communications in mathematical
match
and in computer chemistry

no. 44, October 2001

MATCDY (44) 15 - 30 (2001)

ISSN 0340 - 6253

USE OF MATHEMATICAL STRUCTURAL INVARIANTS IN THE DEVELOPMENT OF QSPR MODELS

Subhash C. Basak* and Denise R. Mills

Natural Resources Research Institute, University of Minnesota
Duluth, MN 55811, USA
E-mail: sbasak@nrri.umn.edu

This paper is dedicated to Professor Alexandru Balaban on his seventieth birthday in appreciation of his many contributions in chemical graph theory and mathematical chemistry.

Abstract. A set of 373 molecular descriptors was calculated for use in the hierarchical quantitative structure-property relationship (QSPR) modeling of normal boiling point for a database consisting of 1015 diverse chemicals. Topological descriptors (partitioned into topostructural and topochemical classes) and geometrical descriptors were utilized in a hierarchical fashion in the model development process. The results indicate that the quickly and easily calculated topological descriptors are sufficient in accounting for the majority of the variance and that the geometrical descriptors improve the quality of the model only minimally.

INTRODUCTION

The last three decades of the twentieth century has witnessed many important developments in the formulation of concepts for the characterization of molecular structure using mathematical invariants [1, 2] Many of these contributions originated from applications of graph theoretical and topological concepts to chemical structure. A graph G = [V, E] consists of an ordered pair where V represents a nonempty set of vertices and E symbolizes a set of edges [3]. When V represents the set of atoms in a molecule and E represents the set of bonds in the molecule, the graph G becomes a molecular graph. The edge set E usually represents covalent chemical bonds between atoms, but that not necessarily have to be the case. In fact, elements of the set E can represent any type of bonds or even weak or

nonbonded interactions among atoms; and it has been argued by Basak et al. that weighted pseudographs constitute a very versatile model capable of representing many diverse classes of chemical species [4].

Mathematical characterization of graphs, including molecular graphs, can be accomplished using graph invariants. A graph invariant is a graph theoretic property which has the same value for isomorphic graphs [3, 5]. A graph invariant may be a polynomial, a sequence of numbers or a single number. A single real number characterizing a molecular graph is usually called a topological index (TI). A particular TI quantifies certain aspects of molecular structure and is sensitive to such chemically interesting features as size, shape, symmetry, branching, cyclicity, heterogeneity of atomic neighborhoods, bonding patterns, etc [5].

The last quarter of the twentieth century has witnessed an upsurge of interest in the use of graph invariants in the formulation of quantitative structure-activity/ property relationship (QSAR/ QSPR) models. This trend in research has been fueled primarily by two factors: a) availability of a large number of TIs which are easily calculable and useful parameters for QSARs/ QSPRs and b) a need for the quick estimation of properties of large numbers of real or hypothetical chemical structures for many practical situations in drug discovery and hazard assessment of chemicals. Various TIs have been defined to characterize different aspects of molecular structure. Many of them have been successfully used in the ordering of isomeric and closely related structures, to quantify the degree of molecular branching, etc. [6-10] Attempts have been made by various authors to define parameters with increased discriminatory power [11-14]. Such TIs have been useful in the QSPRs of congeneric sets of chemicals when the ordering derived from TIs paralleled the ordering of the molecules with respect to certain physicochemical properties of interest. Based on the original work of Randić in the formulation of the connectivity index [8], Kier and Hall developed a set of valence connectivity and electrotopological indices based on molecular graphs properly weighted to capture chemically relevant information [15]. In another recent development, Randić et al have pioneered the development of optimal descriptors for which the associated weights are variables and different values of the variables are used in correlation studies until an optimal model is obtained [16-18].

While such interesting developments in the formulation and interpretation of TIs were taking place, one line of research of our group has been involved in is the use of TIs vis-a-vis physicochemical properties in the formulation of QSARs/QSPRs. The major goal of such studies has been to see whether TIs can be used to produce predictive models as good as those

derived by using measured physicochemical properties. In another frontier we have been active in identifying from the large pool of TIs those indices which are minimally correlated [19-23]. Such parameters and other indices derived from them have been used by our group in the formulation of methods for measuring similarity/ dissimilarity of chemicals [24-33], selecting analogs of chemicals from large and diverse databases, clustering real and virtual chemical libraries [27], as well as in developing hierarchical QSARs [34-41].

In the realm of hierarchical QSAR development, different classes of TIs, viz., topostructural and topochemical indices, have been used in combination with other types of structural parameters such as 3-D and quantum chemical parameters to derive predictive models. In one of our earlier papers, we formulated a QSPR for a set of 1023 chemicals to predict boiling point using a set of 105 TIs [41]. The parameter set in that study did not include some of the important TIs such as the electrotopological and the kappa shape indices. Therefore, in this paper we have attempted to develop QSPRs of essentially the same set of chemicals with an enhanced set of 373 topostructural, topochemical, and geometrical parameters.

METHODS

Database

The database represents a subset of the Toxic Substances Control Act (TSCA) Inventory [42] for which normal boiling point data was available and HB_1, a measure of the hydrogen bonding potential of a chemical, is equal to zero. HB_1 was calculated by a H-Bond, a software program developed by Basak et al. [43]. The normal boiling point of the compounds in the database ranges from -127.8 to 524 ° C. The chemical diversity of the database is summarized in Table 1.

TABLE 1. Summary of Chemical Diversity of Normal Boiling Point Database.

Compound classification	Number of compounds
Hydrocarbons	548
Alkanes, Cyclic alkanes	195
Alkyl benzenes	80
Polycyclic aromatic hydrocarbons	49
Aromatics	288
Fused rings	55
Halogens	365
Thiols	29
Sulfides	73

Calculation of Molecular Descriptors

The majority of the topological descriptors were calculated using software developed by Basak *et al.*, including POLLY 2.3. [44] The topological descriptors include Wiener number [45], molecular connectivity indices developed by Randić [8] and Kier and Hall [15], frequency of path lengths of varying size [15], information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić [10], Roy et al.[46], Basak et al. [4,7, 47-49], as well as those of Raychaudhury *et al.* [7] , parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs [4,7,47-49], and Balaban's J indices [11-14], as well as the triplet indices [23, 50]. The triplets result from a matrix, main diagonal column vector, and free term column vector which are converted into a system of linear equations. The notation used to represent the vectors and matrices is as follows:

A = Adjacency matrix
V = Vertex degree
S = Distance sum
N = Total number of vertices in the graph
Z = Atomic number
D = Distance matrix
1 = Unity matrix.

After the system of N linear equations is solved, the local vertex invariants, x_i, are assembled into a triplet descriptor based on one of the following operations:

1. Summation, $E_i x_i$;
2. Summation of squares, $E_i x_i^2$;
3. Summation of square roots, $E_i x_i^{1/2}$;
4. Sum of inverse square root of cross-product over edges ij, $E_{ij} (x_i x_j)^{-1/2}$;
5. Product, $N(E_i x_i)^{1/N}$.

Additional topological descriptors, including an extended set of molecular connectivity indices, electrotopological state descriptors, general polarity descriptors, and hydrogen bonding descriptors, were calculated by Molconn-Z 3.50 [51]. The set of topological

descriptors was partitioned into two distinct subsets: topostructural (TS) and topochemical (TC). Topostructural descriptors encode information about the adjacency and distances of atoms in molecular structures irrespective of the chemical nature of the atoms, while topochemical descriptors encode information regarding the connectivity as well as specific chemical properties of the atoms making up the molecule. A total number of 363 topological descriptors was calculated for use in the current study.

Ten geometrical descriptors were used, including six kappa shape indices which were also calculated by Molconn-Z. Van der Waals volume, V_W, was calculated using Sybyl 6.2 [52]. In addition, two variants of the 3-D Wiener number, ^{3D}W and $^{3D}W_H$, based on the hydrogen-suppressed and hydrogen-filled geometric distance matrices, respectively, were also calculated by Sybyl using a SPL (Sybyl Programming Language) program developed by our group.

A complete list of the 373 parameters calculated for use in the current study, including brief descriptions, is provided in Table 2.

TABLE 2. Symbols, Definitions and Classification of Molecular Descriptors

	Topostructural (TS)
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
$^h\chi$	Path connectivity index of order h = 0-10
$^h\chi_C$	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 = 3-10
P_h	Number of paths of length h = 0-10
J	Balaban's J index based on topological distance
n rings	Number of rings in a graph
n circ	Number of circuits in a graph
DN^2S_y	Triplet index from distance matrix, square of graph order, and distance sum; operation y = 1-5
$DN^2 1_y$	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1-5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation y = 1-5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation y = 1-5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation y = 1-5
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation y = 1-5
DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1-5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1-5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 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, distance sum, and vertex degree; operation y = 1-5
DSV_y	Triplet index from distance matrix, distance 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 (TC)

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 vertices
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
SIC_r	Structural information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} (r = 0-6) order neighborhood of vertices 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
HB_1	Hydrogen bonding parameter
AZV_y	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1-5
AZS_y	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1-5
ASZ_y	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1-5
AZN_y	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1-5
ANZ_y	Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1-5
DSZ_y	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1-5
DN^2Z_y	Triplet index from distance matrix, square of graph order, and atomic number; operation y = 1-5
nvx	Number of non-hydrogen atoms in a molecule
n elem	Number of elements in a molecule
fw	Molecular weight
$^{h}\chi^{v}$	Valence path connectivity index of order h = 7-10
$^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 7-10
si	Shannon information index
totop	Total Topological Index t
sumI	Sum of the intrinsic state values I
sumdelI	Sum of delta-I values

tets 2	Total topological state index based on electrotopological state indices
phia	Flexibility index (kp1* kp2/nvx)
Idcbar	Bonchev-Trinajstić information index
IdC	Bonchev-Trinajstić information index
Wp	Wiener p
P f	Plattf
Wt	Total Wiener number
knotp	Difference of chi-cluster-3 and path/cluster-4
knotpv	Valence difference of chi-cluster-3 and path/cluster-4
nclass	Number of classes of topologically (symmetry) equivalent graph vertices
numHBd	Number of hydrogen bond donors
numHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 bonded to unsaturated C atoms
SHvin	E-State of C atoms in the vinyl group, =CH-
SHtvin	E-State of C atoms in the terminal vinyl group, $=CH_2$
SHavin	E-State of C atoms in the vinyl group, =CH-, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for –OH, =NH, -NH2, -NH-, -SH, and #CH
SHwHBd	Weak hydrogen bond donor index, sum of C-H Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the E-State values for –OH, =NH, -NH2, -NH-, >N-, -O-, -S-, along with –F and –Cl
Qv	General Polarity descriptor
NHBint$_y$	Count of potential internal hydrogen bonders (y = 2-10)
SHBint$_y$	E-State descriptors of potential internal hydrogen bond strength (y =2-10)
	Electrotopological State index values for atoms types:
	SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, Shother, SHCHnX,
	Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss,Bem,
	SssBH ,SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH,
	SsssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2,
	SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN,
	SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SsssSiH2,

SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SssssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SsssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SssssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb

	Geometrical (3D)
***kp*0**	Kappa zero
***kp*1-*kp*3**	Kappa simple indices
***ka*1-*ka*3**	Kappa alpha indices
V_W	Van der Waals volume
^{3D}W	3D Wiener number based on the hydrogen-filled geometric distance matrix
$^{3D}W_H$	3D Wiener number based on the hydrogen-suppressed geometric distance matrix

HIERARCHICAL QSPR AND STATISTICAL ANALYSIS

QSPR modeling was performed in a hierarchical fashion, utilizing descriptor classes of increasing complexity. The topostructural descriptors are at the lowest level of the hierarchy, followed by the topochemical, and finally the geometrical (3D) descriptors. Initially, a model is developed utilizing only TS descriptors, after which TC descriptors are added to the parameters in the TS model, and the regression analysis is repeated to obtain a TS + TC model. Adding the 3D descriptors to the parameters in the TS +TC model and again performing regression analysis yields a TS + TC + 3D QSPR model.

Prior to analysis, all calculated descriptors with non-negative values were transformed by: ln (descriptor value + 1) due to the fact that some of their scales differed by several orders of magnitude. One was added prior to taking the natural logarithm, as there are descriptors which have a value of zero. A constant was added to the descriptors with negative values, such that their sum was greater than 0, before taking the natural logarithm. Any descriptor with a value of 0 for 98% or more of the compounds was removed from the pool and not used in subsequent analyses. Perfectly correlated descriptors, *i.e.* those having a correlation coefficient of 1.0, were identified using the CORR procedure [53] of the SAS statistical software package. In each case, only one of the perfectly correlated descriptors was retained and used in subsequent analyses. There were 237 descriptors remaining at this point. No

further data reduction was necessary, since the ratio of number of variables to number of observations falls within the limits suggested by Topliss and Edwards [54] for reducing the probability of spurious correlations at the $R^2 \geq 0.7$ level.

Two regression procedures were used for model development, both of which are forms of the REG procedure of the SAS statistical package [53]. When the number of independent variables utilized in the modeling is small, the all possible subsets regression was used. However, when the number of independent variables exceeds approximately 20-25, this regression option is quite time intensive and an alternate option which maximizes the improvement in R^2 was used.

RESULTS

Using only descriptors of the topostructural class, regression analysis yielded the following five-parameter model:

$$BP = 51.3\ {}^{3}\chi - 112.6\ {}^{6}\chi_c - 103.1\ {}^{5}\chi_{ch} + 145.1\ {}^{9}\chi + 56.7\ ASN_4 - 66.8 \qquad (1)$$

$$n = 1015,\ R^2 = 80.4\%,\ s = 39.9\ ,\ F = 827$$

The inclusion of additional parameters in the TS model is not justified, as the resulting improvement in the statistical measures is minimal. For example, even the best 12-parameter model has an R^2 value which explains only an additional 0.9% of the variance and has a standard error only 0.1 ° C lower than that of the five-parameter model above.

When the topochemical descriptors are added to those parameters in the TS model and the regression procedure repeated, we obtain the following eight-parameter TS + TC model:

$$BP = 83.4\ {}^{9}\chi - 39.8\ {}^{3}\chi^{b}_{c} + 200.9\ AZV_4 + 51.9\ HMIN + 24.9\ GMIN - 7.20\ SDSCH - 22.1\ SSF + 39.5\ NUMHBD - 625.8$$

$$n = 1015,\ R^2 = 97.0\ \%,\ s = 15.7\ ,\ F = 4014 \qquad (2)$$

Only one of the parameters from the TS model, ${}^{9}\chi$, has been retained. The improvement in both the explained variance and the standard error are remarkably significant when adding the TC descriptors.

Adding the ten geometrical descriptors to those included in the best TS + TC model given in eq. 2 and performing all possible subsets regression, we obtain the following 11-parameter TS + TC + 3D model:

$$BP = 88.6\ {}^{9}\chi - 54.3\ {}^{3}\chi^{b}_{c} + 212.3\ AZV_4 + 55.4\ HMIN + 18.3\ GMIN - 5.28\ SDSCH - 16.4\ SSF + 38.9\ NUMHBD + 65.9\ V_W - 20.8\ {}^{3D}W - 12.9\ KA_2 - 855.7$$
$$n = 1015,\ R^2 = 97.2\ \%,\ s = 15.1\ ,\ F = 3192 \qquad (3)$$

All of the parameters in the TS + TC model have been retained and three geometrical parameters added. The improvement in the statistical measures due to the addition of the 3D descriptors is not nearly as significant as is seen upon the addition of the TC descriptors. The addition of three geometrical parameters increased the explained variance by merely 0.2 % and the standard error was decreased minimally by 0.6 °C. For this reason, the TS + TC model is felt to be the best of the three models presented in this paper, and a scatter plot of the experimental *vs* the estimated boiling point using eq. 2 is provided in Figure 1.

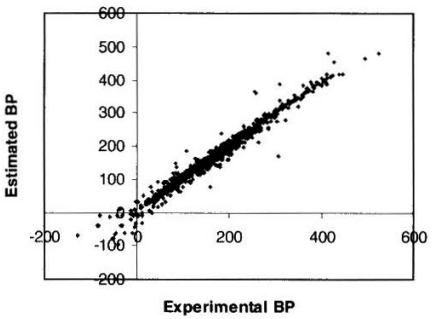

FIGURE 1. Experimental vs Estimated normal boiling point using equation 2.

DISCUSSION

The primary objective of this paper was to study the utility of set if theoretically calculated structural descriptors, including various classes of graph invariants, in the prediction of boiling point of a large and diverse set of 1015 chemicals. Results reported in this paper indicate that the topostructural TIs produce a good predictive model with an R^2 value of 80.4. In an earlier study with a comparable data set, TS indices produced a model of similar quality (Table 3). The addition of TC indices at the next level of hierarchical QSPR model development resulted in a significantly improved model with an R^2 value of 97.0 and standard error of 15.7. This is in line with our earlier published results which show that the addition of TC indices usually results in significant improvement in model quality. [21, 34-41, 55]. The addition of geometrical parameters to those already included in the TS + TC model did not result in significant improvement in the model quality. The descriptors which appear to play a role in the prediction of boiling include the connectivity indices (encoding information regarding size and degree of branching), count of long path lengths, and atom-type electrotopological state indices. Each of the TS and TS + TC models also contain a triplet index which is derived from the adjacency matrix.

TABLE 3. Comparative Results of Current and Previous Normal Boiling Point Models.

	TS Model	TS + TC Model	TS + TC + 3D Model
Previous Study (Ref. 41)	R^2 = 81.2% s = 39.7 8 Descriptors W O $^3\chi$ $^6\chi$ $^6\chi_c$ $^5\chi_{ch}$ $^5\chi_{pc}$ P_{10}	R^2 = 96.1% s = 18.0 8 Descriptors $^6\chi$ $^6\chi^b$ $^3\chi^b_c$ $^0\chi^v$ $^2\chi^v$ $^5\chi^v_{pc}$ P_{10} IC_0	R^2 = 96.3% s = 17.6 10 Descriptors $^6\chi$ $^6\chi^b$ $^3\chi^b_c$ $^0\chi^v$ $^2\chi^v$ $^5\chi^v_{pc}$ P_{10} IC_0 ^{3D}W $^{3D}W_H$
Current Study	R^2 = 80.4% s = 39.9 5 Descriptors $^3\chi$ $^6\chi$ $^5\chi_{ch}$ $^9\chi$ ASN_4	R^2 = 97.0% s = 15.7 8 Descriptors $^9\chi$ $^3\chi^b_c$ AZV_4 HMIN GMIN SDSCH SSF NUMHBD	R^2 = 97.2% s = 15.1 11 Descriptors $^9\chi$ $^3\chi^b_c$ AZV_4 HMIN GMIN SDSCH SSF NUMHBD V_W ^{3D}W KA_2

The major findings of this paper are: a) The TS and TC indices explain most of the variance in the data , b) the addition of geometrical parameters to the TS + TC model did not

result in any significant improvement in model quality. Likewise, our previous hierarchical QSAR/QSPR studies with physicochemical, biomedicinal, and toxicological properties have shown that the addition of geometrical and quantum chemical parameters to TS + TC models produced marginal or no improvement in model quality. c) using the expanded set of descriptors, some improvement in the model was obtained as compared to the earlier study by Basak et al. [41] in which fewer number initial parameters were calculated. This trend was also seen in another study by Basak et al. [55] when using the expanded set of descriptors in developing QSPR models for vapor pressure.

ACKNOWLEDGEMENT

This is contribution number 293 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported by Grant F49620-98-1-0015 from the U. S. Air Force.

REFERENCES AND NOTES

[1] A. T. Balaban, *Chemical Applications of Graph Theory*, Academic Press, London, 1976.

[2] J. Devillers, A. T. Balaban, Eds. *Toplogical Indices and Related Desriptors in QSAR and QSPR*, Gordon and Breach, The Netherlands, 1999.

[3] F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[4] S. C. Basak, V. R. Magnuson, G. J. Niemi, R. R. Regal, *Discrete Appl. Math.* **1988**, *19*, 17-44.

[5] N. Trinajstić, *Chemical Graph Theory*, Vols. 1 and 2, CRC Press, Boca Raton, FL, 1983.

[6] K. Balasubramanian, S.C. Basak, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 367-373.

[7] C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy, S. C. Basak, *J. Comput. Chem.* **1984**, *5*, 581-588.

[8] M. Randić, *J. Am. Chem. Soc.* **1975**, *97*, 6609-6615.

[9] S. C. Basak, Information Theoretic Indices of Neighborhood Complexity and Their Applications. In: *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds.: J. Devillers, A.T. Balaban. Gordon and Breach Science Publishers, The Netherlands, 1999, pp.563-593.

[10] D. Bonchev, N. Trinajstić, *J. Chem. Phys.* **1977**, *67*, 4517-4533.

[11] A. T. Balaban, *Chem. Phys. Lett.* **1982**, *80*, 399-404.

[12] A. T. Balaban, *Pure Appl. Chem.* **1983**, *55*, 199-206.

[13] A. T. Balaban, *Commun. Math. Chem. (MATCH)*, **1986**, *21*, 115-122.

[14] A. T. Balaban, P. Filip, *Commun. Math. Chem. (MATCH)*, **1984**, *16*, 163-190.

[15] a) L. B. Kier, L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research.* Academic Press, New York, 1976; (b) L. B. Kier, L. H. Hall, *Molecular Connectivity in Structure-Activity Studies*, Research Studies Press, Letchworth, 1986; (c) L. B. Kier, L. H. Hall, *Molecular Structure Description: The Electrotopological State,* Academic Press, New York, 1999.

[16] M. Randić, S. C. Basak, *SAR QSAR Environ. Res.* **2000**, *11*, 1-23.

[17] M. Randić, D. Mills, S.C. Basak, *Int. J. Quant. Chem.* **2000**, *80*, 1199-1209.

[18] M. Randić, S. C. Basak, Variable molecular descriptors. In: *Some Aspects of Mathematical Chemistry*, Eds.: D. K. Sinha, S. C. Basak, R. K. Mohanty, I. N. Basumallick, Visva-Bharati University, India, in press.

[19] S. C. Basak, D. K. Harriss, V. R. Magnuson, *J. Pharm. Sci.* **1984**, *73*, 429-437.

[20] S. C. Basak, D.P. Gieschen, V.R. Magnuson, *Environ Toxicol . Chem.* **1984**, *3*, 191-199.

[21] S. C. Basak, L. J. Monsrud, M. E. Rosen, C. M. Frane, V. R. Magnuson, *Acta Pharm. Yugosl.* **1986**, *36*, 81-95.

[22] S. C. Basak, V. R. Magnuson, G. J. Niemi, R. R. Regal, G. D. Veith, *Mathematical Modelling*, **1987**, *8*, 300-305.

[23] S. C. Basak, A. T. Balaban, G. D. Grunwald, B. D. Gute, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891-898.

[24] S. C. Basak, G. D. Grunwald, *Math. Modelling Sci. Comput.* in press.

[25] S. C. Basak, B. D. Gute, G. D. Grunwald, *Mathl. Modelling Sci. Computing*, in press.

[26] S. C. Basak, B. D. Gute, G. D. Grunwald, *Mathl. Modelling Sci. Computing*, in press.

[27] S. C. Basak, D. Mills, B. D. Gute, A. T. Balaban, K. Basak, G. D. Grunwald, Use of Mathematical Structural Invariants in Analyzing Combinatorial Libraries: A Case Study with Psoralen Derivatives. In: *Some Aspects of Mathematical Chemistry*, Eds.: D. K. Sinha, S. C. Basak, R. K. Mohanty, I. N. Basumallick. Visva-Bharati University, India, in press.

[28] B. D. Gute, G. D. Grunwald, D. Mills, S. C. Basak, *SAR QSAR Environ. Res.* in press.

[29] S. C. Basak, B. D. Gute, G.D. Grunwald, *Mathl. Modelling Sci. Computing*, in press.

[30] S. C. Basak, B. D. Gute, G. D. Grunwald, *Mathl. Modelling Sci. Computing*, in press.

[31] S. C. Basak, G. D. Grunwald, *SAR QSAR Environ. Res.* **1995**, *3*, 265-277.

[32] S. C. Basak, G. D. Grunwald, Molecular similarity and estimation of molecular properties.. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 366-372.

[33] S. C. Basak, G. D. Grunwald, *SAR QSAR Environ. Res.* **1994**, *2*, 289-307.

[34] S. C. Basak, B. D. Gute,, G. D. Grunwald, 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*, Eds.: J. Devillers, A.T. Balaban. Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 675-696.

[35] S. C. Basak, B. D. Gute, G. D. Grunwald, Relative Effectiveness of Topological, Geometrical, and Quantum Chemical Parameters in Estimating Mutagenicity of Chemicals. In: *Quantitative Structure-activity Relationships in Environmental Sciences VII,* Eds.: F. Chen, G. Schuurmann. SETAC Press, Pensacola, FL., 1998, pp. 245-261.

[36] S. C. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian, D. Opitz, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 885-890.

[37] S. C. Basak, B. D. Gute, G. D. Grunwald, *SAR QSAR Environ. Res.* **1999**, *10*, 117-129.

[38] B. D. Gute, G. D. Grunwald, S. C. Basak, *SAR QSAR Environ. Res.* **1999**, *10*, 1-15.

[39] B. D. Gute, S. C. Basak, *SAR QSAR Environ. Res.* **1997**, *7*, 117-131.

[40] S. C. Basak, B. D. Gute, G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651-655.

[41] S. C. Basak, B. D. Gute, G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054-1060.

[42] C. M. Auer, J. V. Nabholz, K. P. Baetcke, *Environ. Health Perspect.* **1990**, *87*, 183-197.

[43] S. C. Basak, *H-Bond*, Copyright of the University of Minnesota, 1988.

[44] S. C. Basak, D. K. Harriss, V. R. Magnuson, POLLY 2.3, Copyright of the University of Minnesota, 1988.

[45] N. Wiener, *Cybernetics*, Wiley, New York, 1948.

[46] A. B. Roy, S. C. Basak, D. K. Harriss, V. R. Magnuson, Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In: *Mathematical Modelling in Science and Technology*, 4th Internat. Conf. Zurich, Eds.: X. J. R. Avula, R. E. Kalman, A. I. Liapis, E. Y. Rodin. Pergamon Press, New York, 1983, pp. 745-750.

[47] S. C. Basak, *Med. Sci. Res.* **1987**, *15*, 605-609.

[48] S. K. Ray, S. C. Basak, C. Raychaudhury, A. B. Roy, J. J. Ghosh, *ICRS Med. Sci.* **1982**, *10*, 933-934.

[49] S. C. Basak, V. R. Magnuson, *Arzneimitt.-Forsch. Drug Res.* **1983**, *33*, 501-503.

[50] P. A. Filip, T. S. Balaban, A. T. Balaban, *J. Math. Chem.* **1987**, *1*, 61-83.

[51] (a) MOLCONN-Z, Hall Associates Consulting, Quincy, MA (b) L. H. Hall, M. B. Kier, *Molecular Structure Description: The Electrotopological State*, Academic Press, New York, 1999.

[52] Tripos Associates, Inc., St. Louis, MO, 1994.

[53] SAS/STAT *User's Guide*, Release 6.03 ed., SAS Institute, Cary, NC, 1988.

[54] J. G. Topliss, R. P. Edwards, J. Med. Chem., **1979**, *22*, 1238-1244.

[55] S. C. Basak, D. R. Mills, *J. Chem. Inf. Comput. Sci.*, submit