Prediction of Mutagenicity of Aromatic and Heteroaromatic Amines from Structure: A Hierarchical QSAR Approach

Subhash C. Basak,*,† Denise R. Mills,† Alexandru T. Balaban,‡ and Brian D. Gute†

Natural Resources Research Institute, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811, Organic Chemistry Department, Polytechnic University Bucharest, Splaiul Independentei 313, 77206 Bucharest, Romania, and Department of Oceanography, Texas A&M University at Galveston, Galveston, Texas 77553-1675

Received August 25, 2000

Due to the lack of experimental data, there has been increasing use of theoretical structural descriptors in the hazard assessment of chemicals. We have used a hierarchical approach to develop class-specific quantitative structure−activity relationship (QSAR) models for the prediction of mutagenicity of a set of 95 aromatic and heteroaromatic amines. The hierarchical approach begins with the simplest molecular descriptors, the topostructural, which encode limited chemical information. The complexity is then increased, adding topochemical, geometric, and finally quantum chemical parameters. We have also added log *P* to the set of independent variables. The results indicate that the topological parameters, i.e., the topostructural and topochemical indices, explain the majority of the variance, and that the inclusion of log *P*, geometric, and quantum chemical parameters does not result in significantly improved predictive models.

1. INTRODUCTION

A recent trend in chemistry, risk assessment of chemicals, drug discovery, combinatorial chemistry, and toxicology is the prediction of complex physicochemical, biological, and technological properties of chemicals from their simpler, preferably calculated, properties.[1−11] It is widely recognized that quantitative structure−property (activity) relationship (QSPR/QSAR) models, whether they are derived in a purely empirical fashion from an arbitrary set of molecular descriptors or from a preselected set of descriptors which are expected on theoretical grounds to have a connection with a particular property, can provide insight into the molecular and submolecular origin of properties.[12] From the practical viewpoint, they provide data reduction methods whereby one can predict and study the interrelatedness of numerous hitherto unrelated properties,[12−17] calculate properties which are essential for technological processes, provide necessary data in data-poor situations usually prevalent in the hazard assessment of chemicals by regulatory agencies such as the United States Environmental Protection Agency (USEPA) as well as drug design and discovery, provide quantitative meaning to the concepts of molecular similarity/dissimilarity, and allow us to carry out in silico analysis of large real or hypothetical combinatorial libraries for which the majority or all of the chemicals have no data necessary for assessing their therapeutic or toxic potential.[1]

One important goal of QSAR/QSPR studies has been to predict complex physical, chemical, biological, and technological properties of chemicals from their simpler properties, preferably calculated ones which do not require the input of experimental data. To this end, numerous experimental and computed properties have been developed and used in QSAR/QSPR studies. A specific property, whether it is experimental or calculated, assigns a real number to the chemical and orders the set of chemicals according to the numerical value of the specific property. In other words, a property provides a scale for the chemicals. For example, a particular experimentally determined solvent polarity scale orders the set of selected solvents according to the magnitude of the particular solvent polarity. Similarly, the magnitude of molecular complexity (e.g., first-order information content, IC_1) maps the set of selected chemicals into the set *R* of real numbers and orders them into a scale of molecular complexity.[8,18] If such a scale (independent variable), experimental or calculated, is linearly or nonlinearly related to the magnitude (scale) of a particular physical, chemical, biological, or technological property (dependent variable) of interest for a set of molecules, we will succeed in deriving useful QSAR/QSPR models. Linear models, such as multiple linear regression, are very popular in QSAR/QSPR studies.

Mutagenicity of chemicals is an important toxicological property both for environmental hazard assessment and drug discovery. Physicochemical, quantum chemical, and topological approaches have been used in the prediction of mutagenicity of an important class of compounds, the aromatic amines.[9,19−25] The limited success of the various QSAR studies with the set of 95 aromatic amines indicates that there is need for further studies to investigate whether one can develop QSARs for this set of compounds using algorithmically calculated parameters, i.e., parameters which can be calculated directly from molecular structure without the input of experimental data. Our group has been involved in the development of QSAR models for physicochemical and biological properties of chemicals using topostructural, topochemical, geometric, and quantum chemical param-

* To whom correspondence should be addressed. E-mail: sbasak@nrri.umn.edu. Phone: (218) 720-4230. Fax: (218) 720-4328.

† University of Minnesota Duluth.

‡ Polytechnic University Bucharest and Texas A&M University at Galveston.

10.1021/ci000126f CCC: $20.00
Published on Web 12/14/2000

eters.[4,26−30] We have used techniques such as principal components analysis (PCA) and variable clustering (VC) to derive uncorrelated parameters from different sets of calculated descriptors. While the former, PCA, gives a set of orthogonal variables which are the linear combination of the original variables, the latter, VC, provides a subset of original parameters which are mutually minimally correlated. In our hierarchical QSAR approach, we have first classified the independent variables into four different groups, viz., topostructural, topochemical, geometric, and quantum chemical. We have also investigated whether the hydrophobicity (log *P*) contributed to improving the correlation.

We begin the QSAR model development with the least complex topostructural parameters; indices of higher levels of complexity are utilized for model building in a graduated manner only if their addition leads to a significant improvement in the predictive capability of the resultant models. This process not only leads to a good model at the end but also provides insight into the relative importance of the different classes of variables in explaining the variance of the data. In our previous QSAR study, we carried out a hierarchical QSAR analysis of a set of 95 aromatic amines using a set of 111 descriptors consisting of topological, geometric, and quantum chemical indices. In this paper, we have carried out a further study of the same set of aromatic amines with an augmented set of 211 descriptors consisting of parameters from the same four classes.

2. METHODS

2.1. Database. The set of 95 aromatic and heteroaromatic amines collected by Debnath et al.[21] was used to study mutagenic potency. The mutagenic activities of these compounds in *S. typhimurium* TA98 + S9 microsomal preparation are expressed as the logarithm of *R*, where *R* is the number of revertants/nmol. The compounds used in this study and their experimentally measured mutation rates are listed in Table 1. The results of some previous QSAR analyses on this data set, along with results from the current study, are summarized in Table 2. The log *P* values were also obtained from Debnath et al.,[21] some of which are experimental and others are calculated values, as indicated in Table 1.

2.2. Calculation of Topological Indices. The topological indices used in this study were calculated using POLLY 2.3[31] and other software developed by Basak et al. These indices include the following: Wiener's number;[32] molecular connectivity indices developed by Randić[33] and Kier and Hall;[34] frequency of various path lengths; information theoretic indices based on distance matrixes using the methods of Bonchev and Trinajstić[35] as well as those of Raychaudhury et al.;[36] a set of descriptors based on the neighborhood complexity of vertexes in hydrogen-filled molecular graphs;[8,37−39] Balaban's *J* indices;[40−42] and the triplet indices.[43,44] The triplet indices result from a matrix, a main diagonal column vector, and a free term column vector, converting the matrix into a system of linear equations whose solutions are the local vertex invariants. These local vertex invariants are then used in the following operations in order to obtain the triplet descriptors: (1) summation, $\Sigma_i x_i$; (2) summation of squares, $\Sigma_i x_i^2$; (3) summation of square roots, $\Sigma_i x_i^{1/2}$; (4) sum of inverse square root of cross-product over edges ij, $\Sigma_{ij}(x_i x_j)^{-1/2}$; (5) product, $N(\Sigma_i x_i)^{1/N}$.

A total of 202 topological indices were calculated for use in the current study. A brief description of these indices is provided in Table 3.

2.3. Calculation of Geometric and Quantum Chemical Indices. Three geometric parameters were included in the study. van der Waals volume, V_W, was calculated using Sybyl 6.2.[45] Two variants of the 3-D Wiener number, ^{3D}W and $^{3D}W_H$, based on the hydrogen-suppressed and hydrogen-filled geometric distance matrixes, respectively, were also calculated by Sybyl using a SPL (Sybyl Programming Language) program developed by our group. Six quantum chemical parameters, E_{HOMO}, E_{HOMO-1}, E_{LUMO}, E_{LUMO+1}, ΔHf, and μ, were calculated for the AM1 semiempirical Hamiltonian using MOPAC 6.0 in the Sybl interface.[46]

2.4. Data Reduction and Hierarchical QSAR. Initially, the descriptors were transformed by the natural logarithm due to the fact that their scales differed by several orders of magnitude. Because a descriptor may have a value of 0 for one or more compounds, 1 was added to the descriptor values before taking the natural logarithm. Any descriptor with a value of 0 for all compounds was removed and not used in subsequent analyses. The CORR procedure[47] of the SAS software package was performed in order to identify descriptors which were perfectly correlated, i.e., having a correlation coefficient of 1.0. In this case, only one of the perfectly correlated descriptors was retained and the other(s) discarded and not used in subsequent analyses. The remaining set of topological indices was then divided into two distinct subsets: topostructural indices and topochemical indices (Table 3). Topostructural indices are topological indices which encode information strictly about the adjacency and topological distances of atoms in molecular structures, while topochemical indices encode information about both the topology and chemical properties of the molecular structures.

According to Topliss and Edwards,[48] the use of too many independent variables in the development of QSAR models may result in chance correlations. Using their findings, we have determined that a maximum of 60 independent variables may be used in a regression analysis for a set of 95 compounds, provided that the explained variance of the model is 0.7 or greater. In this case, the probability of chance correlations would be less than 0.01. Therefore, our set of 211 descriptors must be greatly reduced prior to the regression analyses. The VARCLUS procedure[47] of the SAS software package was used independently on the topostructural indices and the topochemical indices to perform variable clustering in order to reduce the number of descriptors to be used in the subsequent regression analysis. Each descriptor is assigned to only one cluster. In each case, we selected from each cluster the one descriptor most correlated with the cluster and all descriptors poorly correlated with the cluster ($R < 0.70$). The selected topological and topochemical descriptors were used with the geometric and quantum chemical descriptors in the subsequent hierarchical model development (Table 4).

The hierarchy begins with the simplest indices, the topostructural. The first model is developed solely on the basis of the topostructural indices. Increasing the complexity, we add the topochemical indices selected by variable clustering to the set of descriptors used in the topostructural model, and modeling is performed again resulting in a topostructural + topochemical model. A third model is

Table 1. Observed and Estimated Mutagenic Potency [log(Revertants/nmol)] for 95 Aromatic and Heteroaromatic Amines

no.	compd	log *P*	obsd log *R*	est log *R* (eq 2)	Δ	no.	compd	log *P*	obsd log *R*	est log *R* (eq 2)	Δ
1	2-bromo-7-aminofluorene	3.92[a]	2.62	2.25	0.37	49	2,6-dichloro-1,4-phenylenediamine	1.79	−0.69	−0.76	0.07
2	2-methoxy-5-methylaniline (*p*-cresidine)	1.74[a]	−2.05	−2.84	0.79	50	2-amino-7-acetamidofluorene	1.72	1.18	1.54	−0.36
3	5-aminoquinoline	1.16[a]	−2.00	−2.77	0.77	51	2,8-diaminophenazine	1.64	1.12	1.14	−0.02
4	4-ethoxyaniline (*p*-phenetidine)[c]	1.24[a]	−2.30	−3.97	1.67	52	6-aminoquinoline	1.28[a]	−2.67	−2.61	−0.06
5	1-aminonaphthalene	2.25[a]	−0.60	−0.86	0.26	53	4-methoxy-2-methylaniline (*m*-cresidine)	1.23[a]	−3.00	−2.73	−0.27
6	4-aminofluorene	2.70	1.13	0.94	0.19	54	3-amino-2′-nitrobiphenyl	2.68	−1.30	−0.16	−1.14
7	2-aminoanthracene	3.26	2.62	1.37	1.25	55	2,4′-diaminobiphenyl	1.58	−0.92	−0.04	−0.88
8	7-aminofluoranthene	3.72	2.88	2.69	0.19	56	1,6-diaminophenazine	1.64	0.20	0.09	0.11
9	8-aminoquinoline	1.79[a]	−1.14	−2.54	1.40	57	4-aminophenyl disulfide	1.99	−1.03	0.38	−1.41
10	1,7-diaminophenazine	1.64	0.75	1.41	−0.66	58	2-bromo-4,6-dinitroaniline	2.78	−0.54	−0.99	0.45
11	2-aminonaphthalene	2.28[a]	−0.67	−0.79	0.12	59	2,4-diamino-*n*-butylbenzene	1.77	−2.70	−2.69	−0.01
12	4-aminopyrene	3.72	3.16	3.25	−0.09	60	4-aminophenyl ether	1.36[a]	−1.14	−0.51	−0.63
13	3-amino-3′-nitrobiphenyl	2.68	−0.55	0.06	−0.61	61	2-aminobiphenyl	2.84[a]	−1.49	−0.78	−0.71
14	2,4,5-trimethylaniline	2.41	−1.32	−0.54	−0.78	62	1,9-diaminophenazine	1.64	0.04	0.20	−0.16
15	3-aminofluorene	2.70	0.89	1.20	−0.31	63	1-aminofluorene	3.18[a]	0.43	0.85	−0.42
16	3,3′-dichlorobenzidine	3.51[a]	0.81	0.05	0.76	64	8-aminofluoranthene	3.72	3.80	2.89	0.91
17	2,4-dimethylaniline (2,4-xylidine)	1.68[a]	−2.22	−1.81	−0.41	65	2-chloroaniline	1.90[a]	−3.00	−2.47	−0.53
						66	2-amino-ααα-trifluorotoluene	2.29[a]	−0.80	−1.85	1.05
18	2,7-diaminofluorene	1.47	0.48	1.15	−0.67	67	2-amino-1-nitronaphthalene	2.95	−1.17	−0.43	−0.74
19	3-aminofluoranthene	4.20[a]	3.31	2.78	0.53	68	3-amino-4-nitrobiphenyl	2.68	0.69	−0.05	0.74
20	2-aminofluorene	3.14[a]	1.93	1.20	0.73	69	4-bromoaniline	2.26[a]	−2.70	−1.98	−0.72
21	2-amino-4′-nitrobiphenyl	2.68	−0.62	−0.39	−0.23	70	2-amino-4-chlorophenol	1.81[a]	−3.00	−2.34	−0.66
22	4-aminobiphenyl	2.86[a]	−0.14	−0.83	0.69	71	3,3′-dimethoxybenzidine	1.81	0.15	−0.63	0.78
23	3-methoxy-4-methylaniline (*o*-cresidine)	1.52	−1.96	−2.80	0.84	72	4-cyclohexylaniline	3.65[a]	−1.24	−0.80	−0.44
						73	4-phenoxyaniline	2.96[a]	0.38	−0.76	1.14
24	2-aminocarbazole	2.30	0.60	0.60	0.00	74	4,4′-methylenebis (*o*-ethylaniline)	3.66	−0.99	−0.43	−0.56
25	2-amino-5-nitrophenol	1.36	−2.52	−2.62	0.10						
26	2,2′-diaminobiphenyl	1.58	−1.52	−0.95	−0.57	75	2-amino-7-nitrofluorene	3.06[a]	3.00	1.61	1.39
27	2-hydroxy-7-aminofluorene	2.03	0.41	1.37	−0.96	76	benzidine	1.34[a]	−0.39	−0.72	0.33
28	1-aminophenanthrene	3.26	2.38	1.41	0.97	77	1-amino-4-nitronaphthalene	2.48	−1.77	−0.62	−1.15
29	2,5-dimethylaniline (2,5-xylidine)	1.83[a]	−2.40	−1.57	−0.83	78	4-amino-3′-nitrobiphenyl	2.68	1.02	0.04	0.98
						79	4-amino-4′-nitrobiphenyl	2.68	1.04	0.11	0.93
30	4-amino-2′-nitrobiphenyl	2.68	−0.92	−0.25	−0.67	80	1-aminophenazine	2.18	−0.01	0.85	−0.86
31	2-amino-4-methylphenol	1.16[a]	−2.10	−2.78	0.68	81	4,4′-methylenebis (*o*-fluoroaniline)	2.50	0.23	−0.50	0.73
32	2-aminophenazine	2.18	0.55	0.96	−0.41						
33	4, 4′-diaminophenyl sulfide	2.18[a]	0.31	−0.29	0.60	82	4-chloro-2-nitroaniline	2.72[a]	−2.22	−1.98	−0.24
34	2,4-dinitroaniline	1.84	−2.00	−1.52	−0.48	83	3-aminoquinoline	1.63[a]	−3.14	−2.38	−0.76
35	2,4-diaminoisopropylbenzene	1.12	−3.00	−2.01	−0.99	84	3-aminocarbazole	2.30	−0.48	0.62	−1.10
36	2,4-difluoroaniline	1.54[a]	−2.70	−2.46	−0.24	85	4-chloro-1,2-phenylenediamine	1.28[a]	−0.49	−1.05	0.56
37	4,4′-methylenedianiline	1.59[a]	−1.60	−0.70	−0.90	86	3-aminophenanthrene[b]	3.26	3.77	1.63	2.14
38	3,3′-dimethylbenzidine	2.34[a]	0.01	−0.66	0.67	87	3,4′-diaminobiphenyl	1.58	0.20	0.42	−0.22
39	2-aminofluoranthene	3.72	3.23	3.04	0.19	88	1-aminoanthracene	3.69[a]	1.18	1.24	−0.06
40	2-amino-3′-nitrobiphenyl	2.68	−0.89	−0.21	−0.68	89	1-aminocarbazole	2.30	−1.04	0.42	−1.46
41	1-aminofluoranthene	3.72	3.35	2.78	0.57	90	9-aminoanthracene	3.26	0.87	0.44	0.43
42	4,4′-ethylenebis (aniline)	2.13	−2.15	−1.39	−0.76	91	4-aminocarbazole[c]	2.30	−1.42	0.37	−1.79
43	4-chloroaniline	1.88[a]	−2.52	−2.70	0.18	92	6-aminochrysene	4.98	1.83	3.22	−1.39
44	2-aminophenanthrene	3.26	2.46	1.60	0.86	93	1-aminopyrene[b]	4.31[a]	1.43	3.44	−2.01
45	4-fluoroaniline	1.15[a]	−3.32	−3.28	−0.04	94	4,4′-methylenebis (*o*-isopropylaniline)	4.46	−1.77	−0.95	−0.82
46	9-aminophenanthrene[c]	3.56[a]	2.98	1.25	1.73						
47	3,3′-diaminobiphenyl	1.58	−1.30	−0.79	−0.51	95	2,7-diaminophenazine[b]	1.64	3.97	1.03	2.94
48	2-aminopyrene	3.72	3.50	2.95	0.55						

[a] Experimental log *P*. [b] Compounds not included when deriving eqs 5 and 6. [c] Compounds not included when deriving eq 6.

Table 2. Predictive Models (Ordered Chronologically) for Mutagenic Potency in *S. typhimurium* TA98 + S9 Based on Aromatic and Heteroaromatic Amines

no. of params	params	*n*	*r*	*s*	investigator(s)	ref
4	log *P*, ϵ_{HOMO}, ϵ_{LUMO}, I_L	88	0.898	0.860	Debnath et al. (1992)	21
9	topological and geometric	95	0.893	0.91	Basak et al. (1997)	9
9	topological, geometric, and quantum chem	95	0.889	0.92	Basak et. al. (1998)	22
6	no. of rings, γ-polarizability, HASA1(SCF/AM1), HDSA (SCF/AM1), E_{tot} (C−C), E_{tot} (C−N)	95	0.913	0.811	Maran et al. (1999)	23
9	electrotopological state indices	95	0.876	0.979	Cash (2000)	24
8	expanded set of topological, geometric, and quantum chem	95	0.891	0.912	Basak et al.	current study
		92	0.910	0.801		
		89	0.923	0.742		

created by adding the geometric descriptors to those descriptors in the topostructural + topochemical model. Likewise, the indices included in the best model from this procedure are added to the quantum chemical indices and modeling is

Table 3. Symbols, Definitions, and Classification of Topological Parameters

	Topostructural
I_D^W	information index for the magnitudes of dists between all possible pairs of vertexes of a graph
$\bar{I}_D^W$	mean information index for the magnitude of dist
W	Wiener index = half-sum of the off-diagonal elements of the dist matrix of a graph
I^D	degree complexity
H^V	graph vertex complexity
H^D	graph dist complexity
$\overline{IC}$	information content of the dist matrix partitioned by frequency of occurrences of dist h
M_1	a Zagreb group param = sum of square of degree over all vertexes
M_2	a Zagreb group param = 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 $h = 3-6$
$h\chi_{PC}$	path−cluster connectivity index of order $h = 4-6$
$h\chi_{Ch}$	chain connectivity index of order $h = 3-6$
P_h	no. of paths of length $h = 0-10$
J	Balaban's j index based on topological dist
DN^2S_y	triplet index from dist matrix, square of graph order, and dist sum; operation $y = 1-5$
DN^21_y	triplet index from dist matrix, square of graph order, and no. 1; operation $y = 1-5$
$AS1_y$	triplet index from adjacency matrix, dist sum, and no. 1; operation $y = 1-5$
$DS1_y$	triplet index from dist matrix, dist sum, and no. 1; operation $y = 1-5$
ASN_y	triplet index from adjacency matrix, dist sum, and graph order; operation $y = 1-5$
DSN_y	triplet index from dist matrix, dist sum, and graph order; operation $y = 1-5$
DN^2N_y	triplet index from dist matrix, square of graph order, and graph order; operation $y = 1-5$
ANS_y	triplet index from adjacency matrix, graph order, and dist sum; operation $y = 1-5$
$AN1_y$	triplet index from adjacency matrix, graph order, and no. 1; operation $y = 1-5$
ANN_y	triplet index from adjacency matrix, graph order, and graph order again; operation $y = 1-5$
ASV_y	triplet index from adjacency matrix, dist sum, and vertex degree; operation $y = 1-5$
DSV_y	triplet index from dist matrix, dist sum, and vertex degree; operation $y = 1-5$
ANV_y	triplet index from adjacency matrix, graph order, and vertex degree; operation $y = 1-5$
	Topochemical
O	order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
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 rth ($r = 0-6$) 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 $h = 3-6$
$h\chi^b_{Ch}$	bond chain connectivity index of order $h = 3-6$
$h\chi^b_{PC}$	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 $h = 3-6$
$h\chi^v_{Ch}$	valence chain connectivity index of order $h = 3-6$
$h\chi^v_{PC}$	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
AZV_y	triplet index from adjacency matrix, atomic no., and vertex degree; operation $y = 1-5$
AZS_y	triplet index from adjacency matrix, atomic no., and dist sum; operation $y = 1-5$
ASZ_y	triplet index from adjacency matrix, dist sum, and atomic no.; operation $y = 1-5$
AZN_y	triplet index from adjacency matrix, atomic no., and graph order; operation $y = 1-5$
ANZ_y	triplet index from adjacency matrix, graph order, and atomic no.; operation $y = 1-5$
DSZ_y	triplet index from dist matrix, dist sum, and atomic no.; operation $y = 1-5$
DN^2Z_y	triplet index from dist matrix, square of graph order, and atomic no.; operation $y = 1-5$

Table 4. Parameters Used for Modeling Mutagenic Potency (log R)

topostructural	topochemical	geometric	quantum chem
I^W_D	DSZ_2	V_W	E_{HOMO}
$^4\chi$	ASZ_3, ASZ_4	^{3D}W	E_{HOMO-1}
DN^2S_2	ANZ_5	$^{3D}W_H$	E_{LUMO}
DNS_1	SIC_0, SIC_2, SIC_4		E_{LUMO+1}
$^4\chi_C$	IC_2, IC_3, IC_6		ΔHf
$^4\chi_{PC}$	$^3\chi^b_C$, $^5\chi^b_C$		μ
$^5\chi_{Ch}$, $^6\chi_{Ch}$	$^4\chi^b_{PC}$		
	$^6\chi^b_{Ch}$		
	$^4\chi^v$		
	J^B		
	O		
	O_{orb}		

performed once again, resulting in a model which is based on topostructural, topochemical, geometric, and quantum chemical descriptors−all levels of the hierarchy. The selected descriptors used in modeling, on the basis of the hierarchical classification, are listed in Table 4.

All models were developed using all possible subsets regression via the REG procedure[47] of the SAS statistical package, and final models were selected on the basis of RSQUARE and Mallow's Cp (CP).

3. RESULTS

All-subset regression using the topostructural indices resulted in a four-parameter model for the estimation of mutagenic potency:

$$\log R = -0.749 I_D^W + 4.38\ ^4\chi_{PC} - 51.2\ ^5\chi_{Ch} + 27.9\ ^6\chi_{Ch} - 3.58 \quad (1)$$

$$n = 95 \quad r = 0.842 \quad s = 1.06 \quad F = 54.9$$

Although RSQUARE does increase with the inclusion of additional topostructural parameters, it does so minimally, and CP is decreased in models containing more than four topostructural parameters.

The topochemical indices were added to the indices used in the above topostructural model, and all-subset regression was performed again, resulting in an eight-parameter model retaining only one topostructural index and acquiring seven topochemical indices:

$$\log R = 8.82\,{}^4\chi_{PC} + 6.51IC_6 - 17.4SIC_2 - 9.57J^B + 13.4ASZ_3 - 2.92ASZ_4 + 2.64\,{}^3\chi^b{}_C - 20.1\,{}^5\chi^b{}_C - 10.0 \quad (2)$$

$$n = 95 \quad r = 0.891 \quad s = 0.912 \quad F = 41.6$$

Adding the three geometric indices to those given in eq 2 above and repeating the regression analysis, we find that the resulting nine-parameter model is not significantly improved. Here, all indices from eq 2 have been retained, and one geometric index has been added:

$$\log R = 8.51\,{}^4\chi_{PC} + 6.32IC_6 - 18.2SIC_2 - 10.3J^B + 12.2ASZ_3 - 2.13ASZ_4 + 2.63\,{}^3\chi^b{}_C - 17.6\,{}^5\chi^b{}_C - 2.76V_W + 4.25 \quad (3)$$

$$n = 95 \quad r = 0.894 \quad s = 0.905 \quad F = 37.8$$

It should be noted, however, that the descriptors ASZ_4 and V_W are highly intercorrelated as can be seen in Table 5.

The addition of six quantum chemical indices to the regression analysis results in a nine-parameter model which again shows only minimal improvement. Here, the geometric and one topochemical descriptor have been replaced by two quantum chemical parameters:

$$\log R = 9.83\,{}^4\chi_{PC} + 5.96IC_6 - 19.2SIC_2 - 12.3J^B + 13.3ASZ_3 - 3.24ASZ_4 - 19.3\,{}^5\chi^b{}_C - 0.387E_{LUMO} - 0.00833\Delta Hf - 3.01 \quad (4)$$

$$n = 95 \quad r = 0.896 \quad s = 0.898 \quad F = 38.5$$

The topostructural and topochemical descriptors alone, which are quickly and easily calculated from molecular structure, explain most of the variance, with the geometric and quantum chemical descriptors improving the model only minimally. Table 1 includes the predicted mutagenic potency based on the model given in eq 2, and a scatter plot of the observed versus the estimated values using this model is presented in Figure 1. The residuals for 3-aminoanthracene, 1-aminopyrene, and 2,7-diaminophenazine (nos. 86, 93, and 95 in Table 1) were greater than twice the standard error of eq 2. Considering these to be outliers, the statistical measures of the topostructural + topochemical model given in eq 2 improved significantly:

$$\log R = 8.31\,{}^4\chi_{PC} + 6.72IC_6 - 17.2SIC_2 - 10.6J^B + 13.8ASZ_3 - 2.97ASZ_4 + 1.79\,{}^3\chi^b{}_C - 16.0\,{}^5\chi^b{}_C - 9.22 \quad (5)$$

$$n = 92 \quad r = 0.910 \quad s = 0.801 \quad F = 50.3$$

If the residuals from the model containing 92 compounds are examined and twice the value of s for that model is considered, three additional compounds must be eliminated, namely, 4, 46, and 91, leaving 89 aromatic mutagenic amines (eq 6). Scatter plots of the observed versus the estimated values using the models given in eqs 5 and 6 are presented in Figure 1. These successive elimination steps with the corresponding progressive improvements of statistical indicators may be seen in the lower portion of Table 2.

$$\log R = 9.07\,{}^4\chi_{PC} + 6.63IC_6 - 17.2SIC_2 - 11.0J^B + 13.8ASZ_3 - 3.15ASZ_4 + 1.87\,{}^3\chi^b{}_C - 17.3\,{}^5\chi^b{}_C - 8.59 \quad (6)$$

$$n = 89 \quad r = 0.923 \quad s = 0.742 \quad F = 57.2$$

In the linear free energy related (LFER) approach to predicting the mutagenic potency of the same set of aromatic amines, Debnath et al.[21] used hydrophobicity (log P, octanol/water), E_{HOMO}, E_{LUMO}, and an indicator variable. We were, therefore, interested to see if adding log P to the topostructural + topochemical model would result in a model superior

Table 5. Intercorrelation Matrix for Descriptors Used in Eqs 2−7[a]

${}^I{}_D{}^W$	ASZ_4	V_W	${}^4\chi_{PC}$	J^B
	0.99578	0.94960	0.77297	−0.76516
	0.0001	0.0001	0.0001	0.0001
${}^4\chi_{PC}$	${}^I{}_D{}^W$	ASZ_3	ASZ_4	${}^5\chi^b{}_C$
	0.77297	0.76162	0.75853	0.70405
	0.0001	0.0001	0.0001	0.0001
${}^5\chi^b{}_{Ch}$	${}^6\chi^b{}_{Ch}$	IC_6	ASZ_3	J^B
	0.84333	0.43690	0.38200	−0.33546
	0.0001	0.0001	0.0009	
${}^6\chi^b{}_{Ch}$	${}^5\chi^b{}_{Ch}$	J^B	ΔHf	ASZ_4
	0.84333	−0.62833	0.61315	0.59764
	0.0001	0.0001	0.0001	0.0001
IC_6	${}^6\chi^b{}_{Ch}$	log P	ASZ_3	${}^4\chi_{PC}$
	0.58929	0.54834	0.46559	0.44770
	0.0001	0.0001	0.0001	0.0001
SIC_2	V_W	ASZ_4	J^B	${}^I{}_D{}^W$
	−0.61000	−0.58334	0.58085	−0.54817
	0.0001	0.0001	0.0001	0.0001
J^B	V_W	ASZ_4	${}^I{}_D{}^W$	${}^6\chi^b{}_{Ch}$
	−0.79787	−0.79587	−0.76516	−0.62833
	0.0001	0.0001	0.0001	0.0001
ASZ_3	${}^4\chi_{PC}$	E_{LUMO}	log P	${}^6\chi^b{}_{Ch}$
	0.76162	−0.64771	0.59018	0.54136
	0.0001	0.0001	0.0001	0.0001
ASZ_4	${}^I{}_D{}^W$	V_W	J^B	${}^4\chi_{PC}$
	0.99578	0.95135	−0.79587	0.75853
	0.0001	0.0001	0.0001	0.0001
${}^3\chi^b{}_C$	${}^5\chi^b{}_C$	${}^4\chi_{PC}$	${}^I{}_D{}^W$	V_W
	0.65013	0.61003	0.46256	0.46238
	0.0001	0.0001	0.0001	0.0001
${}^5\chi^b{}_C$	${}^4\chi_{PC}$	${}^3\chi^b{}_C$	ASZ_3	log P
	0.70405	0.65013	0.47023	0.35327
	0.0001	0.0001	0.0001	0.0004
V_W	ASZ_4	${}^I{}_D{}^W$	J^B	${}^4\chi_{PC}$
	0.95135	0.94960	−0.79787	0.69970
	0.0001	0.0001	0.0001	0.0001
E_{LUMO}	ASZ_3	ΔHf	${}^4\chi_{PC}$	${}^I{}_D{}^W$
	−0.64771	−0.53908	−0.51438	−0.43515
	0.0001	0.0001	0.0001	0.0001
ΔHf	${}^6\chi^b{}_{Ch}$	ASZ_4	E_{LUMO}	ASZ_3
	0.61315	0.54908	−0.53908	0.52605
	0.0001	0.0001	0.0001	0.0001
log P	ASZ_3	${}^4\chi_{PC}$	ASZ_4	IC_6
	0.59018	0.56557	0.55240	0.54834
	0.0001	0.0001	0.0001	0.0001

[a] The four descriptors most highly correlated with each descriptor are listed. V_W and ASZ_4 are the only descriptors used within a given model which are highly intercorrelated (>0.85).

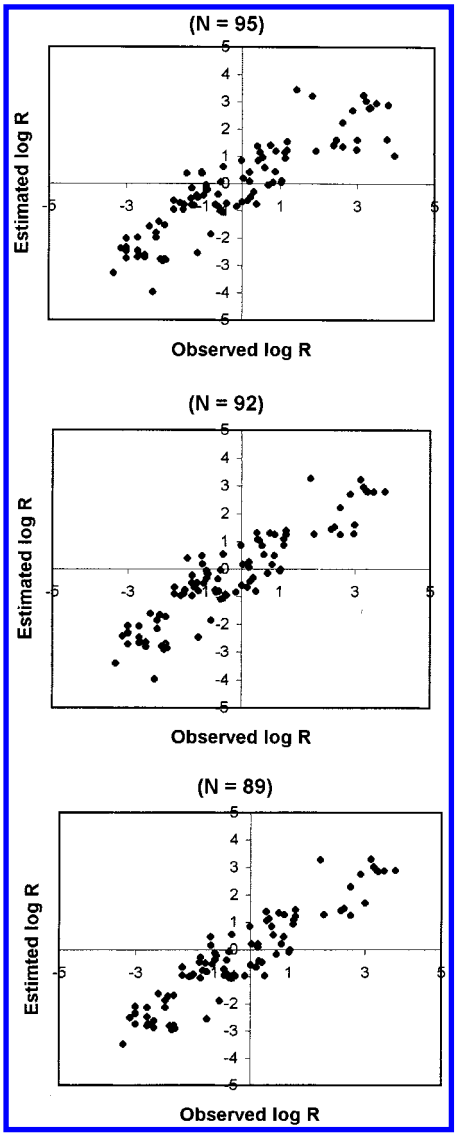

Figure 1. Scatter plot of observed mutagenic potency versus estimated mutagenic potency using the models in eqs 2, 5, and 6 for the set aromatic and heteroaromatic amines ($N = 95$, 92, and 89, respectively).

to that obtained by adding the geometric and quantum chemical descriptors, and we found that this was not the case and, in fact, there was an increase in the standard error:

$$\log R = 9.12\,{}^{4}\chi_{PC} + 6.04\mathrm{IC}_6 - 16.4\mathrm{SIC}_2 - 9.40J^{B} + 12.1\mathrm{ASZ}_3 - 2.95\mathrm{ASZ}_4 + 2.66\,{}^{3}\chi^{b}_{C} - 21.4\,{}^{5}\chi^{b}_{C} + 0.139\log P - 8.01 \quad (7)$$

$$n = 95 \quad r = 0.892 \quad s = 0.914 \quad F = 36.9$$

It is also interesting to note that a model containing log P alone (eq 8) is inferior to a model containing only one topological descriptor (eq 9). Equation 7 shows that the

$$\log R = 1.37\log P - 3.51 \quad (8)$$

$$n = 95 \quad r = 0.632 \quad s = 1.50 \quad F = 62.0$$

$$\log R = 4.85\,{}^{4}\chi - 7.83 \quad (9)$$

$$n = 95 \quad r = 0.787 \quad s = 1.19 \quad F = 152$$

inclusion of log P did not give an improved model although it had a fair correlation with log R (eq 8). log P is an important determinant of toxicokinetics and pharmacokinetics in physiological systems and the cellular milieu, but for many compounds, log P might not be available because of missing fragment values. The fact that log P does not make any improvement in model quality over and above the calculated molecular descriptors probably indicates that calculated variables quantify aspects of molecular structure characterized by log P.

The intercorrelation between all descriptors used in the multiparameter models is given in Table 5. Aside from the high correlation between two descriptors in eq 3, none of the models developed contain highly intercorrelated parameters.

4. DISCUSSION

It was argued that the chemical space of small molecules able to be synthesized by chemists contains from 10^{100} to 10^{200} structures.[49,50] Drug-like molecules belong to this space. These numbers are, however, so huge that they must be scaled down considerably. Modern medicinal chemistry and environmental sciences are based on QSAR and QSPR studies that allow the prediction of biological activities or of physical−chemical properties of $10^{5-}10^{8}$ unknown compounds using data obtained experimentally from small data sets of $10^{1-}10^{2}$ (seldom 10^{3}) structures. It is essential to stress that drug design, toxicity and genotoxicity studies, biodegradability, and hazard assessment need computational methods with low CPU requirements for developing large virtual libraries. On using various filters such libraries can be scaled down to manageable sizes, amenable for synthetic combinatorial chemistry (CC) and high-throughput screening (HTS). At present, the sequence of events for new medicinal drugs has usually become the following one: in silico screening → CC + HTS → in vitro screening → in vivo screening → clinical testing. Thus, the cost and time needed for developing new drugs is no longer escalating.

Molecular formulas may be modeled by topostructural (TS), topochemical (TC), steric and geometrical parameters (3D), electronic factors computed quantum-chemically (QC), hydrophobic (log P), and hydrogen-bonding factors. The electronic charges, hydrophobic/hydrophilic side chains or pockets, and the hydrogen-bonding donor/acceptor sites can be computed globally or locally, but the local alternative is computer-time-intensive and is chosen at the end of the previous in silico screening or when pharmacophore design is detailed enough.

The most important result of the present study is the justification of the hierarchical approach that starts with the simple topostructural (connectivity) information and then continues with the more detailed topochemical data taking into account the chemical nature of atoms and the bond multiplicity. A considerable improvement of the statistical indicators was associated with this second step, and this improvement was not due to the increased number of descriptors because within the class of topostructural parameters no such improvement could be obtained by a similar increase. From the initial topostructural set of four descriptors, only one was maintained after the second step, and seven topochemical parameters joined it, leading to eq 2.

On the other hand, in the following steps, on addition of extra descriptors from the next hierarchical levels (informa-

tion on 3D, QC, or log *P*), insignificant improvements in the statistics were obtained by eqs 3, 4, and 7. One should emphasize that, in each of these steps, all descriptors used in earlier models were still competing, but from the higher hierarchical levels the descriptors were not better than those from eq 2; therefore none or at most one of the TC and TS descriptors of eq 2 could be replaced by descriptors from these levels, and improvements in the statistical measures were negligible.

In the authors' opinion, there is little incentive for going from step 2 to higher hierarchical levels: the standard error (*s*) and the *r* value have no or little improvements. Most of the variance, namely 80−85% (eqs 2, 5, and 6), is explained by TS and especially TC descriptors.

CONCLUSION

The primary objective of this paper was to investigate how far a set of 211 theoretical molecular descriptors can be useful in the formulation of an acceptable QSAR model for predicting mutagenicity of a set of 95 aromatic amines. Another important objective was to examine the relative roles of different levels of molecular descriptors, viz., topostructural, topochemical, geometrical, and quantum chemical, in the development of a useful QSAR model.

Results presented in eqs 1−6 show that the set of theoretical parameters employed in QSAR model building in this paper predict the mutagenic potency of the group of aromatic amines reasonably well. It is interesting to note that of the four classes of descriptors, the first two classes of easily calculable indices, viz., topostructural and topochemical indices, explain most of the variance in the dependent variable. The addition of geometrical and quantum chemical indices to the set of independent variables did not augment the quality of the model significantly. This is in line with earlier studies of Basak et al.,[22,26,28,29] who found in their hierarchical QSAR studies with different physicochemical and biological properties that the addition of geometrical and quantum chemical parameters to the variables selected from topostructural and topochemical indices does not augment the quality of derived QSARs. It is interesting to note that Engelhardt, Clelland, and Jurs[51] found that addition of geometric or quantum chemical information did not do much to improve their QSAR model for vapor pressure. This has substantial implication for the hazard assessment of chemicals as well as for combinatorial chemistry. In both of these cases, one has to have a large number of properties for large and diverse databases of chemicals. The use of geometric and quantum chemical indices in such situations would be extremely costly as compared to the topostructural and topochemical parameters which require much less CPU time. It was argued that the "virtual chemistry space" that a trained chemist could reasonably hope to synthesize (small organic structures) contains from 10^{100} to 10^{200} molecules; this number is so huge that much smaller virtual libraries (but still including billions of structures) are to be screened.

Cash[24] developed a QSAR model for the same set of 95 amines using one class of topochemical parameters, viz., electrotopological indices (Table 2). The set of 211 indices used in this paper did not include the electrotopological indices. It will be interesting to investigate whether the addition of these indices to the set of 211 parameters make any improvement in model quality. Such investigation is in progress which will be communicated in a subsequent paper.

ACKNOWLEDGMENT

This is Contribution No. 276 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this chapter was supported in part by Grant F49620-98-1-0015 from the United States Air Force. The authors are thankful to Greg Grunwald for technical support.

REFERENCES AND NOTES

(1) Grassy, G.; Calas, B.; Yasri, A.; Lahana, R.; Woo, J.; Iyer, S.; Kaczorek, M.; Floc'h, R.; Buelow, R. Computer-Assisted Rational Design of Immunosuppressive Compounds. *Nat. Biotechnol.* **1998**, *16*, 748−752.

(2) Basak, S. C. A Nonempirical Approach to Predicting Molecular Properties Using Graph-Theoretic Invariants. In *Practical Applications of Quantitative Structure−Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; W. Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht/Boston/London, 1990; pp 83−103.

(3) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation Using Graph Invariants. *SAR QSAR Environ. Res.* **1994**, *2*, 289−307.

(4) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Estimation of Molecular Properties. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 366−372.

(5) 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.

(6) Basak, S. C.; Gute, B. D. Use of Graph-Theoretic Parameters in Predicting Inhibition of Microsomal *p*-Hydroxylation of Aniline by Alcohols: A Molecular Similarity Approach. In *Hazardous Waste: Impacts on Human and Ecological Health*; Johnson, B. L., Xintaras, C., Andrews, J. S., Jr., Eds.; Princeton Scientific Publishing Co., Inc.: Princeton, NJ, 1997; pp 492−504.

(7) Basak, S. C.; Gute, B. D. Use of Graph Invariants in QMSA and Predictive Toxicology. In *Discrete Mathematical Chemistry*; DIMACS Series 51; Hansen, P., Fowler, P., Zheng, M., Eds.; American Mathematical Society: Providence, RI, 2000; pp 9−24.

(8) Basak S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605−609.

(9) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of Graph Theoretical and Geometrical Molecular Descriptors in Structure−Activity Relationships. In *From Chemical Topology to Three-dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997; Chapter 4, pp 73−116.

(10) Devillers, J., Balaban, A. T., Eds. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach Science Publishers: Amsterdam, 1999.

(11) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1−8.

(12) Katritzky, A. R.; Petrukhin, R.; Tatham, D.; Basak, S. C.; Benfenati, E.; Karelson, M.; Maran, U. The Interpretation of Quantitative Structure−Property and −Activity Relationships. *J. Chem. Inf. Comput. Sci*. **2000**, in preparation.

(13) Katritzky, A. R.; Tamm, T.; Wang, Y.; Sild, S.; Karelson, M. QSPR Treatment of Solvent Scales. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 684−691.

(14) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. Topological Indices: Their Nature, Mutual Relatedness, and Applications. *Math. Model.* **1987**, *8*, 300−305.

(15) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete Appl. Math. Special volume: Applications of Graph Theory in Chemistry and Physics*; Kennedy, J. W., Quintas, L. V., Eds.; Elsevier Science Publishers BV, North-Holland: New York, 1988; Vol. 19, pp 17−44.

(16) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243−272.

(17) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Estimation of Molecular Properties. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 366−372.

(18) Johnson, M.; Basak, S. C.; Maggiora, G. A Characterization of Molecular Similarity Methods for Property Prediction. *Math. Comput. Model.* **1988**, *11*, 630−634.
(19) Benigni, R.; Passerini, L.; Gallo, G.; Giorgi, F.; Cotta-Ramusino, M. QSAR Models for Discriminating Between Mutagenic and Nonmutagenic Aromatic and Heteroaromatic Amines. *Environ. Mol. Mutagen.* **1998**, *32*, 75−83.
(20) Basak, S. C.; Grunwald, G. D. Tolerance Space and Molecular Similarity. *SAR QSAR Environ. Res.* **1995**, *3*, 265−277.
(21) Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Muagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.* **1992**, *19*, 37−52.
(22) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Relative Effectiveness of Topological, Geometrical, and Quantum Chemical Parameters in Estimating Mutagenicity of Chemicals. In *Quantitative Structure−Activity Relationships in Environmental Sciences VII*; Chen, F., Schuurmann, G., Eds; SETAC Press: Pensacola, FL, 1998; Chapter 17, pp 245−261.
(23) Maran, U.; Karelson, M.; Katritzky, A. R. A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic and Aromatic Amines. *Quant. Struct.-Act. Relat.* **1999**, *18*, 3−10.
(24) Cash, G. G. Prediction of the Genotoxicity of Aromatic and Heteroaromatic Amines Using Electrotopological State Indices. *Mutat. Res.* **2000**, submitted for publication.
(25) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Assessment of the Mutagenicity of Aromatic Amines from Theoretical Structural Parameters: A Hierarchical Approach. *SAR QSAR Environ. Res.* **1999**, *10*, 117−129.
(26) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Hierarchical Approach to the Development of QSAR Models Using Topological, Geometrical and Quantum Chemical Parameters. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 675−696.
(27) 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.
(28) 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.
(29) 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.
(30) Basak, S. C.; Gute, B. D. Characterization of Molecular Structures Using Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 1−21.
(31) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY 2.3*; University of Minnesota: Duluth, MN, 1988.
(32) Wiener, N. *Cybernetics*; Wiley: New York, 1948.
(33) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(34) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, Hertfordshire, England, 1986.
(35) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.
(36) 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.
(37) Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis−A Quantitative Structure−Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneim. Forsch./Drug Res.* **1983**, *33*, 501−503.
(38) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs, and Their Biological Applications. In *Mathematical Modeling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York, 1984; pp 745−750.
(39) Basak, S. C.; Roy, A. B.; Gosh, J. J. Study of the Structure−Function Relationship of Pharmacological and Toxicological Agents Using Information Theory. In *Proceedings of the Second International Conference on Mathematical Modeling*; Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri−Rolla: Rolla, MO, 1980; Vol.2.
(40) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(41) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206.
(42) 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.
(43) Filip, P. A.; Balaban, T. S.; Balaban, A. T.; A New Approach for Devising Local Graph Invariants: Derived Topological Indices with Low Degeneracy and Good Correlation Ability. *J. Math. Chem.* **1987**, *1*, 61−83.
(44) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological Indices: Their Nature and Mutual Relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891−898.
(45) Tripos Associates, Inc., St. Louis, MO, 1994.
(46) Stewart J. J. P. *MOPAC Version 6.00. QCPE No. 455*; Frank J Seiler Research Laboratory, U.S. Air Force Academy: Boulder, CO, 1990.
(47) SAS/STAT *User's Guide*, release 6.03 ed.; SAS Institute: Cary, NC, 1988.
(48) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(49) Weininger, D. Combinatorics of Small Molecular Structures. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; pp 425−430.
(50) Walters, W. P.; Stahl, M. T.; Murcko, M. A. High-throughput "Virtual" Chemistry. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; pp 1225−1237.
(51) McClelland, H. E.; Jurs, P. C. Quantitative Structure−Property Relationships for the Prediction of Vapor Pressures of Organic Compounds from Molecular Structures. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 967−975.

CI000126F