Topological Indices: Their Nature and Mutual Relatedness

Subhash C. Basak,*[,†] Alexandru T. Balaban,[‡] Gregory D. Grunwald,[†] and Brian D. Gute[†]

Natural Resources Research Institute, University of Minnesota−Duluth, Duluth, Minnesota 55811, and Organic Chemistry Department, Polytechnic University Bucharest, Splaiul Independentei 313, 77206 Bucharest, Romania

Received September 9, 1999

We calculated 202 molecular descriptors (topological indices, TIs) for two chemical databases (a set of 139 hydrocarbons and another set of 1037 diverse chemicals). Variable cluster analysis of these TIs grouped these structures into 14 clusters for the first set and 18 clusters for the second set. Correspondences between the same TIs in the two sets reveal how and why the various classes of TIs are mutually related and provide insight into what aspects of chemical structure they are expressing.

INTRODUCTION

A major part of the current research in mathematical chemistry, chemical graph theory, and quantitative structure−activity/property relationship studies involves topological indices. Topological indices (TIs) are numerical graph invariants that quantitatively characterize molecular structure. A graph $G = (V, E)$ is an ordered pair of two sets V and E, the former representing a nonempty set and the latter representing unordered pairs of elements of the set V. When V represents the atoms of a molecule and elements of E symbolize covalent bonds between pairs of atoms, then G becomes a *molecular graph* (or *constitutional graph*, because there is no stereochemical information). Such a graph depicts the topology of the chemical species. A graph is characterized using graph invariants. An invariant may be a polynomial, a sequence of numbers, or a single number. A numerical graph invariant (i.e., a single number) that characterizes the molecular structure is called a topological index.

OVERVIEW OF TOPOLOGICAL INDICES USED IN THE PRESENT STUDY

A large number of topological indices have been defined and used.[1−11] The majority of TIs are derived from the various matrices corresponding to molecular graphs. The adjacency matrix $\mathbf{A}(G)$ and the distance matrix $\mathbf{D}(G)$ of the molecular graph G have been most widely used in the formulation of TIs. Integer-number local vertex invariants (LOVIs) are the vertex degrees (v_i) and the distance sums (distasums, d_i) resulting from summation over rows or columns of entries in the adjacency and distance matrices, respectively. By mathematical operations performed on such LOVIs, one can obtain a molecular descriptor, *i.e.*, a topological index. Wiener's index W (eq 1),[2] the Zagreb group index M_1 (eq 2),[11] Randić's connectivity index, χ (eq 3),[4] the higher-order connectivity indices, ${}^n\chi$, for paths of length n defined by Kier and Hall,[5] and the J index (eq 4)[6] fall in this category.

$$W = (\sum_i d_i)/2 \tag{1}$$

$$M_1 = \sum_i v_i^2 \tag{2}$$

$$\chi = \sum_{ij} (v_i v_j)^{-1/2} \tag{3}$$

$$J = [q/(\mu + 1)] \sum_{ij} (d_i\, d_j)^{-1/2} \tag{4}$$

The summations in formulas 3 and 4 are over all edges $i-j$ in the hydrogen-depleted graph. The numbers q of graph edges and μ of cycles in the graph are introduced into formula 4 in order to avoid the automatic increase of J with graph size and cyclicity. Indeed, for an infinite linear carbon chain it was demonstrated that $J = \pi = 3.14159$. The nature of atoms can be taken into account by means of parameters based on their relative atomic numbers, electronegativities, or covalent radii, with respect to those of carbon atoms, multiplying the corresponding distasum in formula 4 for J.

The mean-square-root distance D derived from all topological distances (denoted by i in the next formula) is defined as[6b]

$$D = [(\sum_i i^2)/(\sum_i i)]^{1/2} \tag{5}$$

For taking into account the chemical nature of atoms symbolized by vertices, Kier and Hall advocated the use of "valence connectivity indices".[5a,b] These are calculated with formulas similar to Randić's (eq 3), but products of edge end point (or path vertex) invariants are no longer of vertex degrees but of weights (valence delta values δ_i) given by formula 5

$$\delta_i = (Z_i^{\mathrm{v}} - H_i)/(Z_i - Z_i^{\mathrm{v}} - 1) \tag{6}$$

where Z_i^{v} stands for the number of valence electrons in atom i, Z_i is its atomic number, and H_i is the number of hydrogen atoms attached to atom i.

The most recent additions to the Kier−Hall armamentary of TIs are electrotopological state indices.[5c]

* Corresponding author. Tel: (218)720-4230. Fax: (218)720-4328. E-mail: sbasak@nrri.umn.edu.
† University of Minnesota.
‡ Polytechnic University Bucharest.

10.1021/ci990114y CCC: $19.00
Published on Web 02/18/2000

Another class of molecular descriptors, the information-theoretic indices, are derived from an entirely different reasoning. In this case, the complexity or mode of partitioning of structural features is decomposed into disjoint subsets using an equivalence relation; a molecular complexity index is then computed using Shannon's idea of information content or complexity.[12] Real-number local vertex invariants (LOVIs), on the other hand, may also be defined starting from different matrices other than **A**(*G*) or **D**(*G*) or by applying information theory at the vertex level. Thus, topological indices *U*, *V*, *X*, and *Y* were defined.[13] Bonchev and Trinajstić described several information-theoretic TIs reviewed thoroughly in Bonchev's book.[7]

The information-theoretic indices developed by Basak and co-workers take into account all atoms in the constitutional formula (hydrogens also being included), and one considers the information content provided by various classes of atoms based on their topological neighborhood. There are three main types of informational indices developed by Basak et al.: IC (mean information content or complexity of a hydrogen-filled graph, with vertices grouped into equivalence classes having *r* vertices; the equivalence is based on the nature of atoms and bonds, in successive neighborhood groups); CIC (complementary information content); and SIC (structural information content), and they are not intercorrelated with other TIs. In the following formula, the summation spans the range from $i = 1$ to $i = r$:

$$\mathrm{IC}_r = -\sum\nolimits_i p_i \log_2 p_i \qquad (10)$$

$$\mathrm{SIC}_r = \mathrm{IC}_r/\log_2 N \qquad (11)$$

$$\mathrm{CIC}_r = \log_2 N - \mathrm{IC}_r \qquad (12)$$

The probability that a randomly selected vertex occurs in the *i*th equivalence class is denoted by p_i. The IC_r, SIC_r, and CIC_r indices can be calculated for different orders of neighborhoods, *r* ($r = 0, 1, 2, ..., \rho$), where ρ is the radius of the molecular graph *G*. At the 0th-order level, the atom set is partitioned solely on the basis of its chemical nature; at the level of the first-order topological neighborhood, the atoms are partitioned into disjoint subsets on the basis of their chemical nature and their first-order bonding topology. At the next level, the atom set is decomposed into equivalence classes using their chemical nature and bonding pattern up to the second-order bonded neighbors. The process is continued until consideration of higher-order neighbors does not yield further increase in the number or composition of disjoint subsets.

A large variety of real-number local vertex invariants, and thence a larger variety of TIs, were described on the basis of converting a matrix (**A** or **D** for instance) into a system of linear equations. This is done by means of two column vectors that can convey topological, chemical, or numerical information. One nonzero vector is the free term of the system of equations. The other one (which may be zero, but this restricts the choices on available supplementary information) becomes the main diagonal of the matrix (if both vectors were zero, then some negative LOVIs would result with difficulties of interpretation). These vectors may be the following integers: *Z* (atomic number of the atom corresponding to each vertex), *V* (vertex degree), *I* (identity), *N* (number of non-hydrogen atoms, or order of the graph), N^k (power *k* of *N*). Less frequently, one may use for periodicity of chemical properties real numbers: *S* (electronegativity) or *R* (covalent radius) of the atom corresponding to each vertex. The resulting matrix with the vector for the main diagonal constitutes the set of coefficients for the *N* unknowns that represent the real-number LOVIs of the *N* vertices. The triplet (matrix, vector for the main diagonal and vector for the free term) also serves as notation for LOVIs and for the derived TIs. After the system of *N* linear equations is solved, the LOVIs (x_i) are assembled into a "triplet TI" based on one of the following operations:

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, $N[\Pi_i x_i]^{1/N}$.

Numbers 1−5 of the above operations after the triplet complete the notation of the triplet TIs.[14]

To conclude this brief review of TIs, one should mention recent progress that includes other matrices such as the reciprocal distance matrix that yields Harary indices,[15] the regressive distance matrices,[16] the Szeged matrix,[17] and the resistance distance matrix that affords Kirchhoff indices.[18] So-called optimal structural descriptors can be obtained from some TIs by varying some parameters and thereby adapting them to the database;[19] alternatively, in Randić-type formulas (eqs 3, 4) the exponent is allowed[20] to differ from $^1/_2$. Three-dimensional molecular descriptors can be derived from geometrical and topological structural features of molecules.[21]

Each of the indices above-discussed is a "global" parameter; i.e., it quantifies certain aspects of the entire molecular structure using a single number.

It is clear from the above discussion that the set of TIs is a group of heterogeneous entities. They have been defined to characterize molecular structure on the basis of distinct objectives and motivations. Despite their distinctive characteristics, TIs share certain common features. A topological index maps a set of chemicals *C* into the set *R* of real or integer numbers. Therefore, TIs quantify some general aspects of molecular architecture such as size, shape, symmetry, bonding type, cyclicity, branching pattern, etc.

Topological indices have been used for isomer discrimination, quantification of the structural similarity/dissimilarity of molecules, and prediction of property/activity from structure.[19] The widespread use of TIs obviously encourages one to ask some fundamental questions about them: What is the fundamental nature of TIs? To what degree are they intercorrelated? How does one extract orthogonal information from TIs?

The intercorrelation of TIs was studied earlier with a limited set of invariants. Thus, Motoc and Balaban[22] described graphically the intercorrelations of the few TIs known until 1981. These aspects were reviewed in the early 1980s.[23] Basak et al. studied the mutual relatedness of a set of 90 TIs calculated for a set of 3692 diverse chemicals.[24] A third study by Todeschini et al. will be discussed in the last section of this paper.

All such studies were limited in the sense that they analyzed data on a smaller and less diverse group of TIs. Therefore, in this paper, we have studied the mutual

Table 1. Summary of Chemical Classes or Features in Databases Analyzed

chemical classes or features	database A (hydrocarbons)	database B (diverse)
total number of compounds	139	1037
hydrocarbons	139	565
alkanes, cyclic alkanes	73	206
aromatics	66	288
alkyl benzenes	29	80
fused rings	37	56
polycyclic aromatics	37	49
non-hydrocarbons	0	472
halogen-containing compounds		359
heteroatom-containing compounds (sulfur or phosphorus)		101
Compounds containing both halogens and heteroatoms		12
organosulfides		105
organophosphorus		8

relatedness of a set of 202 TIs. We have also tried to extract useful and orthogonal structural information from the calculated TIs. This study also reports, for the first time, a comprehensive discussion of Basak's information content indices (IC_r, SIC_r, CIC_r), the triplet indices (proposed by one of the present authors), and Balaban's average distance-based connectivity index J as compared to the traditional and more widely used indices.

The goal of this paper is two-fold: (a) to study the degree of intercorrelation among the various types of topological indices and (b) to extract mutually uncorrelated (orthogonal) topological parameters that can be used for QSAR/QSPR studies, quantitation of intermolecular similarity/dissimilarity, and characterization of real and virtual combinatorial libraries. To this end, we studied the mutual relatedness of a set of more than 200 topological indices in this paper.

METHODS

Chemical Databases. There were two sets of chemicals analyzed in this study: a set of 139 hydrocarbons to represent a moderately homogeneous set of chemicals and a set of 1037 diverse chemicals. The hydrocarbons consisted of 73 C3−C9 alkanes, 29 alkylbenzenes, and 37 polycyclic aromatic hydrocarbons.[25] The diverse set of 1037 compounds consists of those chemicals from the U.S. EPA ASTER system[26] for which a measured boiling point was available and hydrogen-bonding potential (as measured by HB1 = 0) did not exist. The composition of these data sets is indicated in Table 1. Table 2 presents the list of all 202 parameters calculated in this study.

Calculation of TIs. The TIs calculated for this study (some of which are included in Table 2) include Wiener number W,[2] molecular connectivity indices as calculated by Randić[4] and Kier and Hall,[5] frequency of path lengths of varying size,[5] information-theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić,[7] Roy et al.,[27] Basak et al.,[28−31] and Raychaudhury et al.,[32] parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs,[28−32]

Table 2. Symbols and Definitions of Topological Parameters

index	definition
$\overline{I^{W}_{D}}$	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I^{W}_{D}}$	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
$\overline{H^{D}}$	graph distance complexity
$\overline{IC}$	information content of the distance matrix partitioned by frequency of occurrences of distance h
O	order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
I_{ORB}	information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
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
IC_r	mean information content or complexity of a graph based on the rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complementary information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
${}^{h}\chi$	path connectivity index of order $h = 0-6$
${}^{h}\chi$	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 $h = 3-6$
${}^{h}\chi^{b}$	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 $h = 4-6$
P_h	number of paths of length $h = 0-10$
J	Balaban's J index based on distance
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
triplet	Global 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 non-hydrogen atoms (N and N^2), vertex degree (V), or numerical constants (1). Notation is described by triplets (e.g., AZV). Results are weightings for each atom in a molecule. These weights are combined by five possible formulas: 1 = sum of weights, $\sum_i x_i$; 2 = sum of squared weights $\sum_i x_i^2$; 3 = sum of square root of weights $\sum_i x_i^{1/2}$; 4 = sum of cross-products $\sum_i (x_i \cdot x_j)^{-1/2}$; and 5 = product of weights $N \cdot [\sum_i x_i]^{1/N}$.

and Balaban's *J* indices[6] as well as triplet indices.[14] The majority of the TIs were calculated using the program POLLY 2.3.[33] The *J* indices and triplet indices were calculated using software developed in-house by the authors.

STATISTICAL ANALYSIS

For both sets of chemicals, the computed TIs were transformed by the natural logarithm of the index plus a constant, generally 1. This was done since the scale of some indices may be several orders of magnitude greater than that of other indices.

For each set, a technique known as variable clustering was performed using the SAS procedure VARCLUS.[34] The variable-clustering procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional. This is accomplished by a repeated principal-components analysis of the sets of indices. The initial principal-component analysis examines all indices and defines two principal components or eigenvectors. If the eigenvalue for the second component is >1.0, the indices are split into separate clusters by correlating the indices with the first and second principal components. Those indices most correlated with the first component form one cluster and those indices most correlated with the second component form another cluster, thus forming two disjoint clusters. A principal-component analysis is then performed for each cluster of indices, with the cluster being split if the eigenvalue for the second component is >1.0. The procedure is repeated until the second eigenvalue is <1.0 for all clusters.

RESULTS AND DISCUSSION

The first database (denoted by A) consists of 139 hydrocarbons (alkanes, alkylbenzenes, and polycyclic aromatics) and 162 TIs. The number of indices examined was reduced from the original 202 by removing all but one of the degenerate (i.e., correlation of 1.0) indices and those indices that were constant (0.0) for all chemicals. The second database (denoted by B) is a diverse one and contains 1037 chemical structures and 176 nondegenerate, nonconstant indices.

The results of the variable-cluster analysis will be presented, first discussing how the descriptors (variables) for database A become clustered and then surveying the descriptor clustering for database B, as well as the correspondence between these clusters. Intercluster correlation will then be described.

The clusters have been ordered according to decreasing numbers of descriptors in each cluster; when clusters contain the same number of descriptors, the numbering of the corresponding clusters is arbitrary.

In Figure 1, one can see, in graphical form, on the left-hand side the points denoting the clusters that group together the descriptors for the hydrocarbon database A and on the right-hand side those corresponding to the diverse database B. Each cluster is denoted by a letter (A or B) and a number. The total number of variables in each cluster is written under each point. Full lines connect A-type with B-type clusters, having inscribed on them the numbers of descriptors common to each pair of clusters; when no number is inscribed, this indicates a single common descriptor. Dashed side lines denote the descriptors that do not have counterparts in the other set of clusters, and the associated numbers on these

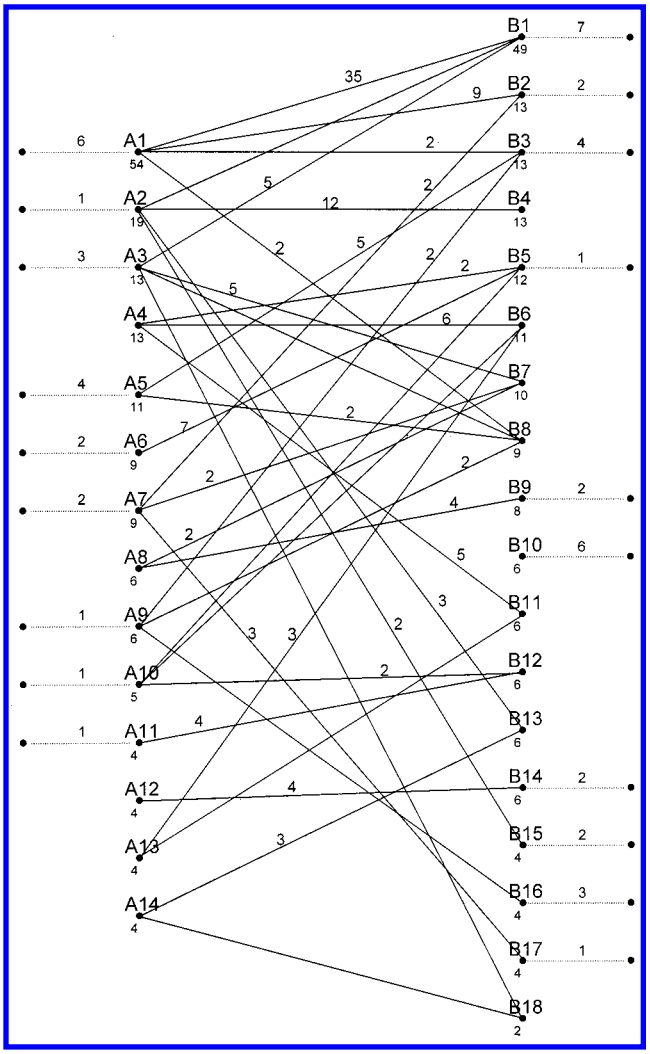

Figure 1. Associations between clusters of descriptors for the hydrocarbon database (A-type clusters) and the database with diverse compounds (B-type clusters). Solid lines connect A-type descriptors with B-type descriptors, and the numbers of common descriptors are indicated on such lines (when no number is indicated, there is just one common descriptor). Dashed lateral lines indicate descriptors that have no correspondence for the other type.

side lines indicate the numbers of such "orphan" descriptors. Because the two data sets differ both in the numbers of compounds and in their structures, it is normal to expect that clusters for one data set will have counterparts in several clusters in the other data set. This is indeed what was found to happen, as will be shown below when the diverse data set will be analyzed.

Only in a single case have we found a one-to-one correspondence between clusters of descriptors corresponding to the two data sets (A12 and B14). Nevertheless, in several instances (A6, A11; B4, B9, B15, B16, and B17), a cluster for one data set (say, A) was found to have all its descriptors in common with only one cluster of the other data set (say, B); however, this latter cluster also contains descriptors found in more than one cluster of the other set.

Clustering of Descriptors for Hydrocarbons. The descriptors for database A are grouped in 14 clusters summarized in Table 3. Cluster A1 has 54 from the total of 162 descriptors; therefore, it groups together about one-third of all variables. These descriptors depend on both the shape and the size (magnitude) of the molecular graph; such

Table 3. Summary of Variable Clustering for 139 Hydrocarbons

cluster	number of variables	representative variables (max. 25% of total listed)
A1	54	DN^2Z_4, DN^2N_4, P_0, AZV_4, ASZ_4, ANN_3, ANN_5, AZN_3
A2	19	${}^6\chi$, P_7, ${}^5\chi$, ${}^6\chi^b$, ${}^6\chi^v$
A3	13	${}^0\chi^b$, ${}^0\chi^v$, ANZ1
A4	13	SIC_6, SIC_5, IC_6
A5	12	DSZ_1, DSZ_5, ASZ_1
A6	9	DSZ_3, DSN_5
A7	9	DSN_3, DN^2N_1
A8	6	${}^5\chi^v{}_C$, ${}^5\chi^b{}_C$
A9	6	DSZ_2, ASZ_2
A10	5	SIC_1
A11	4	CIC_1
A12	4	${}^3\chi^v{}_C$
A13	4	SIC_3
A14	4	${}^5\chi_{Ch}$

descriptors include the Randić connectivity index, the Kier−Hall simple path connectivity indices, the Zagreb group indices, and many triplet indices having as the main diagonal column vector the atomic number Z or the total number N of vertices.

Cluster A2 with about $^1/_8$ of the total number of descriptors includes molecular connectivity indices of order higher than 5, the J indices, and two closely similar triplet indices. Cluster A3 contains mainly valence/bond-corrected molecular connectivity indices. The next cluster, A4, consists mainly of the information-based indices IC (information content), SIC (structural information content), and CIC (complementary information content) for the hydrogen-filled graphs of order higher than 2 for IC and higher than 3 for SIC and CIC. Cluster A5 is composed mainly of triplet indices having as main diagonal unit vectors either distance sums or total number N of vertices.

Each of the remaining clusters has less than 10 descriptors. Clusters A6 and A7 contain mostly triplet descriptors: A6 with the distance sum S and A7 with the order N of the hydrogen-depleted graph, as the main diagonal unit vector; cluster A7 also includes two simple path cluster molecular connectivity indices. Cluster A8 contains simple cluster- and bond/valence-corrected cluster connectivities of high order (4−6). Cluster A9 again consists exclusively of triplet indices, and they are based on summing squares of LOVIs based mainly on distance sum unit vectors on the main diagonal.

Cluster A10 includes three information-theoretic indices IC and SIC of low order (0 and 1) as well as two triplet indices having in common the two unit vectors (distance sum S for the main diagonal, vertex degree V for the free term) and the operation for assembling LOVIs into an index (summation of LOVI square roots).

Interestingly, the four smallest clusters having four descriptors each are pairwise similar in type: A11 with A13, and A12 with A14. Cluster A11 consists of *information TIs* (IC, SIC, CIC) of low order (0−2), whereas A13 includes the same TIs of slightly higher order (2 and 3). Clusters A12 and A14 group together *molecular connectivity indices* based on simple cluster and simple cycle, respectively.

A general remark for the triplet indices is that what groups them together is not the matrix on which they are based (adjacency matrix or distance matrix) but the two unit vectors that convert such matrices into systems of linear equations.

Table 4. Summary of Variable Clustering for 1037 Diverse Chemicals

cluster	number of variables	representative variables (max. 25% of total listed)
B1	49	P_0, ANN_3, ANN_5, $AN1_3$, ANN_1, ANV_4, $AS1_4$, DN^21_4
B2	13	ANV_1, P_3, M_2
B3	13	$AS1_1$, $AS1_5$, $DS1_1$
B4	13	${}^6\chi$, ${}^6\chi^b$, P_7
B5	11	ASN_5, $AS1_3$, ASN_1
B6	10	SIC_3, SIC_4, CIC_4
B7	9	${}^5\chi^b{}_{PC}$, ${}^5\chi_{PC}$
B8	8	ASZ_2, ASZ_1
B9	6	${}^5\chi^b{}_C$, ${}^5\chi_C$
B10	6	${}^3\chi_{Ch}$, ${}^3\chi^b{}_{Ch}$
B11	6	IC_4, IC_5
B12	6	CIC_1, SIC_1
B13	6	${}^6\chi^v{}_{Ch}$, ${}^6\chi^b{}_{Ch}$
B14	6	${}^3\chi^b{}_C$, ${}^4\chi_C$
B15	4	J^B
B16	4	$AS1_2$
B17	4	DN^2N_1
B18	2	ANS_1

Clustering of Descriptors for the Diverse Set of Compounds. There are 18 variable clusters grouping together 176 variables for the database of 1037 diverse compounds (Table 4). Cluster B1, with 49 descriptors, includes 28% of all variables; 35 of these descriptors are common to cluster A1. Some of these indices, e.g., W (Wiener number), P_0 (number of non-hydrogen atoms), and P_1 (number of bonds in the hydrogen-depleted graph), express molecular size. It is interesting that most of the triplet variables (AZV_i, AZN_i, and ANN_i with $i = 1-5$ as well as several other ones) are found to be common to clusters A1 and B1. Five other descriptors (${}^0\chi^b$, ${}^2\chi^b$, ${}^3\chi^b$, ${}^0\chi^v$, and ${}^3\chi^v$) also appear in both clusters A1 and B1.

Cluster B2 has nine variables in common with cluster A1; most of these (${}^3\chi$, ${}^4\chi$, P_2-P_4) are path connectivities of intermediate order. A couple of triplet indices (ANV_1 and ANV_5) are also in common with cluster A1; another pair of triplet indices (ASN_3 and ASN_4) are in common with cluster A7.

Cluster B3 contains triplet indices with distance sums as main diagonal vector; they occur in clusters A5 and A9. In addition, two descriptors ($\bar{I}_D^W$ and H^D) appear also in cluster A1.

Cluster B4 is uniquely associated with cluster A2 and consists of indices ${}^5\chi$, ${}^6\chi$, ${}^5\chi^b$, ${}^6\chi^b$, ${}^5\chi^v$, ${}^6\chi^v$, and P_6-P_{10}. These descriptors are based on long paths; therefore, these variables appear only when large molecules are involved.

Seven of the eleven variables of cluster B5 form exclusively cluster A6; they are related to molecular shape via vertex complexity and graph radius. Five triplet indices such as ASN_1, ASN_5, DSN_1, DSN_5, and ANV_2 also are common to these two clusters.

Very interesting correspondences are manifested by cluster B6, which is mainly associated with two clusters involving the hydrocarbon database, namely, A4 and A13 (plus one descriptor in B6 that appears in A10). All variables are of information-theoretic type. These higher-order variables (SIC_3-SIC_6 and CIC_3-CIC_6) are common to clusters B6 and A4 and represent a true measure of molecular complexity. The lower- and intermediate-order indices such as IC_1 or SIC_2 that appear in clusters B6 and A10 or B6 and A13,

respectively, provide information on lower-order complexity that may be more degenerate than that furnished by the higher-order information indices. One should stress here that information content indices form clusters that are separate from clusters with other descriptors, meaning that such variables convey unique information relative to structure and molecular complexity.

Cluster B7 consists only of path-cluster molecular connectivity descriptors that were included in clusters A3, A7, and A8 for the hydrocarbons.

Cluster B8 includes triplet indices, all of which have the atomic number *Z* for the free-term vector in the system of linear equations. Most of these descriptors appear in clusters A1, A5, and A9.

Cluster B9 consists of high-order connectivity-cluster terms all contained in cluster A8. For hydrocarbons, descriptors ${}^{6}\chi^{b}_{C}$ and ${}^{6}\chi^{v}_{C}$ are perfectly correlated with descriptor ${}^{6}\chi_{C}$; therefore, the former variables did not appear in the hydrocarbon cluster A8. For the diverse-compound database, such a correlation is not perfect because of differences in atom types.

An interesting observation concerns cluster B10: all six variables are absent from the hydrocarbon database because the database does not contain any three- or four-membered rings, unlike the diverse compound database. This is why indices ${}^{3/4}\chi_{Ch}$, ${}^{3/4}\chi^{b}_{Ch}$, and ${}^{3/4}\chi^{v}_{Ch}$ appear only in cluster B10.

Cluster B11 has all but one of its descriptors contained in cluster A4; these information content indices, IC_2-IC_6, measure a high degree of nonredundancy of topological neighborhoods.

Cluster B12 has four of its variables contained in cluster A11; these descriptors (SIC_0, CIC_0-CIC_2) express lower-order redundancy of topological neighborhoods. This is true of indices IC_0 and SIC_1 as well, which are present in cluster A10.

From cluster B13, the six descriptors (simple, bond- and valence-corrected chain molecular connectivity indices) are partitioned equally between clusters A2 and A14, according to the six- versus five-membered ring size, respectively; in the hydrocarbon database A, six-membered chain (or rings) predominate.

Cluster B14 is exclusively associated in a one-to-one relationship with cluster A12. The corresponding descriptors ${}^{3}\chi_{C}$ and ${}^{4}\chi_{C}$ as well as their bond- and valence-corrected counterparts represent connectivity indices on three- and four-vertex structural clusters. For the hydrocarbon database, we have again a case in which the two indices ${}^{4}\chi^{b}_{C}$ and ${}^{4}\chi^{v}_{C}$, perfectly correlated with ${}^{4}\chi_{C}$, do not appear explicitly in cluster A12.

Half of the variables (*J*-type indices) in cluster B15 are contained in cluster A2. These *J* indices again form a cluster apart from all other ones in the case of the diverse database, proving that when heteroatoms are taken into account, the information provided by such *J*-type indices is unique.

Clusters B16, B17, and B18 each have a small number of triplet-type descriptors; the three descriptors of cluster B17 are all contained in cluster A7.

Intercluster Correlations. From each cluster we select 15−25% of the descriptors according to the maximal value of the correlation coefficient with their own cluster. In most cases, the first selected descriptor also has the minimal value of the correlation with the next closest cluster, expressed by the $1 - r^2$ value. When more than one index is chosen from the same cluster, after the first one was selected as indicated above, the next one must also fulfill a third criterion, namely, a low intercorrelation with the previously selected indices of the same cluster.

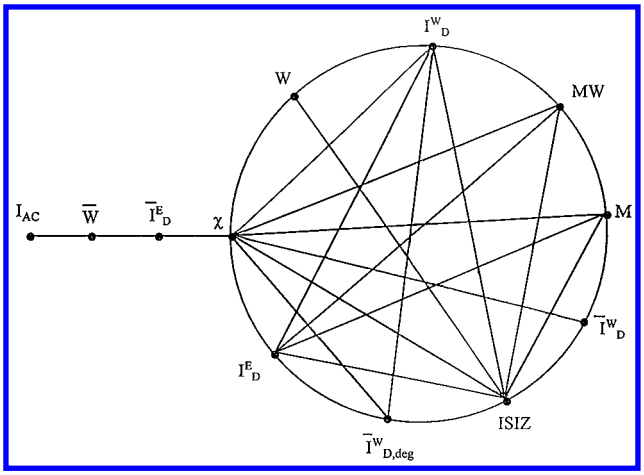

Figure 2. Graph of highly correlated topological indices (TIs) according to Todeschini et al. (notation of TIs as in Table 3 of ref 31). Lines connect TIs with $r > 0.90$.

There were four intercluster correlations within the hydrocarbon data set that were greater than 0.9, and all involved cluster A1. Cluster A1 was positively correlated with A2, A3, and A7. Cluster A1 was correlated negatively with A5. Each of the clusters characterizes some aspect of molecular size and shape.

Cluster B1 showed an intercluster correlation of 0.92 with cluster B2 and −0.90 with cluster B3. These were the only intercluster correlations greater than 0.9. These clusters are the three largest clusters in set B. Like cluster A1, cluster B1 groups TIs expressing molecular size and shape. Interestingly, in set A cluster A1 also had a negative intercluster correlation with cluster A5; it is therefore not surprising that clusters A5 and B3 have the most abundantly populated line connecting them in Figure 1.

In summary, for the hydrocarbon database there are four intercluster correlations with $r > 0.90$ all involving on one hand the first cluster A1 and on the other hand clusters A2, A3, A5, and A7. For the diverse compound database there are only two such intercluster correlations with $r > 0.90$, namely, B1 with B2 and B3. This is not unexpected, as the combination of the first three clusters in each case contains more descriptors than the parameters remaining in all the remaining ones together.

In this context, one should mention that Todeschini and co-workers published an interesting study[35] on 23 TIs for a set of 667 diverse chemicals, 20% of which were hydrocarbons; the above authors excluded 10 of these TIs because they were degenerate, or redundant or had intercorrelation factors higher than 0.90. A graph depicting highly intercorrelated indices using data published by these authors is presented in Figure 2, which is similar to a graph published earlier.[22]

Ten TIs were then selected by Todeschini et al.,[35] namely, the molecular weight (M_W), *J*, IC, CIC, the bonding information content (BIC), mean Randić connectivity (χ), the information content on atomic composition (I_{AC}), the mean Wiener index ($\bar{W}$), and the mean information indices on

equality of distance degree and on the magnitude of distance degree ($\bar{I}^{E}_{D,deg}$ and $\bar{I}^{W}_{D,deg}$, respectively). Then, using principal-component analysis for the above 10 TIs, Todeschini et al. analyzed the composition of the first six principal-components. They found that the first PC is mainly composed of indices that express the size of molecules (M_W, $\bar{W}$, IC, $\bar{I}^{E}_{D,deg}$ and $\bar{I}^{W}_{D,deg}$). This is in agreement with the earlier finding of Basak et al. for a set of 3692 diverse chemicals that the first PC is related to molecular size.[29] Further, Todeschini et al. found that the second PC is dominated by indices expressing information on bonds (IC, CIC, and BIC). This is also analogous to the results reported by Basak et al.[29] that the second axis represents molecular complexity as encoded by higher-order neighborhood complexity indices (IC_2, IC_3, SIC_2, SIC_3, CIC_2, CIC_3, etc.). The IC, CIC, and BIC indices used by Todeschini et al. are based solely on first-order topological bonding/neighborhoods and slightly different equivalence relations as compared to the IC_r, SIC_r, and CIC_r indices defined by Roy et al.[27] In studies by Basak et al.,[29] the first-order complexity indices (IC_1, SIC_1, CIC_1) were usually most correlated with the first PC. Each of the next four PCs in Todeschini et al.'s study[35] is dominated by a single TI,viz., χ, I_{AC}, J (indicating branching), and $\bar{I}^{E}_{D,deg}$ (connected with the position of substituents on the molecular scaffold), respectively.

ACKNOWLEDGMENT

This is contribution number 251 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported by Grants F49620-94-1-0401 and F49620-96-1-0330 from the U.S. Air Force.

REFERENCES AND NOTES

(1) Devillers, J., Balaban, A. T., Eds. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: The Netherlands, 1999. (a) Balaban, A. T.; Ivanciuc, O. Historical Development of Topological Indices. Chapter 2. (b) Kier, L. H.; Hall, L. B. Molecular Connectivity Chi Indices for Database Analysis and Structure−Property Modeling. Chapter 7. (c) Kier, L. H.; Hall, L. B. The Kappa Indices for Molecular Modeling of Molecular Shape and Flexibility. Chapter 10. (d) Kier, L. H.; Hall, L. B. The Electrotopological State: Structure Modeling for QSAR and Database Analysis. Chapter 11. (e) Basak, S. C. Information-Theoretic Indices of Neighborhood Complexity and Their Application. Chapter 12. (f) 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. Chapter 14.

(2) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(3) (a) Balaban, A. T. Chemical Graphs. Part 35. Five New Topological Indices for the Branching of Tree-Like Graphs. *Theor. Chim. Acta* **1979**, *5*, 239−261. (b) Bonchev, D.; Balaban, A. T.; Mekenyan, O. Generalization of the Graph Centre Concept and Derived Topological Indices. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 196−213. (c) Balaban, A. T.; Bertelsen, S.; Basak, S. C. New Centric Topological Indexes for Acyclic Molecules (Trees) and Substituents (Rooted Trees), and Coding of Rooted Trees, *Math. Chem. (MATCH)* **1994**, *30*, 55−72.

(4) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

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

(6) (a) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *80*, 399−404. (b) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206. (c) Balaban, A. T. Chemical Graphs. 48. Topological Index J for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties. *Math. Chem. (MATCH)* **1986**, *21*, 115−122. (d) Balaban, A. T.; Filip, P. Computer Program for Topological Index J (Average Distance Sum Connectivity). *Math. Chem. (MATCH)* **1984**, *16*, 163−190.

(7) Bonchev, D. *Information-Theoretic Indices for Characterization of Chemical Structure*; Research Studies Press: Letchworth, 1993.

(8) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992; pp 225−273.

(9) Balaban, A. T. Using Real Numbers as Vertex Invariants for Third-Generation Topological Indices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 23−28.

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

(11) Gutman, I.; Ruscic, B.; Trinajstić, N.; Wilcox, C. F., Jr. Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62*, 3399−3405.

(12) Shannon, C. A Mathematical Theory of Communication. *Bell Syst. Technol. J.* **1948**, *27*, 379−423.

(13) Balaban, A. T.; Balaban, T. S. New Vertex Invariants and Topological Indices of Chemical Graphs Based on Information on Distances. *J. Math. Chem.* **1991**, *8*, 383−397.

(14) 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 Correlational Ability. *J. Math. Chem.* **1987**, *1*, 61−83.

(15) (a) Ivanciuc, O.; Balaban, T. S.; Balaban, A. T. Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. *J. Math. Chem.* **1993**, *12*, 309−318. (b) Plavsic, D.; Nikolić, S.; Trinajstić, N.; Mihalic, Z. On the Harary Index for Characterization of Chemical Graphs. *J. Math. Chem.* **1993**, *12*, 235−250.

(16) (a) Balaban, A. T.; Diudea, M. V. Real Number Vertex Invariants: Regressive Distance Sums and Related Topological Indices. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 421−428. (b) Diudea, M. V.; Minailiuc, O.; Balaban, A. T. Molecular Topology. Part 4. Regessive Vertex Degrees (New Graph Invariants) and Derived Topological Indices. *J. Comput. Chem.* **1991**, *12*, 527−535.

(17) Diudea, M. V.; Minailiuc, O.; Katona, G.; Gutman, I. Szeged Matrices and Related Numbers. *Math. Chem. (MATCH)* **1997**, *35*, 129−143.

(18) Klein, D. J.; Randić, M. Resistance Distance. *J. Math. Chem.* **1993**, *17*, 147−154.

(19) (a) Randić, M.; Basak, S. C. Optimal Molecular Descriptors Based on Weighted Path Numbers. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261−266. (b) Basak, S. C.; Gute, B. D. Characterization of Molecular Structures Using Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 1−21. (c) Gute, B. D.; Basak, S. C. Predicting Acute Toxicity of Benzene Derivatives Using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117−131. (d) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243−272. (e) Basak, S. C.; Grunwald, G. D. Use of Graph Invariants, Volume and Total Surface Area in Predicting Boiling Point of Alkanes. *Math. Modell. Sci. Comput.* **1993**, *2*, 735−740. (f) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Math. Modell. Sci. Comput.*, in press. (g) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Quantitative Comparison of Five Molecular Structure Spaces in Selecting Analogs of Chemicals. *Math. Modell. Comput. Sci.*, in press. (h) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math. Modell. Sci. Comput.*, in press. (i) 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. (j) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of Graph Theoretic Parameters in Risk Assessment of Chemicals. *Toxicol. Lett.* **1995**, *79*, 239−250. (k) Basak, S. C.; Gute, B. D. Use of Graph Theoretic Parameters in Predicting Inhibition of Microsomal Hydroxylation of Anilines by Alcohols: A Molecular Similarity Approach. In *Proceedings of the International Congress on Hazardous Waste: Impact on Human and Ecological Health*; Johnson, B. L., Xintaras, C., Andrews, J. S., Eds.; Princeton Scientific Publishing Co. Inc.: Princeton, NJ, 1997; pp 492−504. (l) Basak, S. C.; Grunwald, G. D. Predicting Genotoxicity of Chemicals Using Nonempirical Parameters. In *Proceeding of XVI International Cancer Congress*; R. S. Rao, M. G. Deo, L. D. Sanghvi, Eds.; Monduzzi Editore S.p.A.: Bologna, Italy, 1995; pp 413−416. (m) Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A Comparative Study of Molecular Similarity, Statistical and Neural Network Methods for Predicting Toxic Modes of Action of Chemicals, *Environ. Toxicol. Chem.* **1998**, *17*, 1056−1064. (n) Basak, S. C.; Veith, G. D.; Grunwald, G. D. Prediction of Octanol−Water Partition Coefficient (K_{ow}) Using

Algorithmically-Derived Variables G J. *Environ. Toxicol. Chem.* **1992**, *11*, 893−900. (o) Basak, S. C.; Gute, B. D.; Drewes, L. R. Predicting Blood-Brain Transport of Drugs: A Computational Approach. *Pharm. Res.* **1996**, *13*, 775−778. (p) 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. (q) Gute, B. D.; Grunwald, G. D.; Basak, S. C. Prediction of the Dermal Penetration of Polycyclic Aromatic Hydrocarbons (PAHs): A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1999**, *10*, 1−15.

(20) Ivanciuc, O.; Balaban, A. T. Investigation of Alkane Branching with Topological Indices. *Math. Chem. (MATCH)*, in press.

(21) Balaban, A. T., Ed. *From Chemical Topology to Three-Dimensional Geometry*; Plenum Press: New York, 1998.

(22) Motoc, I.; Balaban, A. T. Topological Indices: Intercorrelations, Physical Meaning, Correlational Ability. *Rev. Roum. Chim.* **1981**, *26*, 593−600. Motoc, I.; Balaban, A. T.; Mekenyan, O.; Bonchev, D. Topological Indices: Inter-Relations and Composition. *Math. Chem. (MATCH)* **1982**, *13*, 369−404.

(23) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological Indices for Structure−Activity Correlations. In *Steric Effects in Drug Design*; Charton, M., Motoc, I., Eds.; *Top. Curr. Chem.* **1983**, *114*, 21−55. Balaban, A. T.; Niculescu-Duvaz, I.;. Simon, Z. Topological Aspects in QSAR for Biologically-Active Molecules. *Acta Pharm. Jugosl.* 1987, *37*, 7−36. Voiculetz, N.; Balaban, A. T. Niculescu-Duvaz, I.; Simon, Z. *Modeling of Cancer Genesis and Prevention*; CRC Press: Boca Raton, FL, 1990.

(24) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Characterization of the Molecular Similarity of Chemicals Using Topological Indices. In *Advances in Molecular Similarity*, Vol. 2; R. Carbo-Dorca, P. G. Mezey, Eds.; JAI Press: Stanford, CT, 1998; pp 171−185.

(25) Needham, D. E.; Wei, I. C.; Seybold, P. G. Molecular Modelling of the Physical Properties of Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186−4194. Mekenyan, O.; Bonchev, D.; Trinajstić, N. Chemical Graph Theory: Modelling the Thermodynamic Properties of Molecules. *Int. J. Quantum Chem.* 1980, *18*, 369−380. Karcher, W. *Spectral Atlas of Polycyclic Aromatic Hydrocarbons*; Kluwer Academic Press: Dordrecht, 1988; Vol. 2. pp 16−19.

(26) Russom, C. L. *Assessment Tools for the Evaluation of Risk (ASTER)*, v. 1.0; U.S. Environmental Protection Agency, 1992.

(27) 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 Modelling in Science and Technology*; 4th International Conference, Zurich; Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York; 1983; pp 745−750.

(28) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605−609. Ray, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, J. J. A Quantitative Structure Activity Relationship Study of Tumor Inhibitory Triazenes Using Bonding Information Content and Lipophilicity. *ICRS Med. Sci.* **1982**, *10*, 933−934. Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis. A Quantitative Structure−Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneimitt.-Forsch. Drug Res.* **1983**, *33*, 501−503.

(29) Basak, S. C.; Magnuson, V. R. Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discr. Appl. Math.* 1988, *19*, 17−44.

(30) Balasubramanian, K.; Basak, S. C. Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 367−373.

(31) (a) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Math. Modell. Sci. Comput.*, in press. (b) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math. Modell. Sci. Comput.*, in press. (c) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Quantitative Comparison of Five Molecular Structure Spaces in Selecting Analogs of Chemicals. *Math. Modell. Sci. Comput.*, in press.

(32) Raychaudhury, C.; Ray, S. K.; Roy, A. B.; Ghosh, J. J.; Basak, S. C. Discrimination of Isomeric Structures Using Information Theoretic Indices. *J. Comput. Chem.* **1984**, *5*, 581−588.

(33) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY 2.3*: University of Minnesota, 1988.

(34) SAS Institute Inc. The VARCLUS Procedure. In *SAS/STAT User's Guide*, Version 6, fourth ed.; SAS Institute Inc.: Cary, NC, 1989; Vol. 2, 846 pp.

(35) Todeschini, R.; Cazar, R.; Collina, E. The Chemical Meaning of Topological Indices. *Chemometrics Intell. Lab. Syst.* 1992, *15*, 51−59.

CI990114Y