Discrete Applied Mathematics 19 (1988) 17-44
North-Holland



# DETERMINING STRUCTURAL SIMILARITY OF CHEMICALS USING GRAPH-THEORETIC INDICES

S.C. BASAK and V.R. MAGNUSON
*Department of Chemistry, University of Minnesota, Duluth, MN 55812, USA*

G.J. NIEMI
*Natural Resources Research Institute, University of Minnesota, Duluth, MN 55812, USA*

R.R. REGAL
*Department of Mathematical Sciences, University of Minnesota, Duluth, MN 55812, USA*

Received 5 November 1985
Revised 2 April 1986

Ninety (90) graph-theoretic indices were calculated for a diverse set of 3692 chemicals to test the efficacy of using graph-theoretic indices in determining similarity of chemicals in a large, diverse data base of structures. Principal component analysis was used to reduce the 90-dimensional space to a 10-dimensional subspace which explains 93% of the variance. Distance between chemicals in this 10-dimensional space was used to measure similarity. To test this approach, ten chemicals were chosen at random from the set of 3692 chemicals and the five nearest neighbors for each of these ten target chemicals were determined. The results show that this measure of similarity reflects intuitive notions of chemical similarity.

## Introduction

During the last decade there has been an upsurge of interest in the applications of algebraic graph theory in chemistry [2, 18, 23, 24, 30, 33, 35, 43, 46, 50, 62, 75]. Molecular structures are, in essence, planar graphs where atoms are represented by vertices and covalent chemicals bonds are symbolized by edges [27]. Any pair of atoms in a molecule is involved in a binary relation: either the pair is bonded or not bonded [75]. This pattern of connectedness of atoms in a chemical structure, usually termed molecular topology, is adequately depicted by constitutional graphs. Therefore, it is not surprising that graph-theoretic formalism has been successfully used to illuminate different aspects of molecular structure and properties. To mention just a few, the graph-theoretical approach has found applications in chemical documentation [36, 59, 60], isomer discrimination and characterization of molecular branching [19, 51], enumeration of constitutional isomers associated with a particular empirical formula [2, 27], calculation of quantum chemical parameters [26, 50, 62], structure-physicochemical property correlations [32, 34, 57, 80], and chemical structure-biological activity relationships [7–11, 13, 14, 31, 32, 34, 41, 53–58, 63, 65, 80].

Chemists have long relied on visual perception in order to relate various aspects of constitutional graphs to observable chemical phenomena [42]. But a clear and quantitative understanding of the structural (topological) basis of chemistry necessitates the use of precise mathematical techniques. In recent years, applications of matrix theory, group theory, and information theory to chemical graphs have produced results which are important in chemistry [6, 11, 14–17, 19, 29, 31, 37–40, 44–46, 49, 51–53, 64–66, 75–77].

Mathematical characterization of a chemical graph (structure) may be accomplished by a matrix, a set of numbers or a single numerical index [75]. For example, the adjacency matrix $A(G)$, the distance matrix $D(G)$, and the incidence matrix $T(G)$ of a chemical graph $G$ uniquely determine molecular topology. Among the different matrices used for the representation of chemical structure the adjacency matrix (or connection table) has been most extensively used in chemistry [75]. However, connection tables pose a serious problem in chemical documentation because as the size of the graph increases, they require a disproportionately large number of operations for the testing of graph isomorphism. Specifically, $n^2 \times n!$ operations are required to determine whether two graphs $G_1$ and $G_2$ with $n$ vertices are isomorphic [61]. Another limitation of matrices is that they cannot be used as structural descriptors in the correlation or prediction of properties. Therefore, one of the cherished objectives in contemporary graph-theoretic research has been the discovery of a graph property, preferably a single numerical characteristic or a set of numbers derived from graphs, which would not only be easier to handle than the adjacency matrix itself but also uniquely related to molecular topology [61]. Unfortunately, in spite of numerous attempts, attainment of this goal has remained elusive.

In chronological order, Spialter [70–72] was the first to undertake a search for a graph invariant which could uniquely characterize molecular topology. A graph invariant is a graph-theoretic property which is preserved by isomorphism [27, 61]. Spialter [70–72] asserted that the characteristic polynomial of the adjacency matrix or atom connectivity matrix of a molecule is uniquely related to its topology. This notion was, however, contradicted by later researchers who found that nonisomorphic graphs may possess identical characteristic polynomials [1, 4, 28, 67]. These graphs are called isospectral or cospectral graphs [75]. Later, Randic [48] conjectured that for tree graphs collections of distance degree sequences (DDS) were sufficient to determine isomorphism. Subsequently, it was reported that neither DDS nor PDS (path degree sequence) could uniquely characterize the topology of tree graphs [43, 68]. More recently, Randic [47] developed a single numerical index, the molecular identification number, which was successful in the unique characterization of the topology of a relatively diverse set of structures including constitutional isomers and cyclic analogs. However, a counter-example, i.e. different chemical structures with the same identification number, has already been reported [47].

Under these circumstances, there are two distinct trends in chemical graph-theoretic index research: (a) the simultaneous use of more than one index, i.e., a

superindex to better characterize chemical structure as compared to a single numerical invariant [17], and (b) the development of indices with more discriminating power [3, 5]. Success of the latter approach is evident from the increasing popularity of different topological indices in structure-property relationship (SPR) and structure-activity relationshio (SAR) studies [7–11, 13, 14, 17, 31, 32, 34, 53–58, 63]. This is mainly due to the fact that graph-theoretic molecular descriptors (indices) can be calculated for any real or hypothetical molecular structure whereas physicochemical parameters used in SPR or SAR are not uniformly available [73]. It has also been pointed out by Randic [47] that the nonuniqueness of graph invariants is not a very serious handicap for SPR or SAR. For example, in the alkane series properties like boiling point and octane number are not mutually well correlated and lie in different numerical scales. Therefore, a unique topological index, if discovered, cannot simultaneously correlate with both properties. On the contrary, if a graph-theoretic index shows excellent correlation with a specific property of a congeneric series, then this relationship can help to elucidate the structural (topological) origin of that property. This is an interesting possibility in light of the concept of 'graph like state of matter' proposed by Gordon and Kennedy [23], who found that physico-chemical properties of molecules can be predicted by a general formula

$$P = \sum_i a_i T_i$$

where $P$ is a property, $a_i$ are coefficients determined empirically or calculated by combinatorial methods, and $T_i$ are the topological invariants. This LCGI approach holds for all properties [23] and is more general as compared to Smolenski's additivity function [62, 69].

Topological features of molecules have been used as independent variables in regression models [7–11, 17, 32] and as variables in multivariate pattern recognition models [21, 41, 81]. In regression models, topological features are correlated with physiochemical or biological properties. In multivariate models, topological features have been used to discriminate between given groups of chemicals or to cluster a set of structures into collections of similar structures. In most studies to date, the structures have been relatively homogeneous and often the number of topological features has been small.

In this study we calculated 90 graph-theoretic indices for a set of 3692 molecules and utilized these indices to determine structural similarity of chemicals. The 3692 structures were chosen from a larger Environmental Protection Agency data base of 25,000 industrial chemicals. Because this set of molecules is neither a collection of congeners nor a group of compounds designed for a particular purpose, this data base has a much wider range of structural variation than in previous studies. Investigating 90 indices for 3692 chemicals can be overwhelming. Also, the storage of 90 indices is particularly inefficient if the indices have highly interrelated information. To reduce the dimensionality of the problem, principal component analysis (PCA) is used to reduce the 90 dimensions to 10 uncorrelated linear dimensions which explain the significant parts of the variation between chemicals in the 90

dimensions. These principal components can, in principle, be used to define a distance or dissimilarity between compounds. A distance of zero would imply complete structural (topological) similarity in this principal component space.

To test the efficacy of this numerical definition of structural similarity, we randomly selected ten compounds from our data base and found the five nearest neighbors for each. These results are presented in this paper along with a critical analysis of the utility and limitations of the approach in selecting structural analogs.

## Definitions and basic concepts

A graph $G$ is defined as an ordered pair consisting of two sets $V$ and $R$,

$$G = [V, R]$$

where $V$ represents a finite nonempty set and $R$ is a binary relation defined on the set $V$. The elements of $V$ are called vertices and the elements of $R$, also symbolized by $E(G)$ or $E$, are called edges. Such an abstract graph is commonly visualized by representing elements of $V$ as points and by connecting each pair $x = (v_i, v_j)$ of elements of $V$ with a line if and only if $(v_i, v_j) \in R$. The vertex $v_i$ and line $x$ are incident with each other, as are $v_j$ and $x$. Two vertices in $G$ are called adjacent if they are connected by a line. A walk of a graph is a sequence beginning and ending with vertices in which vertices and edges alternate and each edge is incident with vertices immediately preceding and following it. A walk of the form $v_0, x_1, v_1, x_2, \ldots, v_n$ joins vertices $v_0$ and $v_n$. The length of a walk is the number of occurrences of edges (lines) in it. A walk is closed if $v_0 = v_n$, otherwise it is open. A closed walk with $n$ points is a cycle if all its points are distinct and $n \geq 3$. A path is an open walk in which all vertices are distinct. A graph $G$ is connected if every pair of its vertices is connected by a path. A graph $G$ is a multigraph if it contains more than one edge between at least one pair of adjacent vertices, otherwise $G$ is a linear graph. The distance $d(v_i, v_j)$ between vertices $v_i$ and $v_j$ in $G$ is the length of any shortest path connecting $v_i$ and $v_j$. The degree of a vertex $v_i$ ($\deg v_i$) in $G$ is equal to the number of lines incident with $v_i$. The eccentricity $e(u)$ of a vertex $u$ in $G$ is given by $e(u) = \max_{v \in V} d(u, v)$. The radius $\varrho$ of a graph is given by $\varrho = \min_{u \in V} e(u) = \min \max_{v \in V} d(u, v)$. For a vertex $v \in V$, the first-order neighborhood $\Gamma^1(v)$ is a subset of $V$ such that $\Gamma^1(v) = \{u \in V \mid d(u, v) = 1\}$. The first-order closed neighborhood $N^1(v)$ of $v$ is defined as $N^1(v) = (v) \cup \Gamma^1(v) = \Gamma^0(v) \cup \Gamma^1(v)$ where $(v)$ is the one-point set consisting of $v$ only and may be looked upon as $\Gamma^0(v)$. If $\varrho$ is the radius of a chemical graph $G$, one can construct $\Gamma^i(u)$ and $N^i(u)$, $i = 1, 2, \ldots, \varrho$, for each vertex $u$ in $G$. Two graphs $G_1$ and $G_2$ are said to be isomorphic ($G_1 \simeq G_2$) if there exists a one-to-one mapping of the vertex set of $G_1$ onto that of $G_2$ such that adjacency is preserved. Automorphism is the isomorphism of a graph $G$ with itself.

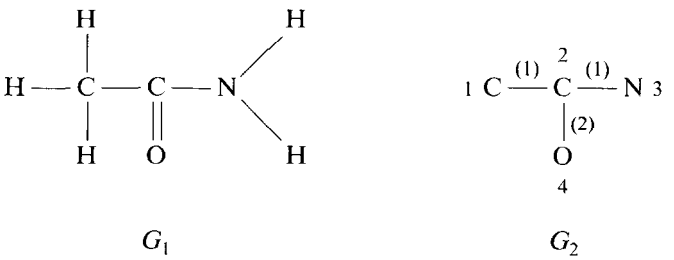
In depicting a molecule by a connected graph $G = [V, E]$, $V$ symbolizes the set of atoms and $E$ represents the set of covalent bonds between adjacent atoms. The set

$V$ may contain either all atoms present in the empirical formula [11, 14, 63] or only nonhydrogen atoms [32]. Hydrogen-filled graphs are preferable to hydrogen-suppressed graphs when hydrogen atoms are involved in critical steric or electronic interactions intramolecularly or intermolecularly or when hydrogen atoms have different physicochemical properties due to differences in bonding topology. In this paper, a single nondirected edge of a graph denotes a covalent bond which is given a weight equal to the bond order: 1 for a single bond, 2 for a double bond, etc. For an aromatic molecule, a bond joining two atoms involved in delocalization is depicted by a single undirected edge with a weight equal to 1.5. Under these conditions, the majority of stable chemical species can be represented by linear graphs. For such molecular graphs the binary relation $R$ depicting the topology of atoms (vertices) is symmetric and anti-reflexive, i.e., for any pair $(v_i, v_j)$ of vertices in $G$,

$$(v_i, v_j) \in R \quad \Leftrightarrow \quad (v_j, v_i) \in R,$$

$$(v_i, v_j) \in R \quad \Rightarrow \quad v_i \neq v_j.$$

The hydrogen-filled molecular graph, $(G_1)$, and labelled hydrogen-suppressed graph, $(G_2)$, for acetamide are shown below. The numbers in parenthesis in $G_2$ represent weights of the different edges.




$G_1$ $G_2$

The ninety topological parameters used in this paper for the calculation of principal components may be conveniently derived from the adjacency matrix $A(G)$, the atom connectivity matrix $A'(G)$ or the distance matrix $D(G)$ of a chemical graph $G$. These matrices are usually constructed from labelled graphs of hydrogen-depleted molecular skeletons. For such a graph $G$ with vertex set $\{v_1, v_2, \ldots, v_n\}$, $A(G)$ is defined to be the $n \times n$ matrix $(a_{ij})$, where $a_{ij}$ may have only two different values as follows:

$a_{ij} = 1$, if vertices $v_i$ and $v_j$ are adjacent in $G$,

$a_{ij} = 0$, otherwise.

Since we are considering graphs which are undirected and devoid of any self-loop, $A(G)$ is a symmetric (0, 1)-matrix in which each diagonal element is zero. It is to be noted that $A(G)$ fails to depict valence bond structures of molecules containing *pi* bonds.

The distance matrix $D(G)$ of a nondirected graph $G$ with $n$ vertices is a symmetric

$n \times n$ matrix $(d_{ij})$, where $d_{ij}$ is equal to the distance between vertices $v_i$ and $v_j$ in $G$. Each diagonal element $d_{ii}$ of $D(G)$ is equal to zero. Since topological distance in a graph is not related to the weight attached to an edge (bond), $D(G)$ does not adequately represent valence bond structures of molecules containing more than one covalent bond between adjacent atoms.

The atom connectivity matrix $A'(G)$ of an undirected chemical graph $G$ with $n$ vertices is a symmetric matrix $(a'_{ij})$, where $a'_{ij}$ is equal to the bond order of the covalent bond connecting atoms $i$ and $j$ [70–72]. All its diagonal elements $a'_{ii}$ are equal to zero. However, sometimes the diagonal of $A'(G)$ is also used to store the chemical identity of the vertex. For the labelled graph $G_2$, the four diagonal elements will be: $a'_{11} = C$, $a'_{22} = C$, $a'_{22} = C$, $a'_{33} = N$ and $a'_{44} = 0$. In principle, the off-diagonal elements $a'_{ij}$ $(i \neq j)$ of $A'(G)$ may be used to represent almost any type of bond, e.g., hydrogen bond or weak bonds present in the transition states of $SN_2$ reactions [71,72]. However, in this paper such bonds are not considered as edges of a graph. It is clear that $A'(G)$ adequately depicts the bonding pattern of a large number of molecules, both organic and inorganic.

The adjacency matrix $A(G_2)$, the atom connectivity matrix $A'(G_2)$, and the distance matrix $D(G_2)$ for the labelled graph $G_2$ may be written as follows:

$$A(G_2) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (2) & (3) & (4) \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

$$D(G_2) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (2) & (3) & (4) \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 1 & 2 & 0 \end{bmatrix} \end{array}$$

$$A'(G_2) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (2) & (3) & (4) \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{bmatrix} \end{array}$$

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 [32]:

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

*Randic's connectivity index* $^1\chi$ is defined as [51]:

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

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 [32]:

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

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 [32].

Bonding connectivity indices are based on the atom connectivity matrix of a chemical graph. From $A'(G)$ it is possible to calculate the valency of the $i$th atom (vertex) $\delta_i^b$ as the sum of all entries in the $i$th row:

$$\delta_i^b = \sum_{j=1}^{n} a'_{ij}. \tag{5}$$

Bonding connectivity indices, $^h\chi^b$, are calculated by replacing $\delta_i$ in eq. (4) with $\delta_i^b$.

Valence connectivity indices are based on vertex-weighted graphs where the weight, $\delta_i^v$, of the $i$th vertex is calculated as follows [32]:

$$\delta_i^v = Z_i^v - h_i \tag{6}$$

where $Z_i^v$ is the number of valence electrons 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 eq. (4) with $\delta_i^v$. It is to be noted, however, that in the case of certain atoms, e.g., chlorine, bromine, iodine, fluorine, sulfur etc., the $\delta^v$ values used are derived empirically through calibration with physicochemical properties [32]. The physical and/or graph-theoretic basis for these empirical adjustments remains far from clear.

The three types of connectivity indices described here present a three-tier approach to the quantification of molecular structure through the progressive integration of the concepts of structural chemistry with the topological aspects of graph theory: simple connectivity indices quantify the adjacency pattern of vertices (atoms) without any regard to their chemical properties or the nature of bonds (edges) connecting them, bonding connectivity indices integrate the nature of covalent bonds with the topology of atoms, and valence connectivity indices not only take care of electrons involved in the formation of covalent bonds but also incorporate lone pairs of electrons within the framework of graph theory.

The $K_h$ ($h = 0, 1, \ldots, 10$) parameters used in this paper represent the number of occurences 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-theoretic algorithms.

The Wiener [79] index $W$, the first topological index reported in the chemical literature, can be calculated from the distance matrix $D(G)$ of a hydrogen-suppressed graph $G$ as the sum of entries in the upper triangular distance submatrix [75]:

$$W = \tfrac{1}{2} \sum_{i,j} d_{ij} = \sum_{h} h \cdot g_h \tag{7}$$

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

Information-theoretic topological indices are calculated by the application of information theory to chemical graphs [7–11, 14, 31, 52, 63]. An appropriate set $A$ of $n$ elements is derived from a molecular graph $G$ depending upon certain structural characteristics. On the basis of an equivalence relation defined on $A$, the set $A$ is partitioned into disjoint subsets $A_i$ of order $n_i$ ($i = 1, 2, \ldots, h$; $\sum_i n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

$$\begin{pmatrix} A_1, A_2, \ldots, A_h \\ p_1, p_2, \ldots, p_h \end{pmatrix}$$

where $p_i = n_i/n$ is the probability that a randomly selected element of $A$ will occur in the $i$th subset.

The mean information content of an element of $A$ is defined by Shannon's relation [18]:

$$\mathrm{IC} = -\sum_{i=1}^{h} p_i \log_2 p_i. \tag{8}$$

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

It is to be noted that information content of a graph $G$ is not uniquely defined. It depends on the way the set $A$ is derived from $G$ as well as on the equivalence relation which partitions $A$ into disjoint subsets $A_i$. For example, when $A$ constitutes 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 orbits of the automorphism group of $G$ whereafter vertices belonging to the same orbit are considered equivalent.

Rashevsky [52] was the first to calculate information content of graphs where 'topologically equivalent' vertices are placed in the same equivalence class. In Rashevsky's approach, two vertices $u$ and $v$ of a graph are said to be topologically equivalent if and only if for each neighboring vertex $u_i$ ($i = 1, 2, \ldots, k$) of the vertex $u$ there is a distinct neighboring vertex $v_i$ of the same degree for the vertex $v$. Subsequently, Trucco [76, 77] defined topological information of graphs on the basis of graph orbits. In this method, vertices which belong to the same orbit of the automorphism group are considered topologically equivalent. While Rashevsky [52] used simple linear graphs with indistinguishable vertices to symbolize molecular structure, weighted linear graphs or multigraphs are better models for conjugated

or aromatic molecules because they more properly reflect the actual bonding patterns, i.e., electron distribution.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar, Roy and Sarkar [66] calculated information content of chemical graphs on the basis of an equivalence relation where two atoms of the same element are considered equivalent if they possess an identical first-order topological neighborhood. Since properties of atoms or reaction centers are often modulated by physicochemical characteristics of distant neighbors, i.e. neighbors of neighbors, it was deemed essential to extent this approach to account for higher-order neighbors of vertices. This can be accomplished by defining open spheres for all vertices of a chemical graph. If $r$ is any non-negative real number and $v$ is a vertex of the graph $G$, then the open sphere $S(v, r)$ is defined as the set consisting of all vertices $v_i$ in $G$ such that $d(v, v_i) < r$. Obviously, $S(v, 0) = \emptyset$, $S(v, r) = v$ for $0 < r < 1$, and $S(v, r)$ is the set consisting of $v$ and all vertices $v_i$ of $G$ situated at unit distance from $v$, if $1 < r < 2$.

One can construct such open spheres for higher integral values of $r$. For a particular value of $r$ the collection of all such open spheres $S(v, r)$, where $v$ runs over the whole vertex set $V$, forms a neighborhood system of the vertices of $G$. A suitably defined equivalence relation can then partition $V$ into disjoint subsets consisting of topological neighborhoods of vertices up to $r$th order neighbors. Such an approach has already been initiated and the information-theoretic indices calculated are called indices of neighborhood symmetry [63].

In this paper chemical species are symbolized by weighted linear graphs. Two vertices $u_0$ and $v_0$ of a molecular graph are said to be equivalent with respect to $r$th order neighborhood if and only if corresponding to each path $u_0, u_1, \ldots, u_r$ of length $r$ there is a distinct path $v_0, v_1, \ldots, v_r$ of the same length such that the paths have similar edge weights, and both $u_0$ and $v_0$ are connected to the same number and type of atoms up to the $r$th order bonded neighbors. The detailed equivalence relation is described in our earlier studies [34, 63].

Once partitioning of the vertex set for a particular order of neighborhood is completed, $IC_r$ is calculated by eq. (8). It is clear that the vertices of a graph belonging to the same equivalence class in terms of the above relation may be permuted without disturbing the relation already defined on the vertex set. Therefore, as pointed out by Mowshowitz [37–40], measures of molecular complexity give information content of structures in relation to a system of transformations leaving the structure invariant.

Subsequently, Basak, Roy and Ghosh [14] defined another information-theoretic measure, structural information content ($SIC_r$), which is calculated as:

$$SIC_r = IC_r / \log_2 n \tag{9}$$

where $IC_r$ is calculated from eq. (8) and $n$ is the total number of vertices of the graph. It is noted that $SIC_r$ is related to Brillouin's [20] measure of redundancy of a system.

Another information-theoretic invariant, complementary information content ($CIC_r$), is defined as [11]:

$$CIC_r = \log_2 n - IC_r. \tag{10}$$

$CIC_r$ represents the difference between maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by $IC_r$.

The information-theoretic index on graph distance, $I_D^W$ is calculated from the distance matrix $D(G)$ of a chemical graph $G$ as follows [19]:

$$I_D^W = W \log_2 W - \sum_h g_h \cdot h \log_2 h. \tag{11}$$

Table 1. Variable definition and symbols.

| | |
|---|---|
| $W$ | Half-sum of the off-diagonal elements of the distance matrix of a graph. |
| $I_D^W$ | 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. |
| $IC_r$ | Mean information content or complexity of a graph based on the $r$th ($r=0, 1, \ldots, 6$) order neighborhood of vertices in a graph. |
| $SIC_r$ | Structural information content of a graph based on $r$th ($r=0, 1, \ldots, 6$) order neighborhood of vertices. |
| $CIC_r$ | Complementary information content of a graph $G$ calculated from $r$th order ($r=0, 1, \ldots, 6$) neighborhood of vertices. |
| ${}^h\chi$ | Path terms of $h$th order ($h=0, 1, \ldots, 6$) calculated from values. |
| ${}^h\chi_C$ | Cluster terms of $h$th order ($h=3, \ldots, 6$) based on values. |
| ${}^h\chi_{PC}$ | Path-cluster terms of $h$th order ($h=4, \ldots, 6$). |
| ${}^h\chi_{CH}$ | Chain or cycle terms of different orders ($h=3, \ldots, 6$) based on values. |
| ${}^h\chi^b$ | Bonding connectivity type path terms of $h$th order ($h=0, 1, \ldots, 6$) calculated from b-values. |
| ${}^h\chi_C^b$ | Bonding connectivity type cluster terms of $h$th order ($h=3, \ldots, 6$) based on b-values. |
| ${}^h\chi_{PC}^b$ | Bonding connectivity type path-cluster terms of $h$th order ($h=4, \ldots, 6$). |
| ${}^h\chi_{CH}^b$ | Bonding connectivity type chain or cycle terms of $h$th order ($h=3, \ldots, 6$). |
| ${}^h\chi^v$ | Valence connectivity type path terms of $h$th order ($h=0, \ldots, 6$) calculated from v-values. |
| ${}^h\chi_C^v$ | Valence connectivity type cluster terms of $h$th order ($h=3, \ldots, 6$) based on v-values. |
| ${}^h\chi_{PC}^v$ | Valence connectivity type path-cluster terms of $h$th order ($h=4, \ldots, 6$). |
| ${}^h\chi_{CH}^v$ | Valence connectivity type chain or cycle terms of $h$th order ($h=3, \ldots, 6$). |
| $O$ | Order or neighborhood when $IC_r$ reaches its maximum value. |
| $K_h$ | Number of paths of length $h$ ($h=0, 1, \ldots, 10$) in the hydrogen-deleted graph. |

The mean information index, $\bar{I}_D^W$, is found by dividing the information index, $I_D^W$, by $W$.

Because the distance matrix of a graph does not account for the chemical nature of edge multiplicity of vertices, indices of neighborhood symmetry are capable of characterizing chemical structure more efficiently as compared to $I_D^W$ or $\bar{I}_D^W$.

The set of 90 topological parameters used in this paper are shown in Table 1.

## Data base

The U.S. Environmental Protection Agency Research Laboratory-Duluth is compiling a data base for approximately 25,000 chemicals selected from the Toxic Substance Control Act inventory of industrial chemicals. Current entries in the data base include physicochemical properties (e.g., *n*-octanol/water partition coefficients), biological endpoints (e.g., lethal concentrations), Chemical Abstracts Registry number, and molecular descriptors (e.g., connectivity indices). In order to minimize the cost of evaluating the utility of topological indices as discriminators of structural similarity, we decided to investigate only a subset of compounds. Faced with making a choice of which compounds to use in these modeling studies, we decided to use compounds which had at least one measured value for boiling point, melting point, or vapor pressure. A subset of 3692 compounds satisfies these criteria. A tabulation of the chemical characteristics for this set of 3692 compounds is given in Table 2. In this table we report the number of occurrences of various important chemical functional groups. The compound 2-(ethylamino)ethanol, $C_4H_{11}NO$, contains both the alcohol and amine functional groups and thus is included in the count for both categories. It is to be noted that this is a very heterogeneous set of compounds. Only 97 compounds are hydrocarbons, that is,

Table 2. Chemical characteristics of the data base of 3692 industrial chemicals.

| No. of occurrences | Functional group | No. of occurrences | Functional group |
|---|---|---|---|
| 221 | nitro | 457 | carboxylic acid |
| 853 | amine | 74 | acid halide |
| 123 | nitrile | 363 | ester |
| 18 | azo | 60 | lactone |
| 75 | imine | 55 | lactam |
| 94 | sulfide | 22 | anhydride |
| 61 | thiol | 8 | imide |
| 3 | sulfate | 94 | amide |
| 19 | sulfonate | 1 | peroxide |
| 917 | halide | 20 | isocyanate |
| 15 | phosphate | 14 | carbonic ester |
| 739 | alcohol | 133 | aldehyde |
| | | 318 | ketone |

containing only hydrogen and carbon atoms. Approximately one-half of the compounds are aromatic; 40% of the compounds are acyclic. The molecular mass range is 26–958 atomic mass units; the average molecular mass is 169 amu. Tables 2 and 5 and Fig. 5 give some indication of the nature of the complexity and diversity in this data base of 3692 chemicals.

## Methods

The data for this investigation can be viewed as $n = 3692$ vectors (chemicals) in $p = 90$ dimensions (indices). These data can be represented by a matix $X$ which has 3692 rows and 90 columns. The data matrix has 90 variables and 3692 cases. Each chemical is represented by a point in $\mathbb{R}^{90}$. If each chemical could be represented in $\mathbb{R}^2$, one could plot and investigate the extent to which similar chemicals are situated near each other according to the two descriptors. In $\mathbb{R}^{90}$ such simple analyses are not possible. However, since the indices are highly interrelated, the 3692 points in $\mathbb{R}^{90}$ will lie nearly on a subspace of lower dimension than 90. The method of principal component analysis (PCA) or the Karhunen-Loeve transformation is a standard linear method for reduction of dimensionality. This method is described in textbooks on multivariate statistics (e.g., Gnanadesikan [22] or Greenacre [25]), or in discussions of pattern recognition (e.g., Varmuza [78]). Other nonlinear methods of reduction of dimensionality and graphical representation such as multidimensional scaling are also possible (e.g., see again Gnanadesikan [22] or Varmuza [78]). However, PCA is the logical starting point in terms of simplicity, ease of interpretation, and ease of computation.

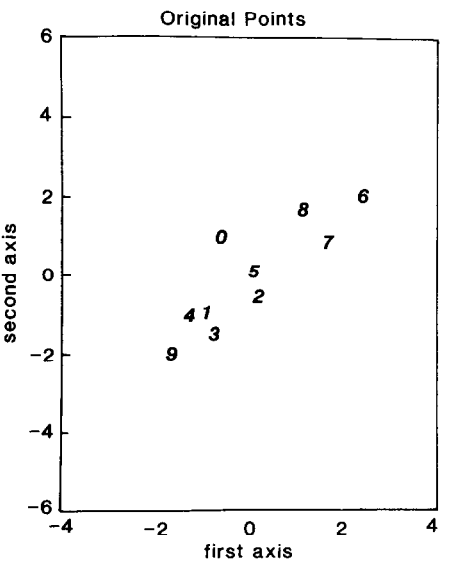



Fig. 1. Sample PCA with ten hypothetical points in two dimensions: Original points.

As an introduction to principal component analysis, consider the 10 points in two dimensions in Fig. 1. The first principal component (PC) is the line which comes closest to the points in the sense of minimizing the sum of the squared Euclidean distances from the points to the line. The one-dimensional representation of the points would be the projection of the points onto this line of closest approximation. Equivalently, projections of the points on the first principal component have maximum variance among all possible linear coordinates. The points in Fig. 1 were constructed so that the mean or center of gravity of both measurements is zero and the closest fitting line is the line passing through (0, 0) at an angle of 45 degrees. The values of principal component number one are given by the projection of the points onto this line. The second principal component is given by projections onto the basis vector orthogonal to the first principal component. Fig. 2 shows the points plotted with the principal components as the axes. The points in Fig. 2 are merely a rotation of the points in Fig. 1.

In general, for points in $\mathbb{R}^n$, the first $r$ principal components give the subspace which comes closest to approximating the $n$ points. The first principal component is the first axis of the points. Successive axes are the major directions orthogonal to previous axes. Since the closest approximating hyperplane must pass through the center of gravity of the points (Greenacre, [25, p. 44]), the first step in finding PC's is to shift the origin to the center of gravity by subtracting the column average from each column of the $n \times p$ matrix $X$. Let $\mu$ be the $p \times 1$ vector of means. Then the translated matrix is $X - \mathbf{1}\mu^{\mathrm{T}}$ where $\mathbf{1}$ is an $n \times 1$ vector of ones. The principal components are then the eigenvectors of the $p \times p$ covariance matrix $(n-1)^{-1}(X - \mathbf{1}\mu^{\mathrm{T}})^{\mathrm{T}}(X - \mathbf{1}\mu^{\mathrm{T}})$.

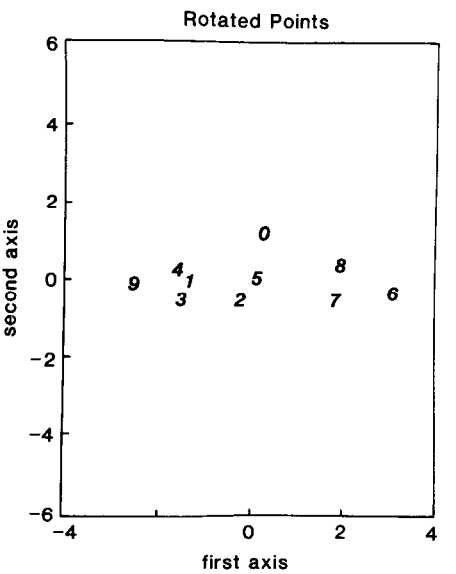



Fig. 2. Sample PCA with ten hypothetical points in two dimensions: Rotated points.

The PC's are the closest approximating hyperplane and because they are calculated from eigenvectors of a $p \times p$ matrix, the computations are relatively accessible. However, there are important scaling choices. First, the principal components depend on the scaling of the original variables. If the values of $X1$ in Fig. 1 were multiplied by say $10^6$, the best fitting line would very nearly be the line which minimizes the sum of the horizontal squared distances from the line to the points. As another example, if $X1$ and $X2$ are uncorrelated, then the first principal component is the variable with the larger variance. Thus, the first PC is scale dependent. To control this dependence, the most commonly used convention is to rescale the variables so that each variable has mean zero and standard deviation one. The effect of this scaling is to make each value correspond to the number of standard deviations from average. The covariance matrix for these rescaled variables is the correlation matrix. The values of $X1$ and $X2$ in Fig. 1 have been standardized so that each has mean zero and variance one.

In addition to changes in linear scaling, nonlinear changes in scale such as logarithmic scale affect the principal components analysis. Outliers will have a large effect on a best fitting plane when using Euclidean distance. For distributions which are positive and highly skewed by large values, a log transformation can be useful in reducing the importance of outliers and approaching a more elliptical pattern such as in Fig. 1. For the data in this investigation, the indices have been transformed by taking the log of the index plus one and then standardizing to mean zero and variance one.

The points in Fig. 2 are merely a rotation of the points in Fig. 1. Hence, distances of points from the origin in Fig. 2 are the same as in the original Fig. 1. In terms of straight Euclidean distance, point 9 is farther from the origin (near point 5) than is point 0. However, in another sense, point 0 is about as far from the general elliptical pattern of points. Another rescaling possibility is to rescale the principal components so that each has standard deviation one. Fig. 3 shows the result of this rescaling. Point 9 and point 0 are now about the same distance from the origin. Distances from points to the origin in Fig. 3 correspond to distances from the points to the origin (Greenacre [25, p. 112]) in the original Fig. 1. Mahalanobis distance imposes a metric defined by the inverse of the covariance matrix and 'sphericizes' the cloud of points so that variances of points along any direction through the centroid is a constant, 1.0.

As an extreme example of the difference between scaling and not scaling the PC's, consider the case of three variables $(X1, X2, X3)$ where $X1 = X2$ and $X3$ is uncorrelated with $X1$ and $X2$. Assume all three variables have mean zero and variance one. The first PC is $(\sqrt{2}/2)(X1 + X2) = \sqrt{2}(X1)$, and the second PC is $X3$. The squared distance from a point to the origin in either the original scale or in terms of these PC's is $2(X1)^2 + (X3)^2$. The redundant variable is used in the distance computation. If the PC's are rescaled, the first PC is $X1$, and the second PC is $X3$. The squared distance to the origin is $(X1)^2 + (X3)^2$. The standardized PC's have eliminated the redundant variable.

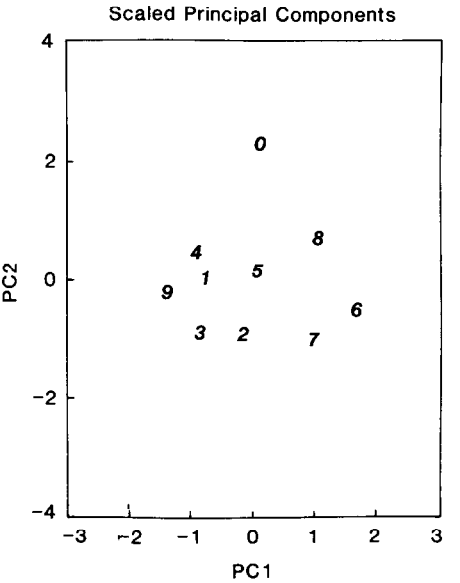



Fig. 3. Sample PCA with ten hypothetical points in two dimensions: Scaled Pricipal Components.

No scaling convention for the original variables or the PC's is by definition correct. The task is to find a convention which works for the problem at hand. The choice made in this investigation was to standardize the log transformed indices and to use standardized principal components.

Another choice in the reduction of dimensionality is to choose the number of principal components retained. Using standardized variables, the variances of the unstandardized PC's are given by their eigenvalues, and the sum of the eigenvalues is $p$, the trace of the correlation matrix or the sum of the variances of the individual standardized variables. If all $p$ standardized variables were uncorrelated, all eigenvalues would be 1.0. The eigenvalue of a PC divided by $p$ is referred to as the variance explained by that PC. The cumulative variance explained by the first $r$ PC's is the sum of their eigenvalues divided by $p$. The hope of PCA is to explain a large percentage of the total variance using a small number of PC's. In choosing the number of PC's retained, there are a number of possible conventions. The convention chosen here was to retain the PC's with eigenvalues greater than one [74].

After reducing the dimensionality of the problem, the question remains as to whether the information in these PC's is useful in picking analogs for a given chemical. A negative answer could indicate either that (1) original indices did not have the required information, (2) linear reduction in dimensionality with PC's was inappropriate, or (3) the choice of some scaling or dimension convention was inappropriate.

To test the ability of these PC's to find analogs in this heterogeneous data base, ten compounds were chosen at random from the data base of 3692 chemicals, and

the five nearest neighbors to each of the target chemicals were found. In finding nearest neighbors, the distance between two points (chemicals) $X_1$ and $X_2$ is given by

$$D = \left[ \sum_{i=1}^{10} [\mathrm{PC}_i(X_1^{\mathrm{T}}) - \mathrm{PC}_i(X_2^{\mathrm{T}})]^2 \right]^{1/2} \tag{12}$$

where $\mathrm{PC}_i$ is the $i$th scaled principal component. This defines a numerical measure of dissimilarity which is calculated solely from the chemical structure. A visual inspection of the resulting 'similar' structures can be used to evaluate the utility and limitations of this approach in selecting structural analogs.

## Results

To compute principal components, each of the 90 variables was transformed by the logarithm of the variable plus one. The principal components were then extracted from the correlation matrix, corresponding to finding PC's after standardizing the log transformed variable to mean zero and standard deviation one. Table 3 shows the resulting eigenvalues and percent of variance explained by the eigenvalues for components with eigenvalues greater than 1.0. The first ten components explain 92.6% of the variance. Correlation of the variables with the ten highest principal components is given in Table 4; for brevity only the ten most highly correlated variables are shown for each of the ten principal components. $\mathrm{PC}_1$ is highly correlated ($0.96 > r > 0.69$) with the path and cluster molecular connectivity indices and with $W$, $I_D^W$, $\bar{I}_D^W$, $K_h$ and $O$. Accordingly, this principal component is related to the size and shape of the molecular graph. It should be noted that $\mathrm{PC}_1$ is also highly correlated ($r = 0.81$) with molecular weight in this data base. The information-theoretic indices ($\mathrm{IC}_r$, $\mathrm{CIC}_r$, and $\mathrm{SIC}_r$) are, as a group, highly correlated with $\mathrm{PC}_2$ with $r_{\mathrm{avg}} = 0.80$. On the other hand the remaining 69 variables are very poorly correlated with $\mathrm{PC}_2$ ($r < 0.29$). $\mathrm{PC}_2$ can be interpreted as an axis that represents the

Table 3. Summary of principal components.

| PC | Eigenvalue | Percent of variance | Cumulative percent |
|---|---|---|---|
| 1 | 39.6 | 44.0 | 44.0 |
| 2 | 14.6 | 16.2 | 60.2 |
| 3 | 9.9 | 11.0 | 71.2 |
| 4 | 6.4 | 7.1 | 78.3 |
| 5 | 3.3 | 3.7 | 81.9 |
| 6 | 3.2 | 3.5 | 85.5 |
| 7 | 1.9 | 2.1 | 87.6 |
| 8 | 1.8 | 1.9 | 89.5 |
| 9 | 1.5 | 1.7 | 91.2 |
| 10 | 1.2 | 1.3 | 92.6 |

Table 4. Correlation coefficients of variables with the principal components (only the 10 most highly correlated listed).

| PC1 | | PC2 | | PC3 | | PC4 | | PC5 | | PC6 | | PC7 | | PC8 | | PC9 | | PC10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K_1$ | 0.959 | $SIC_3$ | 0.973 | $^4\chi_C^b$ | 0.694 | $^4\chi_{CH}$ | 0.848 | $^6\chi_{CH}$ | −0.465 | $IC_0$ | 0.501 | $^4\chi_C^v$ | 0.538 | $K_{10}$ | −0.319 | $^5\chi_{PC}^v$ | 0.292 | $IC_0$ | 0.282 |
| $^2\chi$ | 0.954 | $CIC_4$ | −0.955 | $^3\chi_C^b$ | 0.693 | $^4\chi_{CH}^b$ | 0.844 | $^6\chi_{CH}^v$ | −0.457 | $SIC_0$ | 0.424 | $^3\chi_C^v$ | 0.494 | $^6\chi^v$ | −0.311 | $^6\chi_{PC}^v$ | 0.285 | $^5\chi_C^v$ | 0.282 |
| $^3\chi$ | 0.954 | $CIC_3$ | −0.952 | $^5\chi_C^b$ | 0.683 | $^4\chi_{CH}^v$ | 0.795 | $^6\chi_{CH}^b$ | 0.437 | $^6\chi_{CH}^v$ | −0.374 | $^6\chi_C^b$ | −0.480 | $^5\chi_C^v$ | −0.309 | $^4\chi_{PC}^v$ | 0.282 | $^3\chi^v$ | 0.272 |
| $K_2$ | 0.953 | $SIC_4$ | 0.947 | $^4\chi_C$ | 0.680 | $^3\chi_{CH}$ | 0.751 | $^3\chi_{CH}^b$ | 0.406 | $O$ | −0.349 | $^6\chi_C$ | −0.434 | $K_9$ | −0.301 | $^4\chi_C$ | 0.273 | $^3\chi_C^b$ | −0.233 |
| $K_0$ | 0.949 | $SIC_2$ | 0.940 | $^3\chi_C^v$ | 0.668 | $^3\chi_{CH}^b$ | 0.751 | $^3\chi_{CH}$ | 0.406 | $SIC_1$ | 0.334 | $^5\chi_C^b$ | −0.391 | $^5\chi_{CH}$ | 0.289 | $K_9$ | 0.268 | $K_9$ | −0.232 |
| $^1\chi$ | 0.942 | $CIC_5$ | −0.940 | $^5\chi_C$ | 0.644 | $^3\chi_{CH}^b$ | 0.740 | $^3\chi_{CH}^v$ | 0.391 | $^3\chi_C^b$ | 0.318 | $^5\chi_C$ | −0.343 | $^5\chi_{CH}^b$ | 0.287 | $K_{10}$ | 0.264 | $K_8$ | −0.230 |
| $^3\chi^b$ | 0.938 | $CIC_6$ | −0.922 | $^6\chi_C$ | 0.637 | $^3\chi_{CH}^v$ | 0.718 | $^4\chi_C^b$ | 0.316 | $^3\chi_{CH}^v$ | 0.314 | $^2\chi^v$ | 0.304 | $^6\chi_C^v$ | −0.281 | $SIC_0$ | 0.249 | $^1\chi^v$ | 0.224 |
| $^4\chi$ | 0.935 | $SIC_5$ | 0.915 | $^3\chi_C$ | 0.612 | $^5\chi_{CH}$ | 0.707 | $^5\chi_{CH}^v$ | −0.311 | $^3\chi_{CH}$ | 0.314 | $^4\chi_C$ | 0.274 | $^5\chi_{CH}^v$ | 0.277 | $^5\chi_{CH}$ | 0.239 | $SIC_0$ | 0.222 |
| $^4\chi^b$ | 0.934 | $SIC_6$ | 0.887 | $^6\chi_C^b$ | 0.602 | $^5\chi_{CH}^v$ | 0.672 | $^6\chi_C^b$ | 0.304 | $^3\chi_{CH}^b$ | 0.314 | $^4\chi_C^b$ | 0.232 | $^5\chi^v$ | −0.243 | $IC_0$ | 0.235 | $^2\chi^v$ | 0.218 |
| $^0\chi$ | 0.934 | $CIC_2$ | −0.869 | $^5\chi_C^v$ | 0.600 | $^6\chi_{CH}^b$ | 0.472 | $^6\chi_C$ | 0.310 | $CIC_1$ | 0.312 | $^3\chi_C^b$ | 0.210 | $K_8$ | −0.228 | $K_8$ | 0.228 | $^6\chi_C^v$ | 0.212 |

symmetry of a molecular graph. Symmetry here is taken to mean the degree of redundancy of the neighborhoods of vertices in the molecular graph [63]. Molecular graphs with a high value of $PC_2$ are highly asymmetrical while those with a low value are symmetric. $PC_3$ is most highly correlated with cluster ($0.55 < r < 0.69$) and path/cluster ($0.27 < r < 0.59$) connectivity indices. Since these indices have traditionally been associated with branching [19, 51] in a molecular graph, $PC_3$ is a measure of the degree of branching in a molecular graph. Both acyclic and cyclic graphs having cluster and path/cluster subgraphs may be considered to be branched. As shown in Table 4, $PC_4$ is clearly correlated with cyclic terms of the molecular connectivity indices. A more detailed description of these PC's appears in Basak et al. [12].

Ten compounds were chosen at random. The names and formulas of these target compounds are found in Table 5. A plot of $PC_1$ versus $PC_2$ for all 3692 compounds is given in Fig. 4; the ten compounds chosen at random are indicated. The hydrogen-suppressed structures for these ten target compounds, labeled 1.0, 2.0, etc., are given in Fig. 5. Five nearest neighbors were selected for each target compound using the distance formula from the preceeding section. The names and formulas of the nearest neighbors for each target compound are found in Table 6 and their structures are given in Fig. 5. The notation used in Table 6 and Fig. 5 is: $n.0$ ($n = 1, 2, \ldots, 10$) identifies the $n$th target compound while $n.j$ ($n = 1, 2, \ldots, 10$; $j = 1, 2, \ldots, 5$) identifies the five nearest neighbors for the $n$th target compound.

## Discussion

The purpose of this investigation was to test the efficacy of graph-theoretic indices in the selection of similar structures from a set of diverse chemicals. Path numbers and other topological features of chemical graphs have already been used to determine structural similarity of congeners [21, 80, 81]. At the heart of any SPR or SAR method is the tacit assumption that similarity in structure results in similar physiocochemical or biomedicinal properties [73]. Topological indices correlate well

Table 5. Ten randomly selected chemicals used as target chemicals. Structures shown in Fig. 5.

| No. | Formula | Name |
|---|---|---|
| 1.0 | $C_2H_6$ | Ethane |
| 2.0 | $C_5H_{10}O_2$ | Formic acid, butyl ester |
| 3.0 | $C_9H_8O_3$ | 2-Propenoic acid, 3-(2-hydroxyphenyl)-, (E)- |
| 4.0 | $C_8H_7ClO_2$ | Benzeneacetic acid, 4-chloro- |
| 5.0 | $C_7H_8O$ | Benzenemethanol |
| 6.0 | $C_7H_4Cl_2O$ | Benzaldehyde, 3,4-dichloro- |
| 7.0 | $C_7H_3F_5O$ | Benzene, pentafluoromethoxy- |
| 8.0 | $C_8H_4F_3NO$ | Benzene, 1-isocyanato-3-(trifluoromethyl)- |
| 9.0 | $C_8H_4F_3NO$ | 1-Naphthalenol, acetate |
| 10.0 | $C_{11}H_8O_2$ | Butanamide, N-phenyl- |

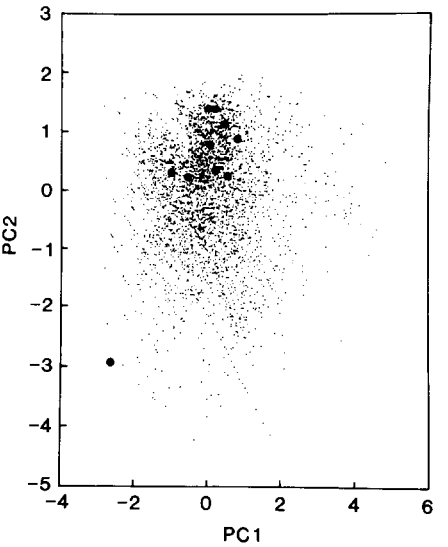



Fig. 4. $PC_1$ versus $PC_2$ for 3692 chemicals. The ten chemicals chosen randomly are indicated with a ●.

with the properties of different congeneric sets of chemicals [7–14, 14, 32]. These results, derived from groups of molecules of limited structural variety, indicate that numerical graph invariants are quantifiers of structural similarity or dissimilarity which strongly determines the behavior of chemical species. However, to our knowledge, topological indices have not been used to select structural analogs from a heterogeneous collection of molecules. This encouraged us to use the topologically derived structure-space in the quantitation of structural similarity.

The ninety topological parameters calculated for 3692 chemicals encode information regarding the size, shape, bonding type, and branching pattern associated with molecular structures. To determine structural similarity one could plot chemicals as points in the 90-dimensional space and use Euclidean distance between a given pair of chemicals in this space to quantify structural similarity. However, we have found that many of the indices are highly intercorrelated and ten principal components derived from the $90 \times 90$ variable matrix explain 92.6% of the variation in the original data [12]. Therefore, we decided to determine structural similarity of molecules in terms of their distance in a 10-dimensional structure-space where 10 PC's constitute individual coordinates. This reduction of dimensionality diminished the magnitude of the problem retaining, at the same time, most of the original structural information.

Fig. 5 depicts hydrogen-suppressed graphs of ten randomly selected target molecules and five nearest neighbors for each of them in the trial universe of 3692 chemicals. It is clear from the results that the family of topological indices considered in this paper has a considerable power of rejecting dissimilar structures. However, some of the neighbors selected by our method have functional groups different from the target species. Consequently, the reactivity profile of a target and its nearest neighbor could be quite different.

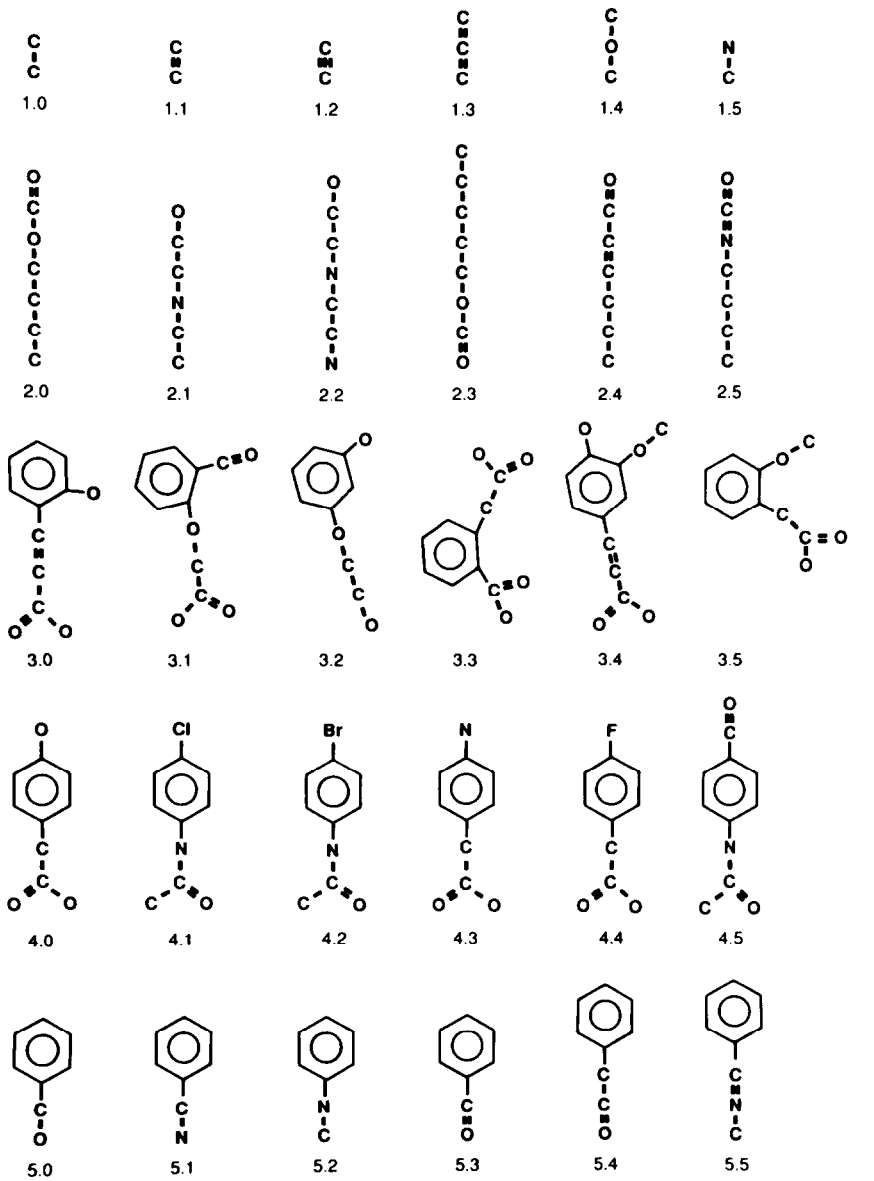



Fig. 5. Hydrogen-suppressed structures of ten target chemicals chosen randomly and their five nearest neighbors. Target chemicals are labeled 1.0, 2.0, etc. Nearest neighbors are labeled 1.1, 1.2, 1.3, 1.4, and 1.5 for compound 1; 2.1, 2.2, ..., 2.5 for compound 2, etc.

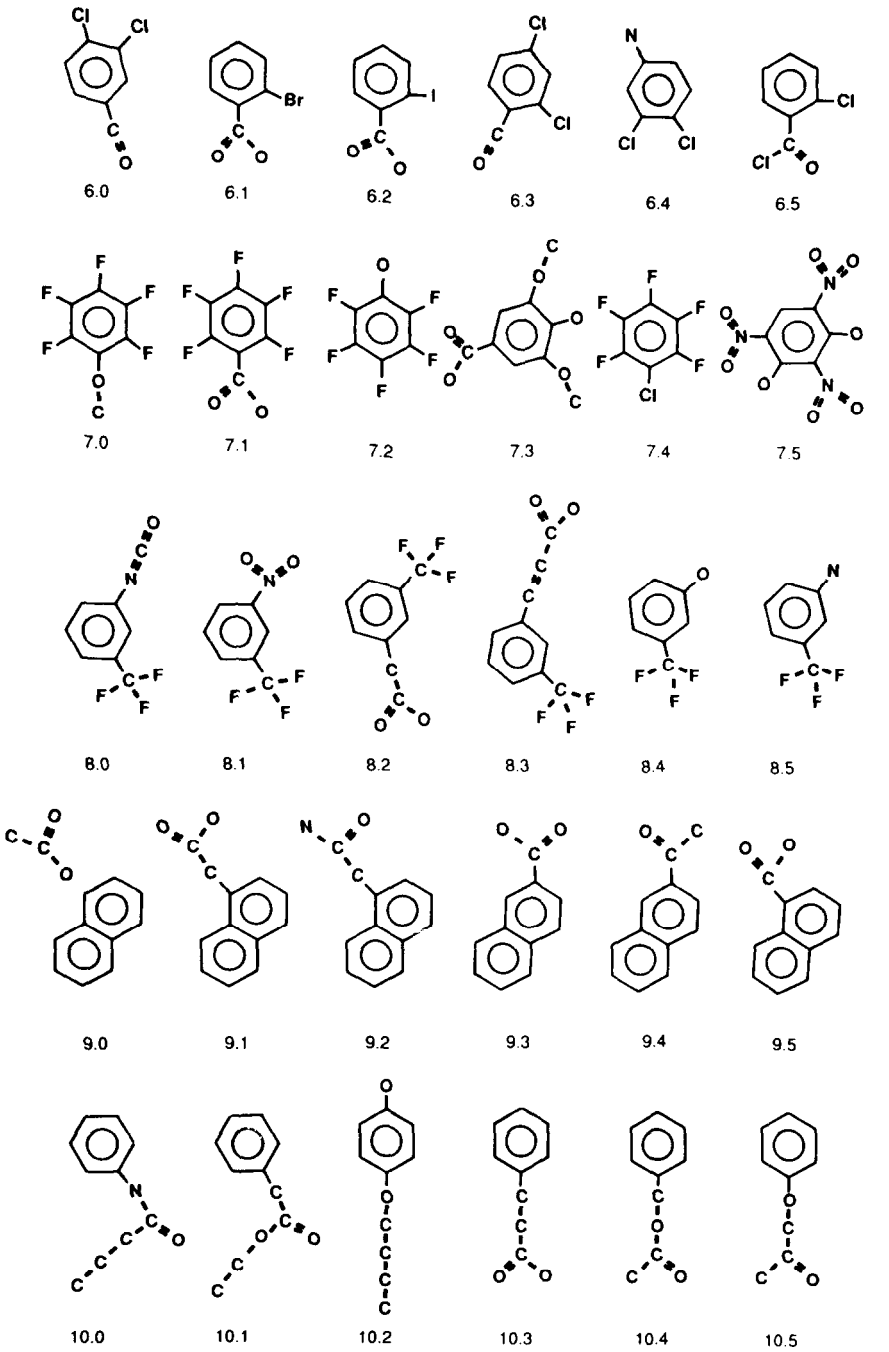

6.0
6.1
6.2
6.3
6.4
6.5
7.0
7.1
7.2
7.3
7.4
7.5
8.0
8.1
8.2
8.3
8.4
8.5
9.0
9.1
9.2
9.3
9.4
9.5
10.0
10.1
10.2
10.3
10.4
10.5

Fig. 5 (contd.).

Table 6. Ten random target chemicals with their five nearest neighbors. Structures shown in Fig. 5.

| Target chemical | No. | Formula | Distance | Name |
|---|---|---|---|---|
| 1 | 1.0 | $C_2H_6$ | | Ethane |
| | 1.1 | $C_2H_4$ | 0.817 | Ethene |
| | 1.2 | $C_2H_2$ | 2.147 | Ethyne |
| | 1.3 | $C_3H_4$ | 2.614 | 1,2-Propadiene |
| | 1.4 | $C_2H_6O$ | 2.773 | Methane, oxybis- |
| | 1.5 | $CH_5N$ | 3.203 | Methanamine |
| 2 | 2.0 | $C_5H_{10}O_2$ | | Formic acid, butyl ester |
| | 2.1 | $C_4H_{11}NO$ | 0.293 | Ethanol, 2-(ethylamino)- |
| | 2.2 | $C_4H_{12}N_2O$ | 0.430 | Ethanol, 2-(2-aminoethyl)amino- |
| | 2.3 | $C_6H_{12}O_2$ | 0.481 | Formic acid, pentyl ester |
| | 2.4 | $C_6H_{10}O$ | 0.488 | 2-Hexenal, (E)- |
| | 2.5 | $C_5H_9NO$ | 0.497 | Butane, 1-isocyanato- |
| 3 | 3.0 | $C_9H_8O_3$ | | 2-Propenoic acid, 3-(2-hydroxyphenyl)-, (E)- |
| | 3.1 | $C_9H_8O_4$ | 0.386 | Acetic acid, (2-formylphenoxy)- |
| | 3.2 | $C_8H_{10}O_3$ | 0.626 | Phenol, 3-(2-hydroxyethoxy)- |
| | 3.3 | $C_9H_8O_4$ | 0.683 | Benzeneacetic acid, 2-carboxy- |
| | 3.4 | $C_{10}H_{10}O_4$ | 0.733 | 2-Propenoic acid, 3-(4-hydroxy-3-methoxyphenyl)- |
| | 3.5 | $C_9H_{10}O_3$ | 0.804 | Benzeneacetic acid, 2-methoxy- |
| 4 | 4.0 | $C_8H_7ClO_2$ | | Benzeneacetic acid, 4-chloro- |
| | 4.1 | $C_8H_8ClNO$ | 0.230 | Acetamide, N-(4-chlorophenyl)- |
| | 4.2 | $C_8H_8BrNO$ | 0.486 | Acetamide, N-(4-bromophenyl)- |
| | 4.3 | $C_8H_8NO_2$ | 0.488 | Benzeneacetic acid, 4-amino- |
| | 4.4 | $C_8H_7FO_2$ | 0.520 | Benzeneacetic acid, 4-fluoro- |
| | 4.5 | $C_9H_9NO_2$ | 0.584 | Acetamide, N-(4-formylphenyl)- |
| 5 | 5.0 | $C_7H_8O$ | | Benzenemethanol |
| | 5.1 | $C_7H_9N$ | 0.129 | Benzenemethanamine |
| | 5.2 | $C_7H_9N$ | 0.204 | Benzenamine, N-methyl- |
| | 5.3 | $C_7H_6O$ | 0.306 | Benzaldehyde |
| | 5.4 | $C_8H_8O$ | 0.322 | Benzeneacetaldehyde |
| | 5.5 | $C_8H_9N$ | 0.326 | Methanamine, N-(phenylmethylene)- |
| 6 | 6.0 | $C_7H_4Cl_2O$ | | Benzaldehyde, 3,4-dichloro- |
| | 6.1 | $C_7H_5BrO_2$ | 0.225 | Benzoic acid, 2-bromo- |
| | 6.2 | $C_7H_5IO_2$ | 0.294 | Benzoic acid, 2-iodo- |
| | 6.3 | $C_7H_4Cl_2O$ | 0.355 | Benzaldehyde, 2,4-dichloro- |
| | 6.4 | $C_6H_5Cl_2N$ | 0.394 | Benzenamine, 3,4-dichloro- |
| | 6.5 | $C_7H_4Cl_2O$ | 0.395 | Benzoyl chloride, 2-chloro- |
| 7 | 7.0 | $C_7H_3F_5O$ | | Benzene, pentafluoromethoxy- |
| | 7.1 | $C_7HF_5O_2$ | 0.468 | Benzoic acid, pentafluoro- |
| | 7.2 | $C_6HF_5O$ | 0.486 | Phenol, pentafluoro- |
| | 7.3 | $C_9H_{10}O_5$ | 0.984 | Benzoic acid, 4-hydroxy-3,5-dimethoxy- |
| | 7.4 | $C_6ClF_5$ | 1.071 | Benzene, chloropentafluoro- |
| | 7.5 | $C_6H_3N_3O_8$ | 1.104 | Benzenediol,2,4,6-trinitro- |

Table 6 (contd.).

| Target chemical | No. | Formula | Distance | Name |
|---|---|---|---|---|
| 8 | 8.0 | $C_8H_4F_3NO$ | | Benzene, 1-isocyanato-3-(trifluoromethyl)- |
| | 8.1 | $C_7H_4F_3NO_2$ | 0.574 | Benzene, 1-nitro-3-(trifluoromethyl)- |
| | 8.2 | $C_9H_7F_3O_2$ | 0.667 | Benzeneacetic acid, 3-(trifluoromethyl)- |
| | 8.3 | $C_{10}H_7F_3O_2$ | 0.922 | 2-Propenoic acid, 3-3-(trifluoromethyl)phenyl- |
| | 8.4 | $C_7H_5F_3O$ | 0.961 | Phenol, 3-(trifluoromethyl)- |
| | 8.5 | $C_7H_6F_3N$ | 1.067 | Benzenamine, 3-(trifluoromethyl)- |
| 9 | 9.0 | $C_{12}H_{10}O_2$ | | 1-Naphthalenol, acetate |
| | 9.1 | $C_{12}H_{10}O_2$ | 0.378 | 1-Naphthaleneacetic acid |
| | 9.2 | $C_{12}H_{11}NO$ | 0.399 | 1-Naphthaleneacetamide |
| | 9.3 | $C_{11}H_8O_2$ | 0.440 | 2-Naphthalenecarboxylic acid |
| | 9.4 | $C_{12}H_{10}O$ | 0.484 | Ethanone, 1-(2-naphthalenyl)- |
| | 9.5 | $C_{11}H_8O_2$ | 0.527 | 1-Naphthalenecarboxylic acid |
| 10 | 10.0 | $C_{10}H_{13}NO$ | | Butanamide, N-phenyl- |
| | 10.1 | $C_{10}H_{12}O_2$ | 0.169 | Benzeneacetic acid, ethyl ester |
| | 10.2 | $C_{10}H_{14}O_2$ | 0.215 | Phenol, 4-butoxy- |
| | 10.3 | $C_9H_{10}O_2$ | 0.289 | Benzenepropanoic acid |
| | 10.4 | $C_9H_{10}O_2$ | 0.315 | Acetic acid, phenylmethyl ester |
| | 10.5 | $C_9H_{10}O_2$ | 0.344 | 2-Propanone, 1-phenoxy- |

The term 'structural similarity' is not explicitly defined in the chemical literature [80]. It is an intuitive concept used by the chemist to classify molecules in terms of certain critical structural features relevant to a particular context. Therefore, there could be more than one measure of similarity and each could be meaningful in its particular context. In this paper we have attempted to derive an operational definition of similarity using a group of topological indices.

Judgements regarding the success of our method must recognize three factors which are at play in the selection of nearest neighbors. The first is that analogous structures can only be selected if they are present in the trial universe. In some cases, the nearest neighbors are not topologically similar because structures analogous to the target are absent in the trial universe and accurate selection can only be expected from larger sets of chemicals. Fig. 4 clearly shows many sparse areas in the chemical space. It is hoped that this method will emerge as a powerful tool in selecting analogs in drug design where molecular manipulation is carried out with a particular 'lead' structure. Secondly, topological similarity of a target and its neighbors is evident in the ability to select similar skeletal graphs, i.e, graphs where the nature of the vertices, and bonding pattern are ignored. In cases where chemical intuition would dispute similarity between a target and its nearest neighbor, a substantial degree of similarity is evident with respect to size, branching pattern, cyclicity, and aromaticity. In our studies the data suggest that chemicals which fall within a distance of 0.3 to 0.4 from a target possess substantial topological similarity with

the target structure. The third factor is that topological indices encode no information about 3-dimensional molecular geometry and little information about electronic characteristics of atoms. Consequently, nearest neighbors with isomorphic skeletal graphs may have quite different biological properties or chemical reactivity profiles. In view of the power of the method in rejecting dissimilar structures, our goal is to improve the selection of neighbors which have similar chemical properties and biological action. We are continuing our exploration of additional molecular descriptors which can be incorporated into structure space coordinates through principal component analysis and improve the chemical meaning of distance.

In all cases, a target and its isomorphic nearest neighbor are not completely similar, i.e., their distance is not zero. This is interesting from the viewpoint of chemistry because substitution of a given atom of a molecule by another atom of equal valency but different electronic nature often drastically alters molecular properties. Substitutions of this type will always produce isomorphic molecular graphs. The chemically meaningful discrimination of such structures by our PC-space is probably due to contributions of valence connectivity and neighborhood indices which take into account the chemical nature of vertices. At the same time, if molecular topology and chemical reactivity have altogether different bases, it may be necessary to select similar chemicals by first selecting a group of topologically similar structures, and then order the set of chemicals with respect to biological property or chemical reactivity using more sophisticated molecular parameters which take care of geometry and electron distribution.

In conclusion, the structure-space constructed from 10 PC's contains topological information useful in ascertaining structural similarity or dissimilarity of molecules. It may be mentioned, however, that none of the topological parameters has any relation to the metric aspects (e.g., bond angle, bond distance, steric strain, etc.) of molecular architecture. Therefore, our structure-space is incapable of discriminating among stereoisomers or taking care of overcrowding effect in molecules with vicinal bulky groups. We hope that within these constraints the method developed in this paper will provide a quantitative basis for intuitive notions of structural similarity.

## Acknowledgments

This research was supported by cooperative agreements (CR-810824-01 and CR-811981-01) between the U.S. Environmental Protection Agency and the University of Minnesota, Duluth. The authors are appreciative of the efforts of Cynthia Frane, Greg Grunwald, Mark Rosen and Jane Zeleznikar for their assistance in the project.

## References


[1] G.A. Baker, Jr., Drum shapes and isospectral graphs, J. Math. Phys. 7 (1966) 2238–2242.

[2] A.T. Balaban, Chemical Applications of Graph Theory (Academic Press, New York, 1976).

[3] A.T. Balaban, Highly discriminating distance-based topological index *J*, Chem. Phys. Lett. 89 (1982) 399–404.

[4] A.T. Balaban and F. Harary, The characteristic polynomial does not uniquely determine the topology of a molecule, J. Chem. Doc. 11 (1971) 258–259.

[5] A.T. Balaban and L.V. Quintas, The smallest graphs, trees, and 4-trees with degenerate topological index *J*, Match 14 (1983) 213–233.

[6] K. Balasubramanian, Symmetry and spectra of graphs and their chemical applications, in: R.B. King, ed., Chemical Applications of Topology and Graph Theory (Elsevier, Amsterdam, 1983).

[7] S.C. Basak, D.P. Gieschen, D.K. Harriss and V.R. Magnuson, Physicochemical and topological correlates of enzymatic acetyl-transfer reaction, J. Pharm. Sci. 72 (1983) 934–937.

[8] S.C. Basak, D.P. Gieschen and V.R. Magnuson, A quantitative correlation of the $LC_{50}$ values of esters in *pimephales promelas* using physicochemical and topological parameters, Environ. Toxicol. Chem. 3 (1984) 191–199.

[9] S.C. Basak, D.P. Gcischen, V.R. Magnuson and D.K. Harriss, Structure-activity relationships and pharmacokinetics: a comparative study of hydrophobicity, van der Waals' volume and topological parameters, IRCS Med. Sci. 10 (1982) 619–620.

[10] S.C. Basak, D.K. Harriss and V.R. Magnuson, Comparative study of lipophilicity *versus* topological molecular descriptors in biological correlations, J. Pharm. Sci. 73 (1984) 429–437.

[11] S.C. Basak and V.R. Magnuson, Molecular topology and narcosis – a quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC), Arzneim. Forsch./Drug Res. 33 (1983) 501–503.

[12] S.C. Basak, V.R. Magnuson, G.J. Niemi, R.R. Regal and G.D. Veith, Topological indices: their nature, mutual relatedness, and applications, in: X.J.R. Avula, G. Leitmann, C.D. Mote and E.Y. Rodin, eds., Proceedings of the Fifth International Conference on Mathematical Modelling (Pergamon Press, Oxford, 1986).

[13] S.C. Basak, S.K. Ray, C. Raychaudhury, A.B. Roy and J.J. Ghosh, Molecular topology and pharmacological action: a QSAR study of tetrazoles using topological information content (IC), IRCS Med. Sci. 10 (1982) 145–146.

[14] S.C. Basak, A.B. Roy and J.J. Ghosh, Study of the structure-function relationship of pharmacological and toxicological agents using information theory, in: X.J.R. Avula, R. Bellman, Y.L. Luke and A.k. Rigler, eds., Proceedings of the Second International Conference on Mathematical Modelling, Vol. II (University of Missouri-Rolla, 1980) 851–856.

[15] S.H. Bertz, Convergence, molecular complexity and synthetic analysis, J. Am. Chem. Soc. 104 (1982) 5801–5803.

[16] S.H. Bertz, On the complexity of graphs and molecules, Bull. Math. Biol. 45 (1983) 849–855.

[17] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures (Research Studies Press, Chichester, 1983).

[18] D. Bonchev and O. Mekenyan, Comparability graphs and electronic spectra of condensed benzenoid hydrocarbons, Chem. Phys. Lett. 98 (1983) 134–138.

[19] D. Bonchev and N. Trinajstić, Information theory, distance matrix and molecular branching, J. Chem. Phys. 67 (1977) 4517–4533.

[20] L. Brillounin, Science and Information Theory (Academic Press, New York, 1956).

[21] K. Enslein and P.N. Craig, A toxicity estimation model, J. Environ. Path. Toxicol. 2 (1978) 115–121.

[22] R. Gnanadesikan, Methods for Statistical Analysis of Multivariate Observations (Wiley, New York, 1977).

[23] M. Gordon and J.W. Kennedy, The graph-like state of matter: 2. LCGI schemes for the thermodynamics of alkane and the theory of inductive inference, J. Chem. Soc. Faraday (II) 69 (1973) 484–504.
[24] M. Gordon and S.B. Ross-Murphy, The structure and properties of molecular trees and networks, Pure & Applied Chem. 43 (1975) 1–26.
[25] M.J. Greenacre, Theory and Applications of Correspondence Analysis (Academic Press, New York, 1984).
[26] I. Gutman and H. Hosoya, On the calculation of the acyclic polynomial, Theoret. Chim. Acta (Berl.) 48 (1978) 279–286.
[27] F. Harary, Graph Theory (Addison-Wesley, Reading, MA, 1969) 3.
[28] F. Harary, C. King, A. Mowshowitz and R.C. Read, Cospectral graphs and digraphs, Bull. London Math. Soc., 3 (1971) 321–328.
[29] G. Karreman, Topological information content and chemical reactions, Bull. Math. Biophys. 17 (1955) 279–285.
[30] J.W. Kennedy and L.V. Quintas, Extremal f-trees and embedding spaces for molecular graphs, Discrete Appl. Math. 5 (1983) 191–209.
[31] L.B. Kier, Use of molecular negentropy to encode structure governing biological activity, J. Pharm. Sci. 69 (1980) 807–810.
[32] L.B. Kier and L.H. Hall, Molecular Connectivity in Chemistry and Drug Research (Academic Press, New York, 1976).
[33] R.B. King, Chemical Applications of Topology and Graph Theory (Elsevier, Amsterdam, 1983).
[34] V.R. Magnuson, D.K. Harriss and S.C. Basak, Topological indices based on neighborhood symmetry: chemical and biological applications, in: R.B. King, ed., Chemical Applications of Topology and Graph Theory (Elsevier, Amsterdam, 1983).
[35] O. Mekenyan, S. Dimitrov and D. Bonchev, Graph-theoretical approach to the calculation of physico-chemical properties of polymers, Eur. Polym. J. 19 (1983) 1185–1193.
[36] H.L. Morgan, The generation of a unique machine description of chemical structures – a technique developed at chemical abstracts service, J. Chem. Doc. 5 (1965) 107–113.
[37] A. Mowshowitz, Entropy and the complexity of graphs: I. an index of the relative complexity of a graph, Bull. Math. Biophys. 30 (1968) 175–203.
[38] A. Mowshowitz, Entropy and the complexity of graphs: II. the information content of digraphs and infinite graphs, Bull. Math. Biophys. 30 (1968) 225–240.
[39] A. Mowshowitz, Entropy and the complexity of graphs: III. graphs with prescribed information content, Bull. Math. Biophys. 30 (1968) 387–414.
[40] A. Mowshowitz, Entropy and the complexity of graphs: IV. entropy measures and graphical structure, Bull. Math. Biophys. 30 (1968) 533–546.
[41] G.J. Niemi, R.R. Regal and G.D. Veith, Applications of molecular connectivity indices and multivariate analysis in environmental chemistry, in: J.J. Breen and P.E. Robinson, eds. American Chemical Society symposium series, in press.
[42] V. Prelog, Forward, in: A.T. Balaban, ed., Chemical Applications of Graph Theory (Academic Press, New York, 1976).
[43] L.V. Quintas and P.J. Slater, Pairs of non-isomorphic graphs having the same path degree sequence, Match 12 (1981) 75–86.
[44] L.V. Quintas and J. Yarmish, Degree distributions for random 4-trees and skeletons of symmetry weighted (1,4)-trees, Congressus Numerantium 36 (1982) 115–125.
[45] L.V. Quintas and J. Yarmish, The number of chiral alkanes having given diameter and carbon automorphism group a symmetric group, Second International Conference on Combinatorial Mathematics (New York, NY, April 1978), Ann. N.Y. Acad. Sci. 319 (1979) 436–443.
[46] L.V. Quintas and J. Yarmish, Valence isomers for chemical trees, Match 12 (1981) 65–73.
[47] M. Randić, On molecular identification numbers, J. Chem. Inf. Comput. Sci. 24 (1984) 164–175.
[48] M. Randić, Characterizations of atoms, molecules and classes of molecules based on paths enumerations, Proc. of Bremen Konferenz zur Chemie Univ. Bremen, 1978, Part II, Match 7 (1979) 5–64.

[49] M. Randić, Symmetry properties of graphs of interest in chemistry. II. Desargus-Levi graph, Int. J. Quant. Chem. 15 (1979) 663–682.

[50] M. Randić, Conjugated circuits and resonance energies of benzenoid hydrocarbons, Chem. Phys. Letters 38 (1976) 68–70.

[51] M. Randić, On characterization of molecular branching, J. Am. Chem. Soc. 97 (1975) 6609–6615.

[52] N. Rashevsky, Life, information theory and topology, Bull. Math. Biophys. 17 (1955) 229–235.

[53] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, A quantitative structure-activity relationship study of tumor-inhibitory triazenes using bonding information content and lipophilicity, IRCS Med. Sci. 10 (1982) 933–934.

[54] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, A quantitative structure-activity relationship (QSAR) analysis of carbamoyl piperidines, barbiturates and alkanes using information-theoretic topological indices, Ind. J. Pharmacol. 13 (1982) 301–312.

[55] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, A quantitative structure-activity relationship study of N-alkylnorketobemidones and triazinones using structural information content, Arzneim. Forsch. 32 (1982) 322–325.

[56] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, The utility of information content, structural information content, hydrophobicity and van der Waals' volume in the design of barbiturates and tumor inhibitory triazenes, Arzneim. Forsch. 33 (1983) 352–356.

[57] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, Quantitative structure-activity relationship studies of bioactive molecules using structural information indices, Ind. J. Chem. 20B (1981) 894–897.

[58] C. Raychaudhury, S.C. Basak, A.B. Roy and J.J. Ghosh, Quantitative structure-activity relationship (QSAR) studies of pharmacological agents using topological information content, Indian Drugs, 18 (1980) 97–102.

[59] R.C. Read, A new system for the designation of chemical compounds. 1. theoretical preliminaries and the coding of acyclic compounds, J. Chem. Inf. Comput. Sci. 23 (1983) 135–149.

[60] R.C. Read, A new system for the designation of chemical compounds. 2. coding of cyclic compounds, J. Chem. Inf. Comput. Sci. 25 (1985) 116–128.

[61] R.C. Read and D.G. Corneil, The graph isomorphism disease, J. Graph Theory 1 (1977) 339–363.

[62] D.H. Rouvray and A.T. Balaban, Chemical applications of graph theory, in: Wilson and L.W. Beineke, eds., Applications of Graph Theory (Academic Press, New York, 1979) 177–221.

[63] A.B. Roy, S.C. Basak, D.K. Harriss and V.R. Magnuson, Neighborhood complexities and symmetry of chemical graphs, and their biological applications, in: X.J.R. Avula, R.E. Kalman, A.I. Liapis and E.Y. Rodin, eds., Mathematical Modelling in Science and Technology (Pergamon Press, New York, 1984) 745–750.

[64] G. Sabidussi, Graphs with given group, and given graph-theoretical properties, Canad. J. Math. 9 (1957) 515–525.

[65] A.K. Samanta, S.K. Ray, S.C. Basak and S.K. Bose, Molecular connectivity and antifungal activity – a quantitative structure-activity relationship study of substituted phenols against skin pathogens, Arzneim. Forsch. 32 (1982) 1515–1517.

[66] R. Sarkar, A.B. Roy and P.K. Sarkar, Topological information content 6f genetic molecules – I., Math. Biosci. 39 (1978) 299–312.

[67] A.J. Schwenk, Almost all trees are cospectral, in: F. Harary, ed., New Directions in the Theory of Graphs (Academic Press, New York, 1973).

[68] P.J. Slater, Counterexamples to Randić's conjecture on distance degree sequences for trees, J. Graph Theory 6 (1982) 89–92.

[69] E.A. Smolenski, Application of the theory of graphs to calculations of the additive structural properties of hydrocarbons, Russ. J. Phys. Chem. 38 (1964) 700–702.

[70] L. Spialter, The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP): a new computer-oriented chemical nomenclature, J. Am. Chem. Soc. 85 (1963) 2012–2013.

[71] L. Spialter, The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP), J. Chem. Doc. 4 (1964) 261–269.

[72] L. Spialter, The atom connectivity matrix characteristic polynomial (ACMCP) and its physico-geometric (topological) significance, J. Chem. Doc. 4 (1964) 269–274.
[73] A.J. Stuper, W.E. Brugger and P.C. Jurs, Computer Assisted Studies of Chemical Structure and Biological Function (Wiley, New York, 1979).
[74] M.M. Tatsuoka, Multivariate Analysis: Techniques for Educational and Psychological Research (Wiley, New York, 1971).
[75] N. Trinajstić, Chemical Graph Theory (CRC Press, Boca Raton, FL, 1983).
[76] E. Trucco, A note on the information content of graphs, Bull. Math. Biophys. 18 (1956) 129–135.
[77] E. Trucco, On the information content of graphs: compound symbols; different states for each point, Bull. Math. Biophys. 18 (1956) 237–253.
[78] K. Varmuza, Pattern Recognition in Chemistry (Springer, New York, 1980).
[79] H. Wiener, Structural determination of paraffin boiling point, J. Am. Chem. Soc. 69 (1947) 17–20.
[80] C.L. Wilkins and M. Randić, A graph theoretical approach to structure-property and structure-activity correlations, Theoret. Chim. Acta (Berl.) 58 (1980) 45–68.
[81] M. Yuan and P.C. Jurs, Computer-assisted structure-activity studies of chemical carcinogens: a polycyclic aromatic hydrocarbon data set, Toxicol. Appl. Pharmacol. 52 (1980) 294–312.