# Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-Activity Relationships

Subhash C. Basak,* Sharon Bertelsen, and Gregory D. Grunwald

Natural Resources Research Institute, University of Minnesota, Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811

Received May 18, 1993*

Two molecular similarity methods have been used to select nearest neighbors from four different sets of chemicals. One of the methods is based on the Euclidean distance of chemicals in the ten dimensional principal components space derived from 97 graph invariants. The second approach is based on the count of atom pairs common to a pair of molecules. Two probe chemicals were selected, and neighbors of each were determined by the two methods for the following four sets of molecules: (a) a combined set of octane and nonane isomers, (b) a relatively more diverse set of 382 chemicals, (c) a diverse set of 3692 chemicals, and (d) the STARLIST data base of log $P$ consisting of 4067 structures. The results show that the measures reflect an intuitive notion of chemical similarity.

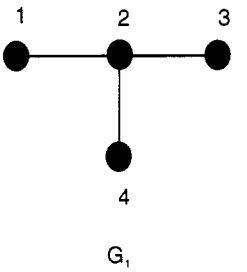
## 1. INTRODUCTION

During the past 2 decades there has been considerable progress in the application of algebraic graph theory in chemistry.[1-24] Molecular structures can be represented by planar graphs $G = [V,E]$, where the vertex set $V$ represents the atoms and the edge set $E$ represents the bonds.[13,25] The pattern of connectedness of atoms in a molecule, called molecular topology, is adequately depicted by chemical graphs. Therefore, it is not surprising that graph theoretical methods have been used in the characterization of moleuclar structure and prediction of properties.

Mathematical characterization of chemical structure can be accomplished by a matrix, a set of numbers, a polynomial, or a single numerical index.[13,25] For example, the distance matrix $\mathbf{D}(G)$, the adjacency matrix $\mathbf{A}(G)$, and the incidence matrix $\mathbf{T}(G)$ of a chemical graph $G$ uniquely determine molecular topology. These matrices, however, cannot be used directly for comparing molecular structures and the prediction of properties from chemical structure. For the purpose of structure-property relationships (SPR) it would be desirable to discover a graph property, preferably a single numerical characteristic or a set of numbers, which would uniquely characterize molecular topology. Unfortunately, in spite of many attempts, attainment of this goal has remained elusive.

Spialter[26-28] was the first to search for a graph invariant which could uniquely characterize the topology of molecular graphs. A graph invariant is a graph theoretical property which is preserved by isomorphism.[25] Spialter[26-28] asserted that the characteristic polynomial of the adjacency matrix $\mathbf{A}(G)$ of a molecular graph $G$ uniquely determines the topology of the molecule. This claim was, however, contradicted by later researchers who discovered nonisomorphic graphs with the same characteristic polynomial.[29-31] Randić[32] conjectured that the distance degree sequence (DDS) is sufficient for determining the isomorphism of tree graphs. Subsequently, it was found that nonisomorphic tree graphs have identical DDS.[33]

In spite of the non-uniqueness of graph theoretical invariants, such parameters (e.g., subgraphs and topological indices) have been employed in the ordering of structures, prediction of properties, and quantification of structural similarity/dissimilarity of molecules.[1-23] In fact, it has been pointed out by Randić[34] that the non-uniqueness of graph invariants is not a very serious handicap for SPR or SAR (structure-activity relationships). In alkanes, for example, properties like boiling point and octane number are not well correlated and lie in different numerical scales. A unique invariant, if discovered, could not simultaneously predict both of the above properties.




$G_1$

**Figure 1.** Labeled hydrogen-suppressed graph of isobutane.

Topological features of molecules have been used in predicting physicochemical and biological properties of molecules and in quantifying structural similarity/dissimilarity of chemical species.[1-12,15-20,35,36] Most studies in this area have been carried out by using relatively small sets of chemicals. In an earlier study, Basak *et al.*[35] developed a new method of quantifying structural similarity/dissimilarity of molecules using principal components (PCs) derived from a large and diverse set of 3692 chemicals. The similarity method developed by Carhart *et al.*[37] using atom pairs has also been used to characterize molecular structural similarity. In this paper, in an effort to understand the relative effectiveness of the PC based and atom pair based approaches, we have carried out a comparative sudy of the two methods in four different groups of chemicals: (a) a collection of octane and nonane isomers, (b) a relatively larger set of 382 chemicals, (c) a diverse set of 3692 structures, and (d) the STARLIST data base of log $P$ comprising chemicals. These are presented in this paper along with an analysis of the utility and limitations of these approaches.

## 2. QUANTIFICATION OF MOLECULAR SIMILARITY

**2.1. Topological Index Based Similarity.** The topological parameters used in this paper may be conveniently derived from the adjacency matrix $\mathbf{A}(G)$ or the distance matrix $\mathbf{D}(G)$ of a hemical graph $G$.

* Abstract published in *Advance ACS Abstracts,* February 15, 1994.

0095-2338/94/1634-0270$04.50/0 

The adjacency matrix $\mathbf{A}(G_1)$ and the distance matrix $\mathbf{D}(G_1)$ of the labeled graph $G_1$ (Figure 1) of isobutane are given below:

$$\mathbf{A}(G_1) = \begin{bmatrix} & (1) & (2) & (3) & (4) \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 1 \\ 3 & 0 & 1 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{D}(G_1) = \begin{bmatrix} & (1) & (2) & (3) & (4) \\ 1 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 1 & 1 \\ 3 & 2 & 1 & 0 & 2 \\ 4 & 2 & 1 & 2 & 0 \end{bmatrix}$$

From the adjacency matrix of a graph with $n$ vertices it is possible to calculate $\delta_i$, the degree of the $i$th vertex, as the sum of all entries in the $i$th row:

$$\delta_i = \sum_{j=1}^{n} a_{ij} \tag{1}$$

Zero-order connectivity index ${}^0\chi$ is defined as[12]

$${}^0\chi = \sum_i (\delta_i)^{-1/2} \tag{2}$$

*Randić's connectivity index* ${}^1\chi$ is defined as[4]

$${}^1\chi = \sum_{\text{all edges}} (\delta_i \delta_j)^{-1/2} \tag{3}$$

A generalized connectivity index ${}^h\chi$ considering paths of the type $v_0, v_1, \ldots, v_h$ of length $h$ in the molecular graph is calculated as[12]

$${}^h\chi = \sum (\delta_{v_0} \delta_{v_1} \ldots \delta_{v_h})^{-1/2} \tag{4}$$

where the summation is taken over all paths of length $h$.

Cluster, path–cluster, and cyclic types of simple connectivity indices are calculated using the method of Kier and Hall.[12]

Valence connectivity indices are based on vertex-weighted graphs where the weight, $\delta_i^v$, of the $i$th vertex is calculated as follows:[12]

$$\delta_i^v = (Z_i^v - h_i)/(Z_i - Z_i^v - 1) \tag{5}$$

where $Z_i^v$ is the number of valence electrons and $Z_i$ is the atomic number of the atom represented by the $i$th vertex of the chemical graph and $h_i$ is the number of hydrogen atoms attached to it. Valence connectivity indices, ${}^h\chi^v$, are calculated by replacing $\delta_i$ in eqs 2–4 with $\delta_i^v$.

The $K_h$ ($h$ = 0, 1, ..., 10) parameters used in this paper represent the number of occurrences of paths of length $h$ in the hydrogen-depleted molecular graph $G$. $K_0$ is the number of vertices and $K_1$ is the number of edges of $G$. Higher-order $K_h$ terms can be calculated using graph theoretical algorithms.

$W$ is calculated as

$$W = {}^1/_2 \sum_{i,j} d_{ij} = \sum_h h g_h \tag{6}$$

where $g_h$ is the number of unordered pairs of vertices whose distance is $h$.

Molecular complexity indices comprise another set of descriptors from molecular graphs.[9,11] The science of information theory has grown mainly out of the pioneering studies of Shannon,[38] Wiener,[39] Ashby,[40] and Kolmogorov.[41] There is more than one version of information theory. In Shannon's[38] statistical information theory, information is measured as reduced uncertainty of the system. In the algorithmic theory of Kolmogorov,[41] the quantity of information is defined as the minimal length of a program which allows a one-to-one transformation of an object (set) into another. In applying an information theoretical formalism on chemical graphs, one looks upon the information content (or complexity) of a graph as a measure of its degree of variety or heterogeneity as suggested by Ashby.[40] An appropriate set A of $n$ elements is derived from a molecular graph $G$, depending on certain preselected criteria. On the basis of an equivalence relation defined on $A$, the set A is partitioned into equivalence classes $A_i$ of order $n_i$ ($i$ = 1, 2, ..., $h$ $\sum_i n_i = n$). A probability scheme is then assigned to the set of equivalence classes:

$$A_1, A_2, \ldots, A_h$$

$$p_1, p_2, \ldots, p_h$$

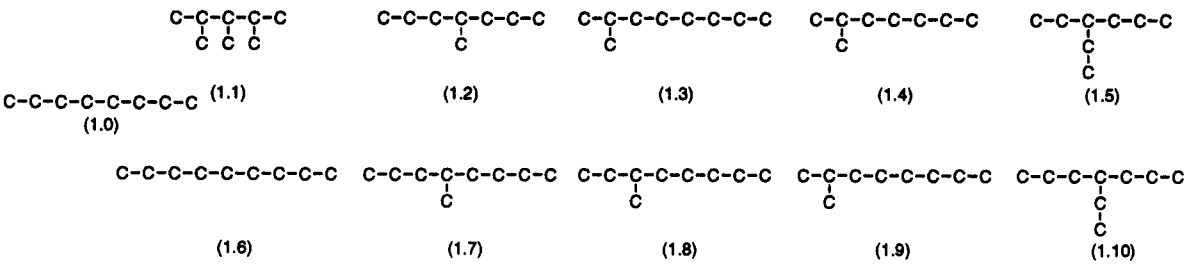
where $p_i = n_i/n$, $n_i$ and $n$ being the cardinalities of $A_i$ and $A$, respectively. The mean information content (or complexity) of an element of A is defined by Shannon's[38] relation:

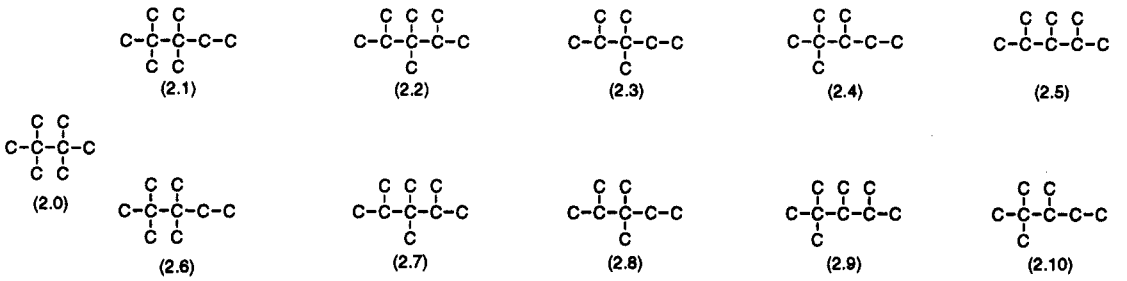
$$\text{IC} = -\sum_i p_i \log_2 p_i \tag{7}$$

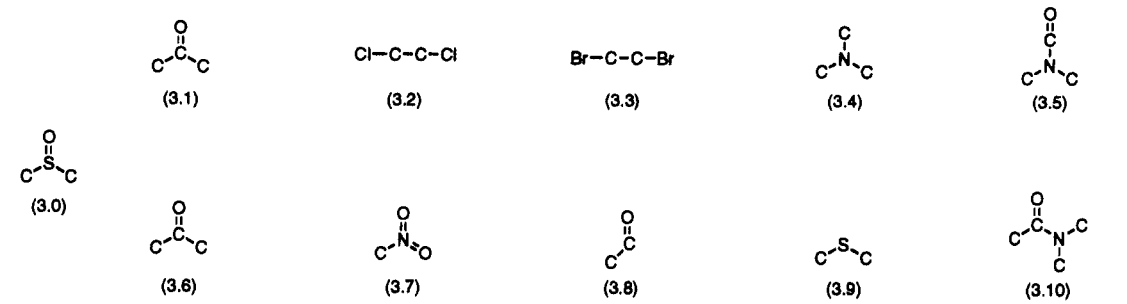
The logarithm is taken at base 2 for measuring the information content in bits. The total complexity of set A is then $n$ times IC.

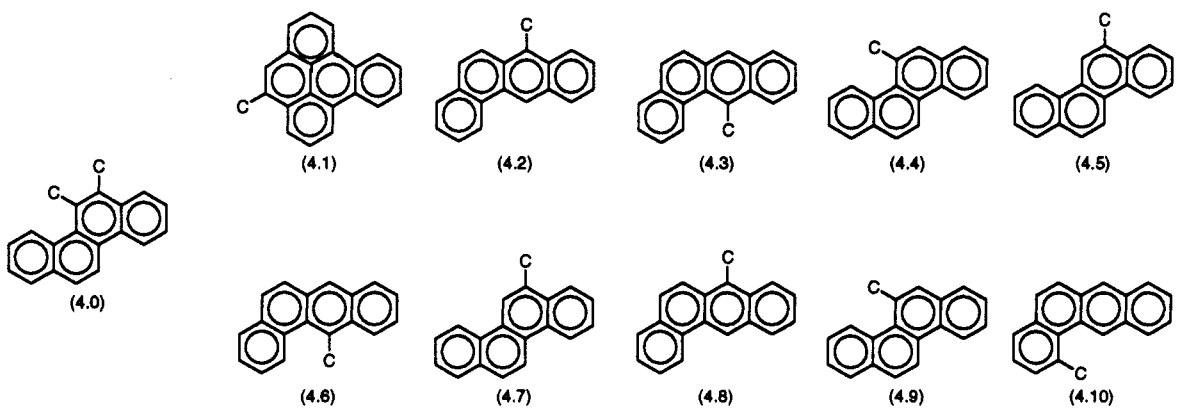
It is to be noted that the complexity of a real object of a model object is not uniquely defined. Complexity of an object would vary depending on the nature of the equivalence relation. For example, when $A$ represents the vertex set of a chemical graph $G$, two methods of partitioning have been widely used: (a) chromatic number coloring of $G$ where two vertices of the same color are considered equivalent and (b) determination of the transitive sets or orbits of the automorphism group of $G$ whereafter vertices are considered equivalent if they belong to the same orbit.

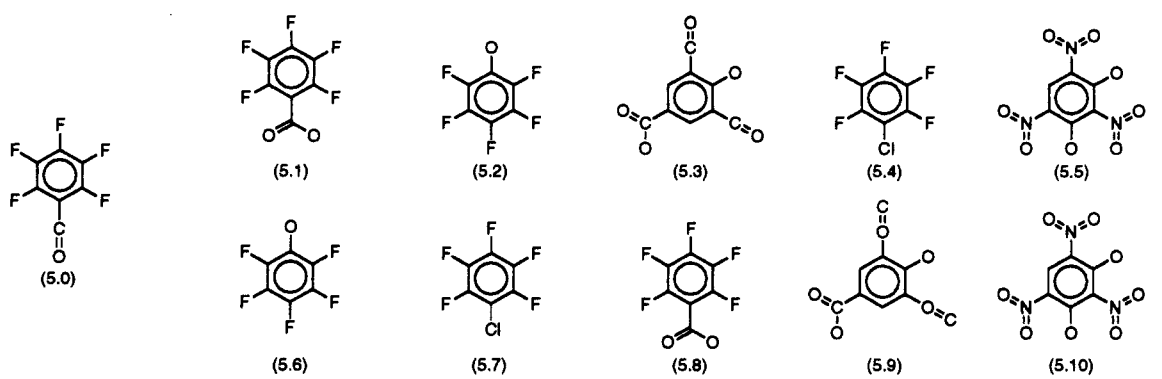
Rashevsky[42] represented molecules by simple linear graphs and calculated molecular complexity. In this approach, two vertices $u$ and $v$ of a graph $G$ are said to be topologically equivalent if and only if for each neighboring vertex $u_i$ ($i$ = 1, 2, ..., $k$) of the vertex $u$ there is a distinct neighboring vertex $v_i$ of the same degree for the vertex $v$. Subsequently, various authors have computed the complexity of molecules where linear graphs or multigraphs with indistinguishable vertices were used to symbolize the chemical species. On the other hand, to account for the unique nature of atoms and their bonding pattern in a molecule, Basak *et al.*,[43] Sarkar *et al.*,[44] and Roy *et al.*[45] calculated the complexity of graphs on the basis of equivalence relations where both the nature of the atom (vertex) and the number and chemical nature of bonded neighbors of all atoms are taken into account. This was accomplished by defining open spheres for all vertices of the molecular graph.[46] If $r$ is any nonnegative real number and $v$ is a vertex of the graph $G$, then the open $r$-sphere $S(v,r)$ is defined as the subset of $V(G)$ consisting of all vertices $v_i$ such that $d(v,v_i) < r$. Obviously, $V(v,0) = \phi$, $S(v,r) = v$ for $0 < r < 1$, and $S(v,r) = (v) \cup \Gamma^1(v) = \mathrm{N}^1(v)$ for $0 < r < 2$. One can construct open $r$-spheres of each vertex of $G$ for all integral values of $r$, $0 \le r \le \rho$. For a particular value of $r$ the collection of all such open spheres $S(v,r)$, where $v$ runs over the entire vertex set $V$, forms a neighborhood system of the vertices of $G$. A suitably defined equivalence relation can then partition $V$ into disjoint subsets based on the equivalence of nature, connectedness, and bonding pattern of neighbors up to $r$th-


(1.0)
(1.1)
(1.2)
(1.3)
(1.4)
(1.5)
(1.6)
(1.7)
(1.8)
(1.9)
(1.10)


(2.0)
(2.1)
(2.2)
(2.3)
(2.4)
(2.5)
(2.6)
(2.7)
(2.8)
(2.9)
(2.10)


(3.0)
(3.1)
(3.2)
(3.3)
(3.4)
(3.5)
(3.6)
(3.7)
(3.8)
(3.9)
(3.10)


(4.0)
(4.1)
(4.2)
(4.3)
(4.4)
(4.5)
(4.6)
(4.7)
(4.8)
(4.9)
(4.10)


(5.0)
(5.1)
(5.2)
(5.3)
(5.4)
(5.5)
(5.6)
(5.7)
(5.8)
(5.9)
(5.10)

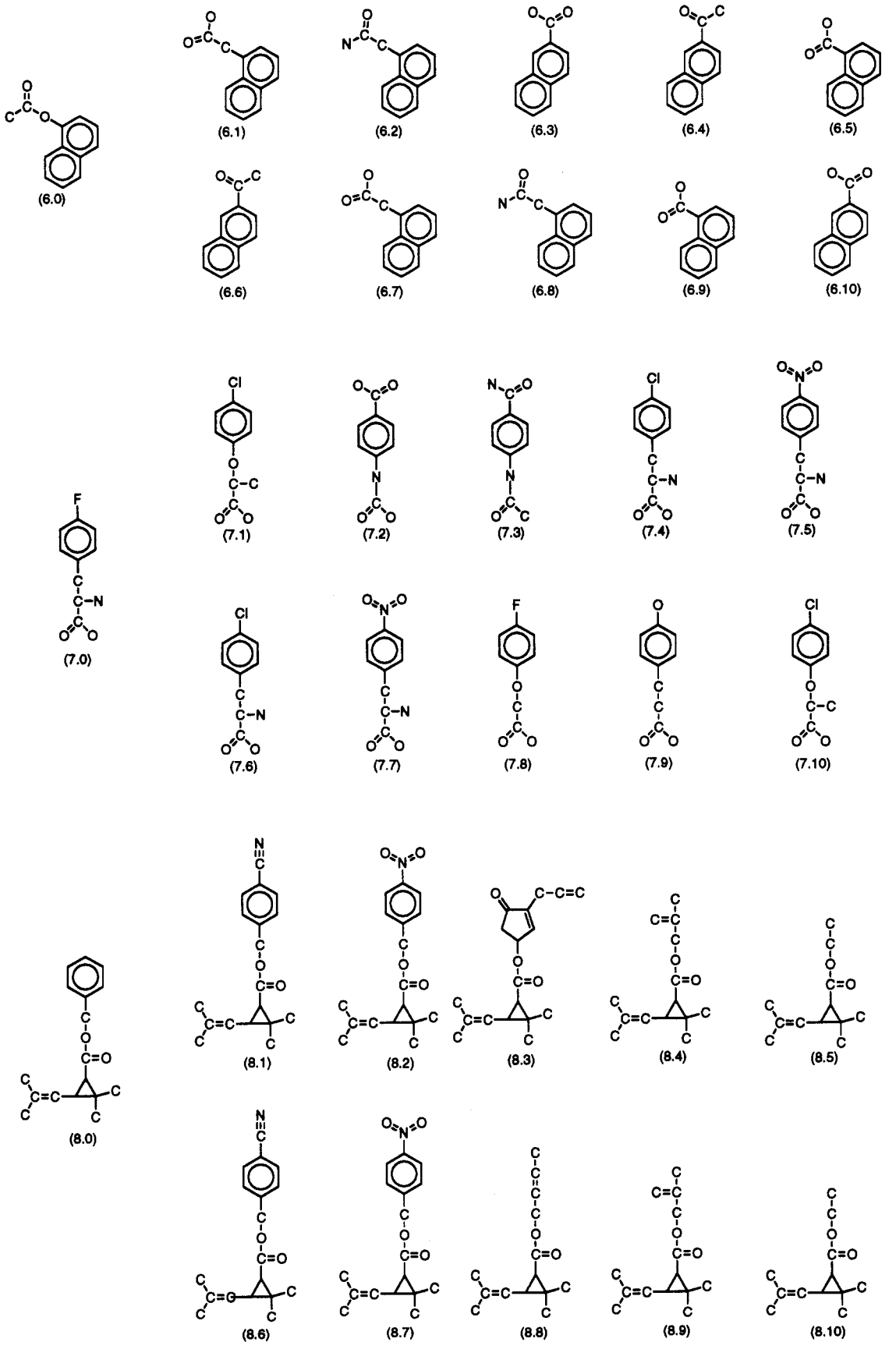



**Figure 2.** Probe compounds and their selected neighbors.

**Table I.** Topological Index Symbols and Definitions

| | |
|---|---|
| $I_D^W$ | information index for the magnitude 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 $h$ |
| $O$ | 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 |
| $O_{ORB}$ | maximum order of neighborhood of vertices for $I_{ORB}$ within the hydrogen-suppressed graph |
| $M_1$ | Zagreb group parameter = sum of square of degree over all vertices |
| $M_2$ | 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 $r$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 |
| $^h\chi$ | path connectivity index of order $h$ = 0.6 |
| $^h\chi_C$ | cluster connectivity index of order $h$ = 3.6 |
| $^h\chi_{CH}$ | chain connectivity index of order $h$ = 3.6 |
| $^h\chi_{PC}$ | path–cluster connectivity index of order $h$ = 4.6 |
| $^h\chi^b$ | bonding path connectivity index of order $h$ = 0.6 |
| $^h\chi_C^b$ | bonding cluster connectivity index of order $h$ = 3.6 |
| $^h\chi_{CH}^b$ | bonding chain connectivity index of order $h$ = 3.6 |
| $^h\chi_{PC}^b$ | bonding path–cluster connectivity index of order $h$ = 4.6 |
| $^h\chi^v$ | valence path connectivity index of order $h$ = 0.6 |
| $^h\chi_C^v$ | valence cluster connectivity index of order $h$ = 3.6 |
| $^h\chi_{CH}^v$ | valence chain connectivity index of order $h$ = 3.6 |
| $^h\chi_{PC}^v$ | valence path–cluster connectivity index of order $h$ = 4.6 |
| $P_h$ | number of paths of length $h$ = 0.10 |

order neighborhoods.[45] It is noteworthy that this approach incorporates the effects of distant neighbors (i.e., neighbors of immediately bonded neighbors) on an atom or a reaction center. After partitioning of the vertices for a particular order ($r$) of neighborhood, $IC_r$ is calculated by eq 7. Subsequently, Basak, Roy, and Ghosh[43] defined another information theoretical measure, structural information content ($SIC_r$), which is calculated as

$$SIC_r - IC_r/\log_2 n \tag{8}$$

where $IC_r$ is calculated by eq 7 and $n$ is the total number of vertices of the graph. It is noted that $SIC_r$ is related to Brillouin's[47] measure of redundancy of a system. Another information theoretical invariant, complementary information content (CIC), was defined as[48]

$$CIC_r = \log_2 n - IC_r \tag{9}$$

The Wiener index $W$,[49] and the information theoretical indices $I_D^W$ and $\bar{I}_D^W$ are calculated from the distance matrix of chemical graphs.[50] The set of topological indices used in this paper are shown in Table I. Topological parameters were calculated by the computer program POLLY[51] where SMILES line notation[52] is the input. The structural similarity of two chemicals X1 and X2 is measured in terms of their distance in the $n$-dimensional PC space following the method of Basak *et al.*,[35] where $n$ is the number of PCs with eigenvalues greater than or equal to 1.

**2.2. Atom Pair Based Similarity.** An atom pair is a way of describing substructural features of a molecule first described by Carhart *et al.*[37] An atom pair is defined as a substructure composed of two non-hydrogen atoms, $i$ and $j$, and their interatomic separation:

$$\langle \text{atom descriptor}_i \rangle - \langle \text{separation} \rangle - \langle \text{atom descriptor}_j \rangle$$

Interatomic separation of two atoms is the number of atoms traversed in the shortest bond-by-bond path containing both atoms. An atom descriptor identifies an atom type, number of non-hydrogen neighbors, and number of bonding $\pi$-electrons.

The atom pairs defined above may then be used to quantify similarity. One measure of similarity is a measure of the number of shared atom pairs between two compounds relative to the total number of atom pairs present within both compounds:

$$S_{ij} = A(i,j)/(T(i) + T(j))$$

where $A(i,j)$ is the number of atom pairs common to compounds $i$ and $j$, $T(i)$ is the total number of atom pairs in compound $i$, and $T(j)$ is the total number of atom pairs in compound $j$.

## 3. RESULTS

**3.1. Alkanes.** For the alkane series we have taken the combined set of octanes and nonanes for analysis. We selected $n$-octane and 2,2,3,3,-tetramethylbutane as the two probe compounds. Five nearest neighbors of each compound were determined by using the Euclidean distance and the atom pair approach.

In Figure 2, structure 1.0 is the probe compound ($n$-octane), structures 1.1–1.5 are the nearest neighbors chosen by the Euclidean method, and structures 1.6–1.10 are the closest neighbors selected by the atom pair method. Euclidean distance and atom pair similarity values are given in Table II. Table III gives the number of principal components (PCs) used in determining Euclidean distance. Table III lists the total variance explained by these components as well. It appears from the data that the Euclidean method has a tendency to maintain the number of vertices in the selection of neighbors whereas the atom pair approach is trying to select neighbors with very similar distance degree sequence.

On the other hand, for structure 2.0 (2,2,3,3-tetramethylbutane) the first three neighbors are identical for both methods. The fourth and fifth chosen neighbors of the two methods reveal that the Euclidean method emphasized molecular size more heavily than the atom pair approach in the selection of analogs.

**3.2. Diverse Data Sets.** Target compounds 3.0 (dimethyl sulfoxide) and 4.0 (5,6-dimethylchrysene) were taken from a set of 382 compounds. Target 3.0 is a small, simple structure with few unique features. The results for the atom pair method reflect an emphasis on maintaining the distance–degree sequence or branching pattern. The Euclidean method is more difficult to interpret, but, again, branching does not seem to be the sole criterion for selection, with molecular size being important as well.

The results for target 4.0 show a high degree of overlap in the two neighbor selection methods, with 4 of the 5 neighbors being the same by each method. The target compound is $C_{20}$ and all neighbors were $C_{19}$, except the first neighbor selected by the Euclidean method, which is $C_{21}$.

For target compounds 5.0 (pentafluoromethoxybenzene) and 6.0 (1-naphthalenol acetate), nearest neighbors were determined by the Euclidean distance method only and taken from a data base of 3692 compounds. The neighbors were then re-ordered by the atom pair method.

**Table II.** Eight Target Chemicals and Five Nearest Neighbors Determined by Atom Pair Similarity and Five Nearest Neighbors Determined by Euclidean Distance

| target chemical | no. | Euclidean distance | atom pair similarity | name | target chemical | no. | Euclidean distance | atom pair similarity | name |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.0 | | | *n*-octane | 5 | 5.0 | | | pentafluoromethoxybenzene |
| | 1.1 | 2.723 | | 2,3,4-trimethylpentane | | 5.1 | 0.468 | | pentafluorobenzoic acid |
| | 1.2 | 2.739 | | 4-methylheptane | | 5.2 | 0.486 | | pentafluorophenol |
| | 1.3 | 2.900 | | 2-methyloctane | | 5.3 | 0.984 | | 4-hydroxy-3,5-dimethoxybenzoic acid |
| | 1.4 | 2.969 | | 2-methylheptane | | 5.4 | 1.071 | | chloropentafluorobenzene |
| | 1.5 | 2.996 | | 3-ethylhexane | | 5.5 | 1.104 | | 2,4,6-trinitrobenzene-1,3-diol |
| | 1.6 | | 0.844 | *n*-nonane | | 5.6 | | 0.764 | pentafluorophenol |
| | 1.7 | | 0.719 | 4-methyloctane | | 5.7 | | 0.764 | chloropentafluorobenzene |
| | 1.8 | | 0.719 | 3-methyloctane | | 5.8 | | 0.651 | pentafluorobenzoic acid |
| | 1.9 | | 0.688 | 2-methyloctane | | 5.9 | | 0.260 | 4-hydroxy-3,5-dimethoxybenzoic acid |
| | 1.10 | | 0.594 | 4-ethylheptane | | 5.10 | | 0.093 | 2,4,6-trinitrobenzene-1,3-diol |
| 2 | 2.0 | | | 2,2,3,3-tetramethylbutane | 6 | 6.0 | | | 1-naphthalenol, acetate |
| | 2.1 | 4.561 | | 2,2,3,3-tetramethylpentane | | 6.1 | 0.378 | | 1-naphthaleneacetic acid |
| | 2.2 | 4.687 | | 2,3,3,4-tetramethylpentane | | 6.2 | 0.399 | | 1-naphthaleneacetamide |
| | 2.3 | 5.199 | | 2,3,3-trimethylpentane | | 6.3 | 0.440 | | 2-naphthalenecarboxylic acid |
| | 2.4 | 5.435 | | 2,2,3-trimethylpentane | | 6.4 | 0.484 | | 1-naphthalenecarboxylic acid |
| | 2.5 | 5.479 | | 2,3,4-trimethylpentane | | 6.5 | 0.527 | | 1-naphthalenecarboxylic acid |
| | 2.6 | | 0.750 | 2,2,3,3-tetramethylpentane | | 6.6 | | 0.769 | 1-(2-naphthalene)ethanone |
| | 2.7 | | 0.531 | 2,3,3,4-tetramethylpentane | | 6.7 | | 0.725 | 1-naphthaleneacetic acid |
| | 2.8 | | 0.464 | 2,3,3-trimethylpentane | | 6.8 | | 0.725 | 1-naphthalenecarboxylic acid |
| | 2.9 | | 0.406 | 2,2,3,4-tetramethylpentane | | 6.9 | | 0.686 | 1-naphthalenecarboxylic acid |
| | 2.10 | | 0.393 | 2,2,3-trimethylpentane | | 6.10 | | 0.663 | 2-naphthalenecarboxylic acid |
| 3 | 3.0 | | | dimethyl sulfoxide | 7 | 7.0 | | | *p*-fluorophenylalanine, *dl* |
| | 3.1 | 0.814 | | acetone | | 7.1 | 0.239 | | 2-(*p*-chlorophenoxy)propionic acid |
| | 3.2 | 1.239 | | 1,2-dichloroethane | | 7.2 | 0.261 | | *p*-carboxyacetanilide |
| | 3.3 | 1.271 | | 1,2-dibromoethane | | 7.3 | 0.379 | | *p*-*N*-acetylaminobenzamide |
| | 3.4 | 1.495 | | trimethylamine | | 7.4 | 0.467 | | *p*-chlorophenylalanine, *dl* |
| | 3.5 | 1.578 | | dimethylformamide | | 7.5 | 0.500 | | *p*-nitrophenylalanine |
| | 3.6 | | 0.500 | acetone | | 7.6 | | 0.846 | *p*-chlorophenylalanine, *dl* |
| | 3.7 | | 0.333 | nitromethane | | 7.7 | | 0.721 | *p*-nitrophenylalanine |
| | 3.8 | | 0.222 | acetaldehyde | | 7.8 | | 0.681 | *p*-fluorophenoxyacetic acid |
| | 3.9 | | 0.222 | dimethyl sulfide | | 7.9 | | 0.625 | *p*-hydroxy-*B*-phenylpropionic acid |
| | 3.10 | | 0.190 | dimethylacetamide | | 7.10 | | 0.577 | 2-(*p*-chlorophenoxy)propionic acid |
| 4 | 4.0 | | | 5,6-dimethylchrysene | 8 | 8.0 | | | benzyl (1*R*)-*trans*-chrysanthemate |
| | 4.1 | 0.851 | | 6-methylbenzo(*e*)pyrene | | 8.1 | 0.478 | | *p*-cyanobenzyl (1*R*)-*trans*-chrysanthemate |
| | 4.2 | 0.885 | | 7-methylbenz(*a*)anthracene | | 8.2 | 0.700 | | *m*-nitrobenzyl (1*R*)-*trans*-chrysanthemate |
| | 4.3 | 0.930 | | 12-methylbenz(*a*)anthracene | | 8.3 | 1.159 | | methylallyl (1*R*)-*trans*-chrysanthemate |
| | 4.4 | 1.034 | | 6-methylchrysene | | 8.4 | 1.278 | | ethyl (1*R*)-*trans*-chrysanthemate |
| | 4.5 | 1.080 | | 5-methylchrysene | | 8.5 | 1.362 | | ethyl (1*R*)-*trans*-chrysanthemate |
| | 4.6 | | 0.881 | 12-methylbenz(*a*)anthracene | | 8.6 | | 0.803 | *p*-cyanobenzyl (1*R*)-*trans*-chrysanthemate |
| | 4.7 | | 0.870 | 5-methylchrysene | | 8.7 | | 0.766 | *m*-nitrobenzyl (1*R*)-*trans*-chrysanthemate |
| | 4.8 | | 0.864 | 7-methylbenz(*a*)anthracene | | 8.8 | | 0.667 | crotyl (1*R*)-*trans*-chrysanthemate |
| | 4.9 | | 0.859 | 6-methylchrysene | | 8.9 | | 0.632 | methylallyl (1*R*)-*trans*-chrysanthemate |
| | 4.10 | | 0.820 | 1-methylbenz(*a*)anthracene | | 8.10 | | 0.538 | ethyl (1*R*)-*trans*-chrysanthemate |

**Table III.** Summary of Principal Component Analysis (PCA) for Four Sets of Data Used in Similarity Probes

| data set (probe ID's) | no. of PCs extracted[a] | total variance explained |
|---|---|---|
| (a) 53 $C_8$–$C_9$ alkanes (1.0 and 2.0) | 6 | 94.1 |
| (b) 382 diverse compounds (3.0 and 4.0) | 8 | 93.8 |
| (c) 3692 diverse compounds (5.0 and 6.0) | 10 | 92.6 |
| (d) 4067 STARLIST compounds (7.0 and 8.0) | 8 | 93.5 |

[a] Components with eigenvalues >1.0 were retained.

In the case of target 5.0, the Euclidean method selected some compounds with quite different substituents on the benzene ring, specifically 2,4,6-trinitrobenzene-1,3-diol and 4-hydroxy-3,5-dimethoxybenzoic acid. The corresponding similarity values for these compounds are quite low, 0.093 and 0.260, respectively. The atom pair similarity method indicates a strong bias in maintaining the branching pattern with the same substituents.

However, for target compound 6.0, although the ordering of compounds is slightly different for the two similarity methods, the similarity values for the atom pair method are quite high for all of the neighbor compounds.

Probe compounds 7.0 (*p*-fluorophenylalanine) and 8.0 (benzyl (1*R*)-*trans*-chrysanthemate) were used to search a data base of 4067 compounds. Both selection methods resulted in a high degree of overlap (4/5 neighbors in common) for both of the probe compounds.

## 4. DISCUSSION

The purpose of this paper was to compare the relative effectiveness of two graph theoretical methods, viz., PC based and atom pair based approaches, in the selection of similar structures from different sets of molecules with increasing diversity.

For the probe compound *n*-octane, the nearest neighbor selected by the PC based method is an octane isomer, viz., 2,3,4-trimethylpentane. On the other hand, the atom pair approach selected *n*-nonane as the closest analog of *n*-octane. This may be interpreted as the tendency of the PC method to weight more heavily on size (number of vertices) than on the pattern of branching in the molecular graph. In the case of dimethyl sulfoxide, the first chosen neighbors are chemically different from the probe molecule. This is because any similarity method is constrained by the presence of similar structures in the database from which the neighbors are being selected. In cases where the Euclidean distance is sufficiently low (or the atom pair based similarity scores are high), there is a substantial degree of structural similarity between the probe chemical and the selected neighbors in terms of molecular size, branching pattern, cyclicity, and aromaticity.

The term "structural similarity" is not uniquely defined. It is an intuitive concept developed by chemists to compare and classify molecules. Any measure of similarity is dependent upon three important factors. Firstly, similarity methods will depend on the way molecular structures are represented. Graph theoretical and geometric models of molecules, for example, will represent different aspects of molecular architecture and are expected to give different measures of similarity for the same pair of molecules. Secondly, similarity has to be measured in terms of certain characteristics of chemical structure. The selection of the set of descriptors will have an impact on the final outcome of the similarity measure. Thirdly, even if one begins with the same set of descriptors, the function used to map the set of descriptors into the set of real numbers will determine how far two molecules are from each other in a particular similarity scale.

In many practical situations in drug design and hazard assessment of chemicals, one has to compare molecules with chemicals in large data bases. To use empirical properties is often not practical in such situations. The graph theoretical approach described in this paper is a viable approach in such cases. In many instances one has to assess risk of chemicals and predict the toxic potential of molecules in the face of limited or unavailable test data. In the area of industrial chemicals, the National Research Council explored the availability of toxicity end points and concluded that many of these chemicals have been subjected to a minimum of testing or, in many cases, no testing at all. For example, risk assessment of chemicals present in the Toxic Substances Control Act (TSCA) Inventory as well as those submitted for Premanufacture Notification (PMN) reviews have very little test data.[53] When test data are available, they mostly consist of acute toxicity and skin and eye irritation test results. In an analysis of an environmental database of more than 30 000 chemicals, Basak *et al.*[54] found that the number of chemicals which had a measured value of either boiling point, melting point, or vapor pressure was 3692. Under these circumstances, a three tier strategy for the risk assessment of chemicals have been proposed: (1) critical evaluation of test data, (2) identification of potential analogs of a chemical, and (3) estimation of properties using quantitative structure–activity relationship (QSAR) models. In such cases the similarity methods studied in this paper may be used to select analogs of a candidate chemical and properties of the analogs may be used for hazard assessment. In the case of drug design, one has to screen a large number of chemicals and choice of strictures similar/dissimilar to a selected "lead" is very important. In recent years the Euclidean distance based approach developed by Basak *et al.*[35] has been successfully used in the identification of specific inhibitors of human immunodeficiency virus (HIV) through the analysis of large data bases.[55,56]

## ACKNOWLEDGMENT

The authors wish to acknowledge their appreciation to the U.S. Environmental Protection Agency (Cooperative Agreement CR 819621-01-0) for support. This paper is contribution no. 107 for the Center for Water and the Environment.

## REFERENCES AND NOTES

(1) Needham, D. E.; Wei, I. C.; Seybold, P. G. *J. Am. Chem. Soc.* **1988**, *110*, 4186.
(2) Rouvray, D. H.; Tatong, W. *Z. Naturforsch.* **1986**, *41a*, 1238.
(3) Gao, Y.; Hosoya, H. *Bull. Chem. Soc. Jpn.* **1988**, *61*, 3093.
(4) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(5) Mekenyan, O.; Dimitrov, S.; Bonchev, D. *Eur. Polym. J.* **1983**, *19*, 1185.
(6) Basak, S. C.; Niemi, G. J.; Veith, G. D. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Publishers: New York, 1990; p 235.
(7) Ray, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, J. J. *Ind. J. Chem.* **1981**, *20B*, 894.
(8) Randić, M. *J. Chem. Educ.* **1992**, *69*, 713.
(9) Magnuson, V. R.; Harriss, D. K.; Basak, S. C. in *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; p 178.
(10) Baker, R. J.; Acree, W. E.; Tsai, C. C. *Quant. Struct.-Act. Relat.* **1984**, *3*, 10.
(11) Mekenyan, O.; Bonchev, D.; Trinajstic, N. *Int. J. Quantum Chem.* **1980**, *18*, 369.
(12) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Letchworth, Hertfordshire, U.K., 1986.
(13) Trinajstic, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vols. I and II.
(14) Kennedy, J. W.; Quintas, L. V. *Applications of Graphs in Chemistry and Physics*; North-Holland: Amsterdam, 1988.
(15) Randić, M. *Int. J. Quantum Chem. Quantum Biol. Symp.* **1984**, *11*, 137.
(16) Sabljic, A.; Trinajstic, N. *Acta Pharm.* Yugosl. **1981**, *31*, 189.
(17) Basak, S. C.; Gieschen, D. P.; Harriss, D. K.; Magnuson, V. R. *J. Pharm. Sci.* **1983**, *72*, 934.
(18) Basak, S. C.; Monsrud, L. J.; Rosen, M. E.; Frane, C. M.; Magnuson, V. R. *Acta Pharm. Yugosl.* **1986**, *36*, 81.
(19) Basak, S. C. *Med. Sci. Res.* **1987**, *15*, 605.
(20) Rouvray, D. H. *J. Mol. Struct. (THEOCHEM)* **1989**, *185*, 187.
(21) Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517.
(22) Balasubramanian, K. *J. Math. Chem.* **1991**, *7*, 353.
(23) Herndon, W. C. *J. Org. Chem.* **1975**, *40*, 3583.
(24) Dias, J. R. *J. Chem. Educ.* **1989**, *66*, 1012.
(25) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA 1969.
(26) Spialter, L. *J. Am. Chem. Soc.* **1963**, *85*, 2012.
(27) Spialter, L. *J. Chem. Doc.* **1964**, *4*, 261.
(28) Spialter, L. *J. Chem. Doc.* **1964**, *4*, 269.
(29) Baker, G. A. *J. Math. Phys.* **1966**, *7*, 2238.
(30) Balaban, A. T.; Harary, F. *J. Chem. Doc.* **1971**, *11*, 258.
(31) Schwenk, A. J. In *New Directions in the Theory of Graphs*; ed. Haray, F., Ed.; Academic Press, New York, 1973.
(32) Randić, M. *Match* **1979**, *7*, 5.
(33) Slater, P. J. *J. Graph Theory* **1982**, *6*, 89.
(34) Randić, M. *J. Chem. Int. Comput. Sci.* **1984**, *24*, 164.
(35) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; *Discrete Appl. Math.* **1988**, *19*, 17.
(36) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(37) Carhart, R.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64.
(38) Shannon, C. E. *Bell Syst. Tech. J.* **1948**, *27*, 379.
(39) Wiener, N. *Cybernetics*; Wiley: New York, 1948.
(40) Ashby, W. *An Introduction to Cybernetics*; Wiley: New York, 1956.
(41) Kolmogorov, A. N. *Probl. Peredachi. Inf.* **1969**, *5*, 3.
(42) Rashevsky, N. *Bull. Math. Biophys.* **1955**, *17*, 229.
(43) Basak, S. C.; Roy, A. B.; Ghosh, J. J. In *Proceedings of the Second International Conference on Mathematical Modelling*; Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri—Rolla: Rolla, MO, 1980; Vol. II, p 851.
(44) Sarkar, R.; Roy, A. B.; Sarkar, P. K. *Math. Biosci.* **1978**, *39*, 299.
(45) Roy, A. B.; Basak, S. C.; Hariss, D. K.; Magnuson, V. R. In *Mathematical Modelling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York, 1984; p 745.
(46) Long, P. E. *An Introduction to General Topology*; Charles E. Merrill: Columbus, OH, 1971.
(47) Brillouin, L. *Science and Information Theory*; Academic Press: New York, 1956.
(48) Basak, S. C.; Magnuson, V. R. *Arzneim.-Forsch./Drug Res.* **1983**, *33*, 501.
(49) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(50) Bonchev, D.; Trinajstic, N. *J. Chem. Phys.* **1977**, *67*, 4517.
(51) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. POLLY: Copyright of the University of Minnesota, 1988.
(52) Anderson, E.; Veith, G. D.; Weininger, D. Report No. EPA/600/M-87-021; Environmental Research Laboratory: Duluth, MN, 1987.
(53) Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. *Environ. Health Perspect.* **1990**, *87*, 183.
(54) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. *Math. Model.* **1987**, *8*, 300.
(55) Lajiness, M. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova: New York, 1990; p 299.
(56) Romero, D. L.; Busso, M.; Tan, C. K.; Reusser, F.; Palmer, J. R.; Poppe, S. M.; Aristoff, P. A.; Downey, K. M.; So, A. G.; Resnick, L.; Terpley, W. G. *Proc. Natl. Acad. Sci.* **1991**, *88*, 8806.