity index of order h = 4–6	[23]
$^hX^b$	Bonding path connectivity index of order h = 0–6	[23]
$^hX^b_C$	Bonding cluster connectivity index of order h = 3–6	[23]
$^hX^b_{Ch}$	Bonding chain connectivity index of order h = 3–6	[23]
$^hX^b_{PC}$	Bonding path-cluster connectivity index of order h = 4–6	[23]
$^hX^v$	Valence path connectivity index of order h = 0–6	[23]
$^hX^v_C$	Valence cluster connectivity index of order h = 3–6	[23]
$^hX^v_{Ch}$	Valence chain connectivity index of order h = 3–6	[23]
$^hX^v_{PC}$	Valence path-cluster connectivity index of order h = 4–6	[23]
P_H	Number of paths of length h = 0–10	—
J	Balaban's J index based on distance	[31]
J^B	Balaban's J index based on multigraph bond orders	[31,32]
J^X	Balaban's J index based on relative electronegativities	[33]
J^Y	Balaban's J index based on relative covalent radii	[33]
U	J index formula based on mean local information on the magnitude of distances	[34]
V	J index formula based on local information on the magnitude of distances	[34]
X	J index formula based on extended local information on distance magnitude	[34]
Y	J index formula based on extended mean local information on distance magnitude	[34]
LOVI	Local vertex invariants based on solutions of linear equation systems using the adjacency matrix (A), distance matrix (D), and column/row vectors: distance sums (S), atomic number (Z), number of nonhydrogen atoms (N), vertex degree (V), or numerical constants. Notation is described by triplets (e.g., AZV)	[35]