On Molecular Identification Numbers[†]

MILAN RANDIĆ

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311, and Ames Laboratory–DOE, Iowa State University, Ames, Iowa 50011

Received December 20, 1983

It is proposed to assign identification numbers (I.D.) to molecular skeletons. Numbers should be (1) easy to derive, (2) unique, and (3) have structural significance. We outlined an attempt in this direction, based on the count of suitably weighted paths in the structure. The desired properties 1 and 3 are demonstrated, while 2, uniqueness, remains to be proven, or disproven (e.g., by demonstration of a counterexample: a pair of nonisomorphic structures having the same I.D. numbers). Application to the known cases that have been troublesome in the past in failing to differentiate some related structures has not revealed a single case of limited discriminatory power of the scheme proposed here. Examined cases include isospectral graphs, isocodal graphs using different coding procedures, graphs having identical topological indexes, path degree sequences, and distance degree sequences. Hence, the scheme deserves a wider critical examination and evaluation with respect to its use in SAR (structure–activity relationship) and possibly as supplementary reference in *Chemical Abstracts*.

INTRODUCTION

Chemical names and codes serve several purposes, which, in a way, explain why different schemes continue to be used. Trivial names are convenient, sometimes descriptive, allowing one to relate effectively to derivatives of a compound. Standard names try to organize the plethora of structures more systematically—as a policy, they aim at a more satisfactory task. The problem, however, occurs in numerous overlapping areas that require hierachical rules in determining (by a convention) priority among alternatives. With time, the rules appear have grown so much that their application is by no means straightforward. Computer manipulation of chemical structures introduced novel coding schemes. The codes no longer necessarily suit a common use, although they may maintain some resemblance to traditional approaches. Illustrations are Wiswesser line notation or Dyson's system,[1] which are based on a convention. Alternative are the schemes attempting a more structural approach, using adjacency relations in the molecular graph as the starting point. An illustration of such an approach is Morgan's algorithm for labeling atoms,[2] which results in some canonical form for the structure. Although such schemes have reached the stage of practical implementation, an interest in more advanced schemes, whether they can be used to name compounds or derive molecular codes for computer use, remains.[3] An important advantage of structural approach based on mathematical analysis is anticipation of unusual situations and, hence, no need for continuous updating, which typifies empirical schemes. However, mathematical codes are unusually long even for relatively small structures, if they are to encode completely all structural features. The important question then is, is there anything of practical use for chemical documentation in the large gap between the CAS Registry Numbers[4] and very detailed chemical codes? Registry numbers, being historical rather than structural, offer little if any use for *comparison* of structures. In order to keep track of individual compounds, it would be desirable to have simple identification numbers (I.D.'s). Unique canonical labels[5] (illustrated in Table I), even though suitable for computer processing,[6] cannot be considered simple, despite the fact that some have quite elegant interpretation.[7] As we will show, one can devise a coding scheme that is both structural and brief.

Table I. Illustration of Binary Codes and Their Growth with Size: The Cases of Canonical Numbering Based Code and "Walk Around" Codes of Read

structure	binary
	Minimal Binary Number
n-butane	0 0 1 1 0 1
n-pentane	0 0 0 1 0 1 0 1 1 0
n-hexane	0 0 0 0 1 0 0 1 0 1 0 1 1 0 0
h-heptane	0 0 0 0 0 1 0 0 0 1 0 0 1 0 1 1 1 0 0 0 0
	Walk Around
n-butane	0 0 0 1 1 1
n-pentane	0 0 0 0 1 1 1 1
n-hexane	0 0 0 0 0 1 1 1 1 1
n-heptane	0 0 0 0 0 0 1 1 1 1 1 1

Recently, Read[8] reviewed useful qualities that chemical coding systems ought to acquire. These include (1) that codes be a linear string of symbols, (9) that codes should be pronouncable, and (10) that the symbols used in the code be familiar. All these are automatically satisfied if codes are represented by numbers—as we intend to do. The remaining conditions we will collect in two groups. Group I includes (2) codes should be unique, (3) codes should be reconstructable, (4) preferentially codes should be derived by hand, (5) decoding should be possible by hand calculations, and (12) coding and decoding algorithms should be efficient. Group II includes (6) the coding process should not depend on properties of compounds, (7) the codes should be free from conventional (nonsystematic) items, (8) the codes should be brief, and (11) the codes should be easily comprehensible.

In the first group of conditions item 2 is fundamental, but as Read points out, "with some systems it is very difficult to be sure whether they possess attribute (2) or not."[8] From a practical point of view, such systems can be used, and when problems arise, some modification may upgrade the system and resolve the situation—pragmatism demands that such exceptions be few. Condition 3 is desirable (prerequisite for reconstruction is uniquenes of the codes); if not possible, one can always recover the structure from the catalog of structures. Conditions 4 and 5 are for the convenience of bench chemists. With the present proliferation of small calculators and personal computers, these two conditions will be as archaic in the near futute as log tables are today. Finally, condition 12 appears essential. Theoretically, an algorithm is considered efficient

[†] Dedicated to Professor R. S. Hansen, who graciously continues to support M.R.'s stubborn devotion to chemical graph theory. Ames Laboratory–DOE is operated for the U.S. Department of Energy by Iowa State University under Contract W-7405-Eng-82. This work was supported in part by the office of R. S. Hansen, Director.

0095-2338/84/1624-0164$01.50/0 © 1984 American Chemical Society

if it involves polynomial rather than exponential (or worse) growth in the number of operations with an increase in the size of a structure. From the practical point of view, an algorithm is acceptable, regardless of its intrinsic character, if it gives results in practical time. This will be measured in seconds and minutes for structures of great frequency, but it may be longer for occasional structures. The program ALL-PATH,[9] which enumerates paths of different length in molecular skeletons, is nonpolynomial (NP). Already, finding the longest path between two vertices is a proven NP-complete problem;[10] hence, listing all paths is only worse. Yet, for molecules of chemical interest having for instance, 40–50 atoms and five to six rings, the program typically gives results in a fraction of a minute.[9] There is lot of chemistry within the indicated range of size and ring number, the latter being the critical limitation. Symmetry of a molecule has not been incorporated in the ALL-PATH program, which treats each vertex (atom) separately and constructs the sequence of path numbers for such atoms. The program is applicable to *all* chemical structures, including asymmetrical molecules having more than 10 rings. Moreover, the program is applicable to any graph, whether it represents a molecular system, a chemical isomerization system, or any abstract concept. Eventually, the size of the system will be too large, but as was reported elsewhere,[11] enumeration of paths is practical (with current computers) when the total number of paths is of the order of 10^6. This limit encompasses many graphs, some of visible complexity. For example, the total number of paths in dodecahedron (12-membered ring system) is just above 250 000 while "cube-in-cube" (a four-dimensional cube) having 24 fundamental rings has a little less than one and a half million paths.

Our current approach represents a modification of path counts. The derived molecular I.D. numbers satisfy the second group of conditions reviewed by Read. The first group of conditions is in part not satisfied (less essential conditions 3, 4, 5, and 12) and in part not resolved (condition 2). No case has yet been found of two structures having the same code, which shows that the discrimination power of molecular I.D. numbers is very high. While it should not be suprising that different structures may eventually be found that will have the same I.D. numbers, because the construction of I.D. numbers involves considerable condensation of the structural information, it is interesting to observe that no such structures have been detected among a relatively sizable group of closely structurally related systems. In fact, it is suprising that no such structures have been detected among classes of compunds hitherto examined. While a hunt for such nonisomorphic structures that will have the same I.D. number continues, the finding of such graphs needs not dramatically upset the use of I.D. numbers and needs not cause confusion. First, there will be few such instances; second and more importantly, because the approach is strictly structural, we may succeed in characterizing such special situations, which will no doubt have definite structural cause. Finally, once the situation is recognized, some remedy may be possible, because, as will be seen, the approach of weighted path counts has flexibilities that can be explored, starting with modifications in the weighting factors used. Before we continue with an outline of the approach, let us mention that one can consider additional conditions and constraints on the codes, as well as one may feel that some of those suggested are of a lesser importance. A referee[12] suggested that identification numbers need to show minimal overlap between I.D. ranges for different classes of compounds, because in the opposite case the retrieval of compounds will be hardly possible and the coding system will represent a filter of low efficiency. As we will see later, our system satisfies this additional requirement too. One expects that condensation of structural information to a single number ought to be accompanied by considerable loss of information. What is remarkable is how much of the essential information has not been lost! In that respect, I.D. numbers can be viewed as an additional topological index, an index that combines some features of the connectivity index and some features of the total path counts in a structure. Hence, the importance of molecular I.D. numbers may be in SAR (structure–activity relationship) and in organizing molecular data, while their use as real identification numbers needs further research.

TOPOLOGICAL INDEXES

In order to appreciate the problem of deriving a compact structural parameter for representing a structure, we will briefly review the topic of topological indexes. The task is to represent each structure by a single parameter, yet to maintain as much as possible a high discriminatory power. The purpose of topological indexes is (a) to classify structures and (b) to serve for structure–property correlations. They failed to be unique, although different topological indexes show different degrees of discrimination among closely related structures.[13] Platt's and Wiener's pioneering work[14] demonstrated that graph invariants, paths in particular, provide a useful basis for such correlations. Hosoya's Z index illustrates multiple use of such indexes: designed for classification of structures and found useful for correlation of thermodynamical properties of hydrocarbons.[15] The connectivity index,[16] which is based on differentiation of bond types (vide infra) and uses different weight factors for different bonds, appears to be most successful in producing structure–property correlations and regressions, particularly when suitably expanded to account for differentiation among heteroatoms.[17] Because a number of properties do not correlate among themselves, a single index cannot equally satisfactorily portray such diverse features. An example is boiling points vs. octane numbers: the connectivity index well correlates with the boiling points for alkanes, while the centric index[18] better represents octane numbers. One should hardly expect that yet another, undiscovered, index may be superior to the two indexes mentioned with respect to *both* properties, because the relative magnitudes of the boiling points and the octane numbers vary. For instance, compare the data for 2-methyl and 3-methyl alkanes: the boiling points for the latter (and the connectivity indexes) are greater; the octane numbers (and the centric indexes) are smaller. While topological indexes are useful for correlating structure and property, and for classification, they failed to provide unique characterizations for a structure. The question is if such expectation is realistic at all. Nonuniqueness is not necessarily a disadvantage when properties and classification are the prime target, because there are compounds with similar (or practically *same*) properties, which requires then a similar or a same index. However, interest continues in trying to devise an index that would be unique. Bonchev, Mekenyan, and Trinajstić[19] considered a "superindex", a collection of a dozen indexes, as a surrogate approach. While such a collection has been shown to have a greater discriminatory power than any of the individual component, the approach avoided the issue of construction of a *single* structural parameter. Among numerous topological indexes suggested in the past, the recently proposed index of Balaban[20] has the greatest discriminatory or selectivity power. The index J, as it is called by Balaban, can be viewed as a generalization of the connectivity index. Instead of being based on the adjacency matrix and differentiation of bond types, it is derived in an analogous way from the distance matrix. Very recent comparative study of several topological indexes by Razinger, Chretien, and Dubois[21] confirmed high selectivity of the J index, which was able to differentiate all alkanes with 11 and less carbon atoms, and among 355 dodecanes, there were only four duplicate indexes. Can an index do even better and yet maintain simple structural origin remains an open question.

STRUCTURAL SEQUENCES

An alternative to a single index is collection of indexes. If elements of a collection can be *ordered*, where the ordering rule has structural origin, we obtain structural sequences. Clearly, the sequences have greater structural content, at the cost of involving a larger number of data. Paths (i.e., self-avoiding walks) have been found very useful in structure–property studies: Wiener numbers P and W (paths of length three and the number of all paths, respectively) suffice for accurate correlations of many compounds (not alkanes alone).[12] The numbers P and W do not form a sequence, rather just a collection. First short sequence suggested for study of structure–activity relationships consists of paths of length two and three (p_2 and p_3, respectively).[22] Their use has led to organization of data on isomers into "Periodic Table of Isomers".[23] The count of all paths of different length in a molecule (expressed as a sequence: P_1, P_2, P_3, ..., P_k) serves as a molecular signature and characterization and allows one to make comparison of structures and to measure the degree of structural similarity at a quantitative level.[24] Such approach has been applied to several structure–activity studies in which a few selected compounds, identified as standards, allowed ranking of the remaining structures, thus suggesting which among numerous possibilities can be expected to show similar properties and even surpass the standards in some desired qualities.[25] Finally, path sequences P_1, P_2, P_3, ..., P_k can serve as a basis for nonempirical pattern recognition,[26] technique in which one aims at clustering of compounds in smaller lots of closely related ones or even attempts to have maximal discrimination among the structures. Empirical pattern recognitions and clustering schemes may employ a large number of parameters, most or all of which are devoid of any structural interpretation. Here, in contrast, selected graph invariants (paths of different length) appear capable of achieving a great degree of clustering without introducing even a single empirical parameter! The situation well illustrates the difference between the graph theoretical approach and the empirical "curve-fitting" approch characterizing conventional schemes. Of course, each approach has some advantages, and frequently, one complements the other. Empirical schemes represent an unbaised approach, an approach without preconceived concepts on model or mechanisms involved, and they can point to the dimensionality (i.e., degree of independence of the collection of data). Graph theoretical sequences presuppose the importance of selected structural invariants and can at best show the degree of success of such underlying assumptions. Hence, the two approaches can be combined, each contributing to the partial answer of complex structure–activity studies. Clearly, we are at the beginning of studies of structure–activity relationships, and one can expect other structural invariants to be examined, uncovered, and elaborated. For instance, random walks may be of use in characterization of immediate atomic environment.[27] Here, we wish to consider *condensation* of structural information, the question of how few parameters can suffice to characterize a structure. What can a single parameter do in representation of a structure? The results on topological indexes and on simple sequences are so much encouraging that one ought to further examine possibilities of simple constructions, short sequences, and even a single parameter. Molecular I.D. numbers appear promising as novel structural characterizations for molecules. We will see that they surpass all previous constructions in containing a considerable amount of structural information and appear to have also unprecedent selectivity characteristics for a single datum for a structure.

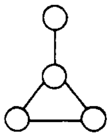
Table II. Output of ALL-PATH Program for a Simple Graph[a]

	0 1 0 0	
	1 0 1 1	
	0 1 0 1	
	0 1 1 0	
vertex	atomic path sequence	sum
1	1 1 2 2	6
2	1 3 2 0	6
3	1 2 3 1	7
4	1 2 3 1	7
	4 4 5 2	
	tot no. of paths is 15	

[a] Adjacency matrix serves to verify input. Vertexes were labeled 1–4, and for each, atomic path sequences are shown and the sum of all atomic paths. The last two lines give information for a molecule (graph) as a whole.

OUTLINE OF THE APPROACH

We will now try to combine the two successful approaches for characterization of molecules: the path numbers and the connectivity index. We want to see if, in this way, we can arrive at modified path sequences that can even better characterize structural features of molecules. Enumeration of paths leads to the following four outcomes: (a) sequence of path numbers for individual atoms; (b) sequence of path numbers for a molecule as a whole; (c) total number of paths for an individual atom; (d) total number of paths for a molecule as a whole. In Table II we show the computer output of the ALL-PATH program for the simple graph



which shows all four types of path numbers. The adjacency matrix serve the purpose of verifying the input, which is specified as the list of neighbors. The first item under the short dashed line gives the label for each vertex. The row that follows represents atomic path numbers a_0, a_1, a_2, ...; a_0 is always 1. The last sequence represents molecular path numbers m_0, m_1, m_2, ...; m_0 is always n, the number of vertexes in the molecular graph. Atomic path sums follow atomic sequences and are shown in the last column as the corresponding sum while the total number of paths in a molecule is shown in the last line of the computer output. Atomic path sums represent condensation of atomic path numbers, a condensation of a sequence to a single entry. Necessarily, such a step is accompanied by loss of information, but the individual atomic path sums still show regularities and variations that reflects some structural content: for instance, all atoms on a pending fragment attached to a polycyclic frame have the same sum.[28] The so-derived numbers can be viewed, as Seybold pointed out,[29] as an index of "compactness" or centrality of the particular site. The numbers can be related to chemical activity of atoms and eventually correlated with such properties as carcinogenity.[29] The sum of all atomic path sums gives the molecular path sum, the number that appeared in Wiener's analysis, W, which Platt tried to relate to molecular volume. By outlined condensation of *integers*, it ought not to be suprising that numerous nonisomorphic structures may result in having the same W, i.e., that degeneracy (duplicate outcomes), particularly as the size of structures increases, will be a rule rather than exception. The origin of the degeneracy in Hosoya's Z index is of the same kind: condensation of integers into a single entry. In contrast, the origin of the degeneracy of the connectivity index is in the occurrence of the same bond types with the same frequency in some nonisomorphic structures. In other words, structures have been

Table III. Bond Weights Used[a]

bond type	atomic factors	bond weight
(1, 1)	(1)(1)	1.0000
(1, 2)	$(1)\,(^1/_2)^{1/2}$	0.7071
(1, 3)	$(1)\,(^1/_3)^{1/2}$	0.5774
(1, 4)	$(1)\,(^1/_4)^{1/2}$	0.5000
(2, 2)	$(^1/_2)^{1/2}(^1/_2)^{1/2}$	0.5000
(2, 3)	$(^1/_2)^{1/2}(^1/_3)^{1/2}$	0.4083
(2, 4)	$(^1/_2)^{1/2}(^1/_4)^{1/2}$	0.3536
(3, 3)	$(^1/_3)^{1/2}(^1/_3)^{1/2}$	0.3333
(3, 4)	$(^1/_3)^{1/2}(^1/_4)^{1/2}$	0.2887
(4, 4)	$(^1/_4)^{1/2}(^1/_4)^{1/2}$	0.2500
	Bonds of Higher Valency	
(1, 5)	$(1)\,(^1/_5)^{1/2}$	0.4472
(2, 5)	$(^1/_2)^{1/2}(^1/_5)^{1/2}$	0.3162
(3, 5)	$(^1/_3)^{1/2}(^1/_5)^{1/2}$	0.2582
(4, 5)	$(^1/_4)^{1/2}(^1/_5)^{1/2}$	0.2236
(5, 5)	$(^1/_5)^{1/2}(^1/_5)^{1/2}$	0.2000

[a] Only four digits are shown; all the calculations were made by single precision arithmetics, i.e., eight to nine digits.

broken down into too small of fragments, and different structures show occasionally the same number of fragments of each kind. Bond types imply differentiation of the immediate environment of each bond; path counts preserve some information on more distant neighbors. Path counts, in fact, keep track of sequential distribution of bonds, its limitation being that the outcome is given as an integer; the connectivity index loses the information on the serial distribution of bonds but compensates with noninteger weights. If we combine the two approaches, i.e., use weighted bond types and then count the paths, we may reduce, if not eliminate totally, the degeneracy due to both ingredients.

Indeed, this is what happens. Our scheme therefore consists of (1) use of weighted bond types followed by (2) enumeration of all paths in a molecular graph so weighted. The weights selected for each bond type are shown in Table III; bond types are defined by the valency of the terminal vertices defining a bond. Bond (m, n), in which one end atom has m nearest neighbors while the other end atom has n nearest neighbors, is assigned weighting factor $(m, n)^{-1/2}$—this is precisely the weighting scheme used in defining the connectivity index χ.[16] The available ALL-PATH program (the version designed for study of graphs with multiple connections)[9b] allows one to derive weighted path counts simply by using the entries in Table III as part of the input. After one indicates the pair of vertices that are connected, the input requires a multiplicity number, which is 1 for single bond, 2 for double bond, and 3 for triple bond. However, the program allows fractional inputs as well, like 1.5 for an aromatic bond. Use of nonintegral bond weights has been illustrated in reference 9b on selected benzenoid systems using Pauling bond orders as the weights for the aromatic CC bond. A consequence of using weighting factors smaller than 1 is gradual attenuation of the role of paths of longer length, i.e., giving dominance to local features, environments of a few bonds around. This is a natural and more satisfactory limitation of the role of distant parts on a local scene than would be an abrupt truncation of the path codes. Hence, a summation of suitably weighted paths will diminsh the role of distant neighbors. The total sum of all paths can be alternatively viewed as an average atomic path sum (when properly normalized, i.e., divided by n, the number of atoms).[30] Therefore, the weighted molecular path sum represents an average atomic path count in which local features are more pronounced. Clearly, other weighting schemes are possible that can even more accentuate local character vs. global character of unweighted path sums.

In Table IV, the output of a *modified*[31] ALL-PATH program is shown for the simple graph already considered in Table II. The modification consists in a subroutine that, given input that consists in the list of neighbors, finds bond types for all connections and inserts the corresponding weighting factors in the adjacency matrix, which is subsequently used in the counting algorithm. In this way, the weights are automatically absorbed in enumeration of the paths. The program has been written in BASIC and is suitable for personal computers (such as an Apple IIe, on which all the computations reported here were performed). The output of the modified ALL-PATH program should be compared to the output of the simple ALL-PATH program shown in Table II: again, individual rows, except the last two, correspond to data for individual atoms, while the last two rows summarize the same information for a molecule as a whole. The total number of paths—here 7.17586859—is the number that we propose to use as the molecular I.D.

Table IV. Computer Print of ALL-PATH Program Modified to Include Weighing Factors for Individual Bonds for the Simple Graph of Table II

```
]
                    CONNECTION MATRIX

0 1 0 0

1 0 1 1

0 1 0 1

0 1 1 0




-----
1

1                   .577350269       .471404521       .23570226
2.28445705
-----
2

1                   1.39384685       .408248291       0
2.80209514
-----
3

1                   .908248291       .606493073       .11785113
2.6325925
-----
4

1                   .908248291       .606493073       .11785113
2.6325925
----------
4                   1.89384685       1.04631948       .23570226

TOTAL NUMBER OF PATHS: 7.17586859
```

The present scheme appears promising on count of it (1) being easy to derive, involving simple and basic concepts such as total number of paths and differentiation of bond types, (2) being of high discriminatory power, although uniqueness remains an open question most likely to be resolved by first encounter of a counterexample and (3) having definite structural origin, thus possibly being used for comparison of structures, classification, and ordering, including structure–activity research. We have already indicated conceptual and computational simplicity of the proposed molecular I.D. numbers. By computational simplicity, we mean here simple algebra behind the computation of the weighted paths, not to be confused with the computational complexity mentioned earlier that concerns the growth of the number of operations to be made with increasing the size of the structure. In Table V we fully disclose the involved algebra for enumeration of weighted paths for the simple graph of Tables II and IV. Paths are first listed by using adopted labeling of atoms (ALL-PATH program has as an option possibility the ability list all paths explicitely as we have done in Table V). We have the following bond types: a–b = (1, 3); b–c and b–d = (2, 3); c–c = (2, 2). Paths of length 1 are simply given by summing the corresponding weighting factors. For other paths, one has to

Table V. List of Paths of Different Length for Each Vertex of the Simple Graph of Table II Illustrating Computational Steps in Deriving Weighted Path Numbers

	paths	length	weight
vertex a	a	0	1
	a–b	1	$1/\sqrt{3}$
	a–b–c	2	$(1/\sqrt{3})(1/\sqrt{6})$
	a–b–d	2	$(1/\sqrt{3})(1/\sqrt{6})$
	a–b–c–d	3	$(1/\sqrt{3})(1/\sqrt{6})(^1/_2)$
	a–b–d–c	3	$(1/\sqrt{3})(1/\sqrt{6})(^1/_2)$
vertex b	b	0	1
	b–a	1	$1/\sqrt{3}$
	b–c	1	$1/\sqrt{6}$
	b–d	1	$1/\sqrt{6}$
	b–c–d	2	$(1/\sqrt{6})(^1/_2)$
	b–d–c	2	$(1/\sqrt{6})(^1/_2)$
vertexes c and d	c	0	1
	c–b	1	$1/\sqrt{6}$
	c–d	1	$^1/_2$
	c–b–a	2	$(1/\sqrt{6})(1/\sqrt{3})$
	c–b–d	2	$(1/\sqrt{6})(1/\sqrt{6})$
	c–d–b	2	$(^1/_2)(1/\sqrt{6})$
	c–d–b–a	3	$(^1/_2)(1/\sqrt{6})(1/\sqrt{3})$
	Sequences		
vertex a	1; 1; 2; 2		1.00000000; 0.57735026; 0.47140452; 0.2350226
vertex b	1; 3; 2		1.00000000; 1.3938467; 0.40824829
vertex c	1; 2; 3; 1		1.00000000; 0.90824829; 0.60649296; 0.11785113

multiply weighting factors for each consecutive bond added in the path. Thus, for path a–b–c, we multiply factors associated with a–b and b–c, i.e., factors for bond types (1, 3) and (2, 3). A similar approach has been considered in introducing the connectivity series concept[32] although there a different rule for combining valency numbers has been adopted. From Table V we see that for relatively simple structures one can derive the molecular I.D. number by hand, but the number of paths proliferates fast with the size of a structure, and the count of paths ought to be delegated to a computer. In the following sections, we will illustrate molecular I.D. numbers for various classes of structurs, discuss uniqueness of the I.D. numbers, and point to their potential use in structure–activity work by illuminating some of their structural properties.

ARE THE I.D. NUMBERS UNIQUE?

We have examined over 400 structures and have not detected a duplicate code. Four hundred is small number of structures, but they have been selected from structurally related compounds that may be expected, in view of apparent similarity, to show the same path-weighted counts. The families of structures and graphs that have been examined are: (a) acyclic alkane molecular graphs up to $n = 10$ (149 cases), (b) monocyclic graphs up to $n = 8$ (122 cases), (c) bicyclic graphs up to $n = 7$ (79 cases), (d) all graphs on five vertexes, (e) sesquiterpenes (30 cases), (f) miscellaneous polycyclic, and (g) selected counterexamples to other schemes. Some of the above structures overlap the simple classification intended to illustrate variations in the structure considered.

In order to illustrate variations in the novel I.D. numbers in Table VI, we list the numbers for all five-vertex graphs (Figure 1). The smallest I.D. number 8.5 corresponds to the carbon skeleton of neopentane, and the complete graph on five vertexes, K_5, has the largest number of paths for its size and the largest I.D. of 10.5468. I.D.'s are real numbers, and immediately, one has to decide how many digits to include in reporting and using such numbers. Graphs in Table VI already differ in the second or third digit after the decimal point, and for discussion of such compounds, one needs not to use an extended number of digits. Razinger, Chretien, and Dubois[21] have shown that the use of four digits may be too few when real-numver topological indexes are examined for their structural selectivity, while the use of eight decimal places appears sufficient. It seems therefore prudent to follow their finding and keep all the digits that a single precision calculation on computers offers. The trend in Table VI points to an increase in the I.D. numbers as the number of cycles increases, but the dependence is not as simple as reflected by cases in which a structure with fewer rings comes after a structure with more rings. Roughly speaking, the I.D. measures the "complexity" of a graph.[33] Among the graphs of Table VI, we find the smallest pair of nonisomorphic graphs with the same *distance degree sequence* (graphs belong to carbon skeletons of bicyclo[1.1.1]pentane and bicyclo[2.1.0]pentane). This illustrates the discriminatory power of the new codes, which succedded where distance degree sequences fail.

Table VI. Numbers for Graphs Having $n = 5$ Vertexes (Shown in Figure 1)

graph	I.D.	graph	I.D.
1	8.5	12	9.82715413
2	8.69680194	13	9.83823053
3	8.84987374	14	9.9128876
4	8.99632035	15	9.932653
5	9.11591895	16	10.1224553
6	9.14899514	17	10.1283511
7	9.41451053	18	9.49632035
8	9.45469847	19	10.2851529
9	9.51759609	20	10.3830119
10	9.6875	21	10.546875
11	9.70373865		

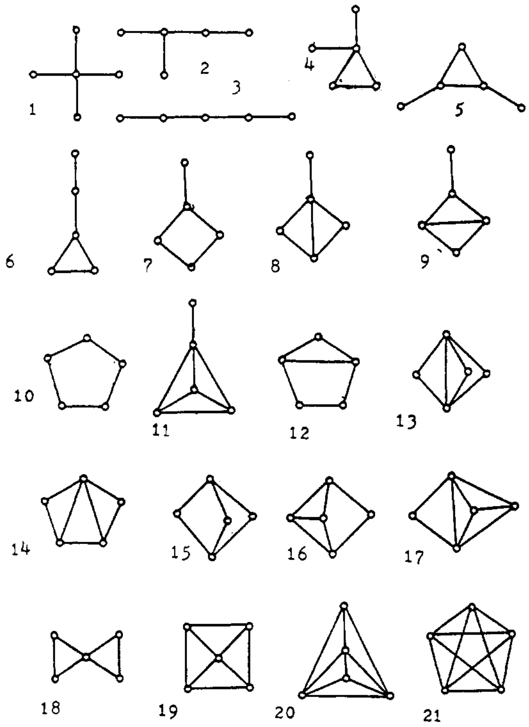


Figure 1. Graphs having $n = 5$ vertexes.

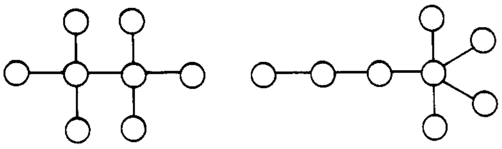
COUNTEREXAMPLES OF OTHER SCHEMES

In order to assess the selectivity power of the proposed I.D. numbers in the following, we will only discuss cases that some other scheme could not differentiate. In Table VII, we list examples of (a) graphs having the same connectivity index, (b) graphs having the same Hosoya's Z index (c) isospectral graphs having the same characteristic polynomial, (d) isocodal graphs having same path sequence, (e) graphs having the same Balaban's J index, (f) graphs having the same distance sequence, and, finally, (g) graphs having a same distance degree

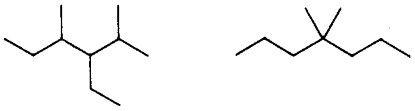
Table VII. I.D. Numbers for Graphs Having the Same Connectivity Indexes (22–23, 24–25), the Same Hosoya's Z Index (26–27, 28–29, 30–31), Isospectral Graphs (32–33, 34–35), Isocodal Codes Based on Path Sequences (36–37, 38–39), the Same Balaban's J Index (40–41, 42–43, 44–45, 46–47, 48–49, 50–51), the Same Distance Sequence (52–53), and the Same Collection of All Vertex Path Sequences (54–55)

graph	I.D.	graph	I.D.
22	14.5011376	39	18.2695267
23	14.4916058	40	22.084102
24	14.6601375	41	22.088966
25	14.658474	42	22.0805576
26	14.440483	43	22.0209309
27	14.2716797	44	22.0808422
28	14.5019207	45	22.0239752
29	14.4310606	46	22.3250773
30	14.4960414	47	22.2707371
31	14.4301407	48	22.3289495
32	20.3310016	49	22.2752
33	20.3129466	50	22.1597477
34	20.0818978	51	22.1052347
35	20.0819973	52	18.0230497
36	16.3223013	53	18.0063762
37	16.4238591	54	33.1242956
38	18.2558656	55	33.1196609

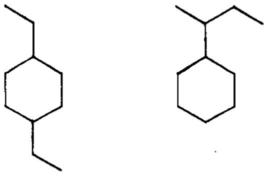
sequence. In all cases, we see that corresponding molecular I.D. numbers are different. A same connectivity index occurs already among pairs of octanes. The same is also true for Hosoya's Z index, if we disregard cases of molecules having a different number of atoms (such as propane and neopentane carbon skeleton graphs). Next, we will consider isospectral graphs. Again, the smallest pair of nonisomorphic *acyclic* isospectral graphs has eight vertexes:



The corresponding I.D. numbers will necessarily be different, because they involve different irrational numbers: $1/\sqrt{5}$ appearing for one of the graph only. If we confine attention only to graphs representing carbon skeletons of alkanes, the smallest isospectral pair is graphs of 2,4-dimethyl-3-ethylpentane and 4,4-dimethylheptane:



with I.D. numbers 16.3335319 and 16.4220982, respectively. Isospectral decane isomers are listed in Table VII. The pool of isospectral graphs has expanded considerably[34] since the early work of Collatz and Sinogowitz,[35] who were first to recognize this class of structures. However, one needs not consider many of these, examples like[36]



because they have different bond types and will involve different weighting factors. There is, for instance, no (3, 3) bond type in 1,4-divinylbenzene.

The next class of graphs having potential counterexample to our own scheme is graphs with the same path enumeration. In Table VII we illustrate I.D. numbers for nonane and decane isomers having the same path sequences: 8, 10, 10, 6, 2 for 2,3,4-trimethylhexane and 3,3-dimethylhexane and 9, 12, 11, 9, 4 common to 2,4-dimethyl-4-ethylhexane and 2,2-dimethyl-3-ethylhexane. The corresponding I.D. numbers in each case already differ at the first decimal position! Hence, in order to have the same or similar I.D. numbers, a pair of structures has to possess considerable structural coherence. Balaban's J. index, presently the most selective topological index, may point to some such structures. A number of such graphs are illustrated in Figure 2; some have, as can be seen, the same bond types. Again, we do not find any instance of duplicate I.D. numbers. Finally, we have examined few graphs having the same distance sequence d_0, d_1, d_2, ... and two graphs having the same distance degree sequence s_0, s_1, s_2, ...; both groups are illustrated in Figure 2. The later pair of graphs is particularly interesting as they have the same collection of atomic path sequences a_0, a_1, a_2, ..., a_k. The graphs have been constructed by Slater[37] as a counterexample to a conjecture proposed by this author[38] that, for acyclic graphs, lists of atomic codes (i.e., the collection of all atomic path sequences) may be unique. The conjecture has been verified by Shelly and Trulson[39] for alkane graphs up to 14 carbon atoms, but Slater was able to construct graphs on 18 vertices (maximal degree 5; hence, of no immediate chemical interest but nevertheless fatal for the conjecture), which have considerable internal similarity and which produce the same sequences of atomic codes. As one sees, the graphs differ in arrangement of fragments around the central bond. The two graphs have the same bond types and the same path sequences for individual atoms and thus represent the most serious threat to producing duplicate I.D. numbers. However, we find the I.D.'s different, the difference occurring in the third decimal place—hence our approach passed this rather important test.

STRUCTURAL PROPERTIES OF I.D. NUMBERS

In Table VIII we list I.D. numbers for alkanes up to 10 carbon atoms, lexically ordered. Structures are represented by an abbreviated code indicating methyl (M), ethyl (E), propyl (P), and isopropyl (I) groups using standard chemical numbering along the main chain. One can see a definite regularity in the ordering of isomers: first (smaller I.D numbers) more branched isomers appear and later those with fewer pending bonds. Moreover, for the same residual (such as a dimethyl-substituted group), a more centrally substituted structure has a lower I.D. Similarly, if two groups of substitutions are present, isomers with adjacent substitution sites have a smaller I.D. Such observations allow one to narrow intervals on I.D. numbers for which certain structural characteristics appear. For example, in the case of nonane isomers we find the following subsclassification:

I.D. interval	structural characteristics
16.04–16.10	tetrasubstituted
16.25–16.27	trisubstituted with quaternary carbon
16.32–16.33	trisubstituted (at three different sites)
16.42–16.43	disubstituted with quaternary carbon
16.48–16.49	disubstituted (at two different sites)
16.65–16.66	monosubstituted
16.82	unsubstituted

Observe the somewhat larger gaps between tetra-, tri-, di-, mono-, and unsubstituted cases and a finer subdivision of each of tri- and disubstituted cases depending on the number of sites involved. Hence, the molecular I.D. numbers can serve not only as structural codes for individual molecules but also as a simple basis for *classification* of structurally related compounds. Molecules of similar structural composition display similar I.D. numbers! Thus, I.D. numbers appear to preserve some important structural characteristics remarkably well for

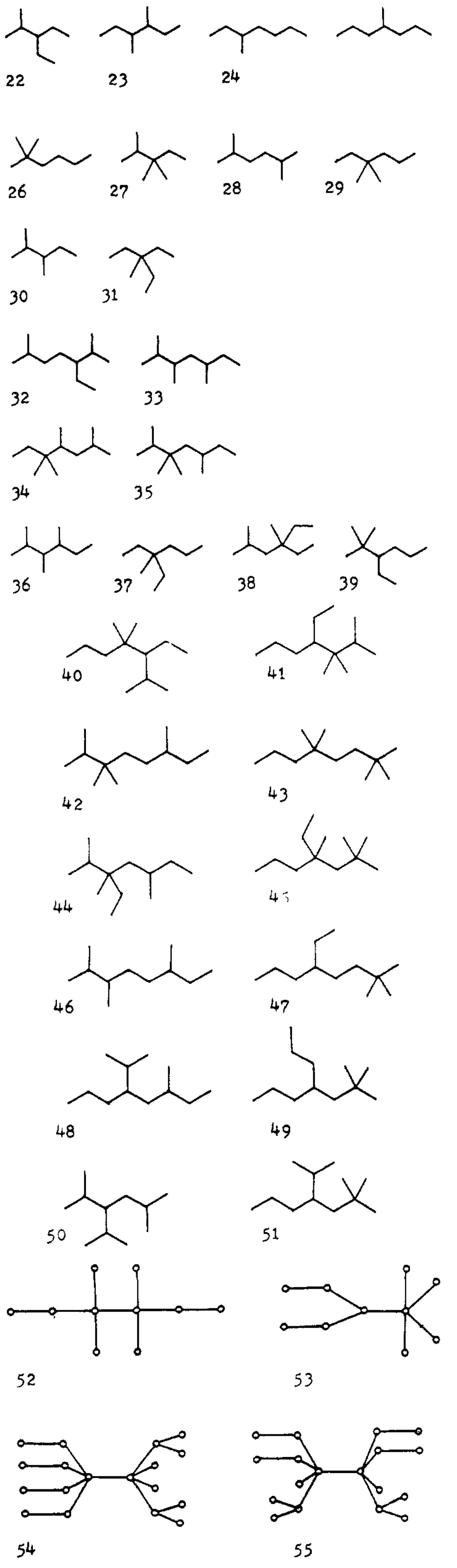


Figure 2. Acyclic graphs having some same property (having same connectivity index, having same Hosoya's Z index, being isospectral, having same path sequences, having same Balaban's J index, having same distance sequence, and having same distance degree sequence).

a difficult proposition of supressing a structure to a single number.

Cyclic structure represent a more challenging group. Many properties and procedures valid for acyclic systems no longer hold for cyclic and polycyclic structures. It is therefore of interest ot see whether the observed regularities for I.D. numbers for acyclic graphs will give analogous observations.

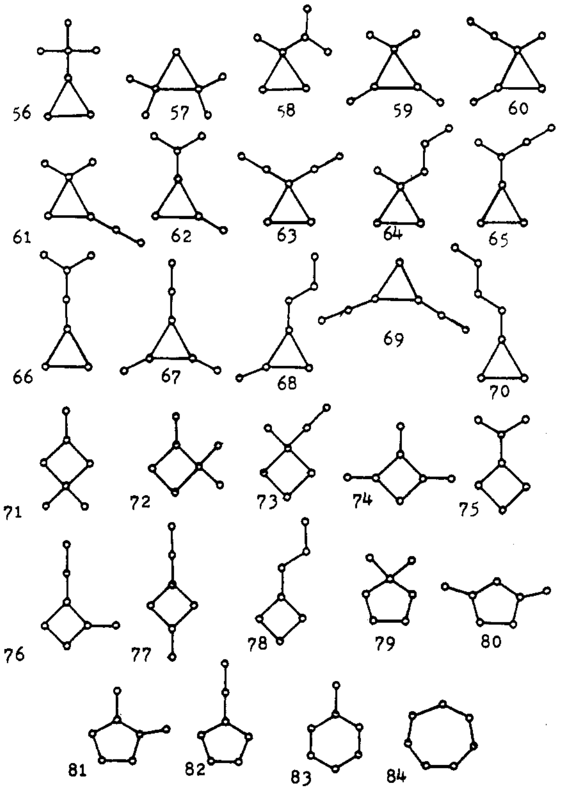


Figure 3. Monocyclic graphs having n = 7 vertexes.

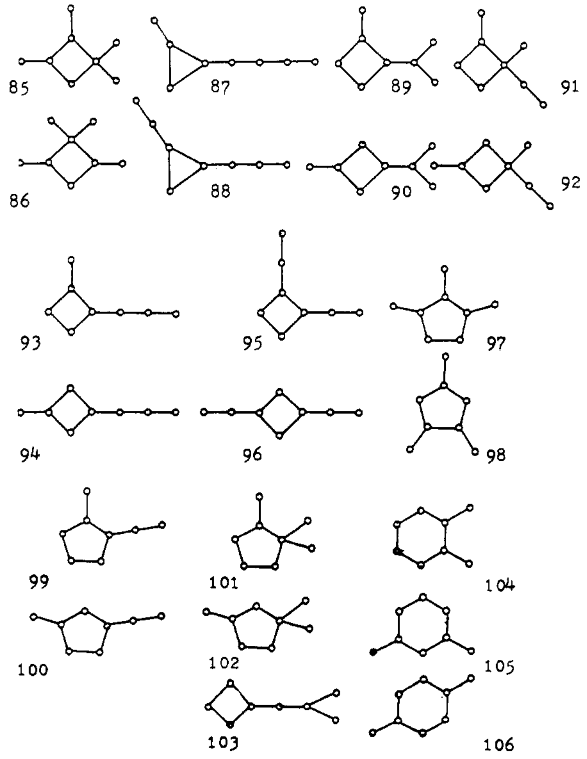


Figure 4. Selected monocyclic graphs having n = 8 vertexes and having similar I.D. values (pairwise, except for the last two triplets).

We have examined monocyclic graphs having n = 7 and n = 8 vertexes (Figure 3 and Table IX). By inspection, one can verify that with an increase in the size of the ring the I.D. number also increases, that longer chains lead to larger I.D.'s, and that substitution at more sites again produces, for the same type of the substitutent, greater I.D. values. One can even among substituents of a same size (e.g., isopropyl and n-butyl) discern heierarchical order, but clearly, more and larger

Table VIII. I.D. Numbers for Alkanes from *n* = 2 (Ethane) to *n* = 10 (Decanes)[a]

alkane	substitution	I.D.	alkane	substitution	I.D.	alkane	substitution	I.D.
C_2		1	C_9	2,2-M_2-3-E	16.2713728	C_{10}	2,4-M_2-4-E	18.2558656
C_3		4.91421356	C_9	2,2,5-M_3	16.2720936	C_{10}	3,3-M_2-4-E	18.2571464
C_4	2-M	6.73205081	C_9	2,3,4-M_3	16.3223013	C_{10}	2,2,4-M_3	18.2605982
C_4	n	6.87132035	C_9	2,3,5-M_3	16.3275	C_{10}	2-M-3,3-E_2	18.2610989
C_5	2,2-M_2	8.5	C_9	2,4-M_2-3-E	16.3335319	C_{10}	2,2,3-M_3	18.2615726
C_5	2-M	8.69680194	C_9	4,4-M_2	16.4220982	C_{10}	2,2,5-M_3	18.2626061
C_5	n	8.84987374	C_9	3-M-3-E	16.4238591	C_{10}	2,2,6-M_3	18.2686
C_6	2,2-M_2	10.4659903	C_9	3,3-M_2	16.425239	C_{10}	2,2-M_2-4-E	18.2687351
C_6	2,3-M_2	10.5236459	C_9	3,3-E_2	16.4283009	C_{10}	2,2-M_2-3-E	18.2695267
C_6	3-M	10.6758508	C_9	2,2-M_2	16.4362317	C_{10}	3,4,5-M_3	18.3111022
C_6	2-M	10.6791775	C_9	3,4-M_2	16.4847872	C_{10}	2,3,4-M_3	18.3159074
C_6	n	10.8391504	C_9	3,5-M_2	16.486489	C_{10}	2,4,5-M_3	18.3166979
C_7	2,2,3-M_3	12.293055	C_9	2-M-4-E	16.4984038	C_{10}	2,3,5-M_3	18.3178068
C_7	3,3-M_2	12.4427038	C_9	2,4-M_2	16.4903701	C_{10}	2,4,6-M_3	18.3223797
C_7	2,2-M_2	12.4489854	C_9	2,3-M_2	16.4914406	C_{10}	2,3,6-M_3	18.3237466
C_7	2,3-M_2	12.5052429	C_9	3-M-4-E	16.4922664	C_{10}	2,3-M_2-4-E	18.3238113
C_7	2,4-M_2	12.5091624	C_9	2,5-M_2	16.4923299	C_{10}	3-M-4-I	18.3255101
C_7	3-M	12.66	C_9	2,6-M_2	16.4982999	C_{10}	2,5-M_2-3-E	18.331845
C_7	3-E	12.6691974	C_9	2-M-3-E	16.499085	C_{10}	2-M-3-I	18.3327533
C_7	2-M	12.6703653	C_9	4-M	16.6550235	C_{10}	4,4-M_2	18.417617
C_7	n	12.8337888	C_9	3-M	16.6575186	C_{10}	4-M-4-E	18.4202584
C_8	2,2,3,3-M_4	14.0625	C_9	2-M	16.6637561	C_{10}	3-M-3-E	18.4207183
C_8	2,3,3-M_3	14.2716797	C_9	3-E	16.6642074	C_{10}	3,3-M_2	18.4223282
C_8	2,2,3-M_3	14.2750651	C_9	4-E	16.6661184	C_{10}	3,3-E_2	18.427381
C_8	2,2,4-M_3	14.2790808	C_9	n	16.8297675	C_{10}	2,2-M_2	18.4341061
C_8	2,3,4-M_3	14.335089	C_{10}	2,2,3,3,4-M_5	17.8708061	C_{10}	3,4-M_2	18.4813779
C_8	3-M-3-E	14.4301407	C_{10}	2,2,3,4,4-M_5	17.8748926	C_{10}	3,5-M_2	18.4814164
C_8	3,3-M_2	14.4310606	C_{10}	2,2,3-M_3-3-E	18.0310922	C_{10}	2,4-M_2	18.4872381
C_8	2,2-M_2	14.440483	C_{10}	2,2,3,3-M_4	18.0309016	C_{10}	2,5-M_2	18.4875345
C_8	3,4-M_2	14.4916058	C_{10}	2,2,4,4-M_4	18.0322026	C_{10}	2,3-M_2	18.4891402
C_8	2,3-M_2	14.4960414	C_{10}	2,2,5,5-M_4	18.0422754	C_{10}	2,6-M_2	18.4901778
C_8	2,4-M_2	14.4966342	C_{10}	2,3,3,4-M_4	18.0872176	C_{10}	3-M-5-E	18.4906415
C_8	2-M-3-E	14.5011376	C_{10}	2,3,4,4-M_4	18.087821	C_{10}	3-M-4-E	18.4925967
C_8	2,5-M_2	14.5019207	C_{10}	2,3,3,5-M_4	18.0922132	C_{10}	2,7-M_2	18.4964894
C_8	4-M	14.658474	C_{10}	2,3-M_2-3-I	18.0922552	C_{10}	2-M-5-E	18.4970369
C_8	3-M	14.6601375	C_{10}	2,2,3,4-M_4	18.0923348	C_{10}	3,4-E_2	18.4976929
C_8	2-M	14.665992	C_{10}	2,2,3,5-M_4	18.097593	C_{10}	2-M-3-E	18.4980586
C_8	3-E	14.6658707	C_{10}	2,2,4,5-M_4	18.0976891	C_{10}	2-M-4-E	18.4992885
C_8	n	14.831108	C_{10}	2,2,4-M_3-3-E	18.1038407	C_{10}	4-I	18.5006067
C_9	2,2,3,3-M_4	16.0414344	C_{10}	2,3,4,5-M_4	18.1531482	C_{10}	5-M	18.6524666
C_9	2,2,4,4-M_4	16.049017	C_{10}	2,4-M_2-3-I	18.1663803	C_{10}	4-M	18.6532983
C_9	2,3,3,4-M_4	16.1009962	C_{10}	3,4,4-M_3	18.2483637	C_{10}	3-M	18.6562091
C_9	2,2,3,4-M_4	16.1049849	C_{10}	3,3,4-M_3	18.2500564	C_{10}	2-M	18.6626546
C_9	3,3,4-M_3	16.2572642	C_{10}	3,4-M_2-3-E	18.2501865	C_{10}	3-E	18.6633757
C_9	2,3,3-M_3	16.260992	C_{10}	3,3,5-M_3	18.251577	C_{10}	4-E	18.6662423
C_9	2,3-M_2-3-E	16.2610276	C_{10}	2,4,4-M_3	18.2536269	C_{10}	4-P	18.6699406
C_9	2,4,4-M_3	16.2621115	C_{10}	2,3,3-M_3	18.2556482	C_{10}	n	18.8290973
C_9	2,2,3-M_3	16.2660701	C_{10}	2,3-M_2-3-E	18.2557016			
C_9	2,2,4-M_3	16.2667591	C_{10}	2,5,5-M_3	18.2573274			

[a] Standard chemical numbering assumed and substitutions along the main chain indicated with M (methyl), E (ethyl), P (propyl), and I (isopropyl).

structures ought to be examined before attempting to formulate the rules.

Monocyclic structures with *n* = 8 vertexes offer more comparisons. In Figure 4, we have collected the most similar pairs of I.D. numbers and have shown the corresponding molecular graphs. In all cases indeed, we find that the corresponding graphs are rather similar. Thus, the question of molecular similarity, which is beginning to receive some attention in the literature,[40] can be approached also with use of molecular I.D. numbers—at least Figure 4 is a promising sign in that direction.

Bicyclic structures introduce novel comparisons and novel regularities. We see that structures with more equilized branches (e.g., a bicyclo[2.1.1] system as compared to [2.2.0] or [3.1.0] systems) have higher I.D. numbers. In addition, a substitution at the carbon atom of lower valency (secondary vs. tertiary carbon) gives a higher I.D., spiro compounds (isomers) that involve a larger ring size have a greater I.D., etc. Such regularities are an asset when one becomes interested in structure–activity work. Observe how I.D. numbers can filter out a fraction of compounds having special structural features. For example, bicyclic structures of Figure 5 (Table X) having six vertexes and displaying various ring sizes can be classified by their ring types as suggested in the *Ring Index*:[41]

ring type	I.D. range
(3, 3)	11.35–11.40
(3, 4)	11.70–11.79
(3, 5), (4, 4), (4, 5)	11.80 and higher

Unusual Ring Compounds. It is difficult to select a set of representative cyclic and polycyclic structures; the pool of structures is so large. In order to be objective as possible in trying to include different cases and not exclude potential problem structures, we decided to examine *all* structures that have been mentioned in a brief review on unusual ring systems in organic chemistry.[42] In Table XI and Figure 6 we show computed I.D. values and molecular graphs, respectively. The selected structures show impressive variety in size, shape, and ring structure. The purpose of reporting I.D.'s for such a group is to show the range of values as well as to examine the already mentioned property of I.D. numbers in that intuitively similar

Table IX. Monocyclic Graphs with $n = 7$ Vertexes (56–84) and Selected Monocyclic Graphs with $n = 8$ Vertexes (85–106) That Show Similar I.D. Numbers

graph	I.D.	graph	I.D.
56	12.7469829	82	13.5852364
57	12.767767	83	13.6768253
58	12.7967764	84	13.890625
59	12.8698709	85	15.0180056
60	12.9231036	86	15.0182681
61	12.9323757	87	15.0915444
62	12.9326343	88	15.0918215
63	12.9515873	89	15.1592795
64	12.9547281	90	15.1596747
65	12.9617205	91	15.1235927
66	12.9662897	92	15.1235214
67	13.0437615	93	15.3270861
68	13.0950265	94	15.3276597
69	13.0952112	95	15.3344925
70	13.12884	96	15.3341519
71	13.1285051	97	15.3217194
72	13.1289923	98	15.3213754
73	13.2154514	99	15.4652381
74	13.2186104	100	15.4656827
75	13.2447313	101	15.2437603
76	13.3256371	102	15.2430295
77	13.3259335	103	15.2437182
78	13.410812	104	15.5334097
79	13.3704004	105	15.5329582
80	13.453399	106	15.5328292
81	13.4539269		

Table X. I.D. Numbers for Bicyclic Graphs Having $n = 6$ Vertexes

graph	I.D.	graph	I.D.
107	11.350627	116	11.735597
108	11.3543155	117	11.7433104
109	11.4120045	118	11.7965985
110	11.432653	119	11.8017705
111	11.4349802	120	11.882575
112	11.4468102	121	12.0020974
113	11.4853111	122	12.1089994
114	11.5072408	123	12.1742346
115	11.7051539		

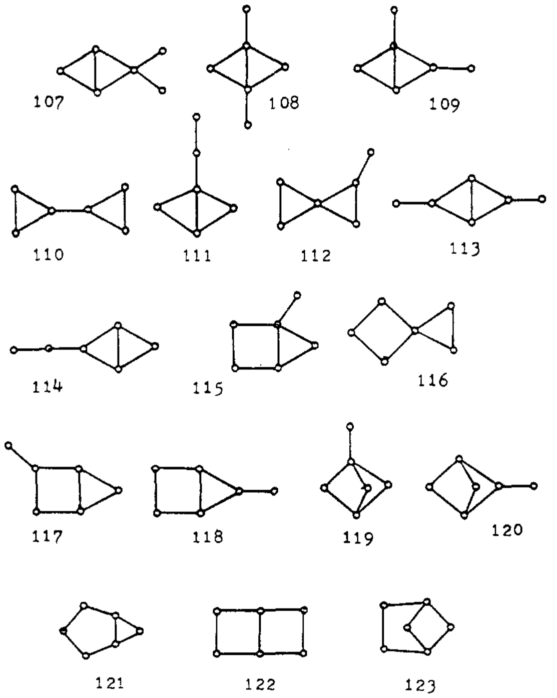


Figure 5. Bicyclic graphs having $n = 6$ vertexes.

structures give numerically similar results for I.D. values. Consider, for example, tricyclic spiro octanes with the I.D. numbers of 15.9613 and 15.9616. Similar I.D. numbers are also found among the spiro compounds having four C_3 rings, connected in a chain (I.D. = 18.0975) or joined to the central ring (I.D. = 18.0379). At first look, the two compounds appear less similar, but they correspond to a butane–isopropane kind of variation if attention is focused on the C_3 rings. Thus, even if compounds may not at first glance look alike, similar I.D. numbers may suggest some structural relationship. Generally, structurally similar compounds are expected to have similar I.D.'s, but of course, the converse needs not be true. By contracting a structure to a single number, one should not be suprised to come across diverse structures having close I.D. values. It is therefor more meaningful to make comparisons within a class of structurally related compounds, which would

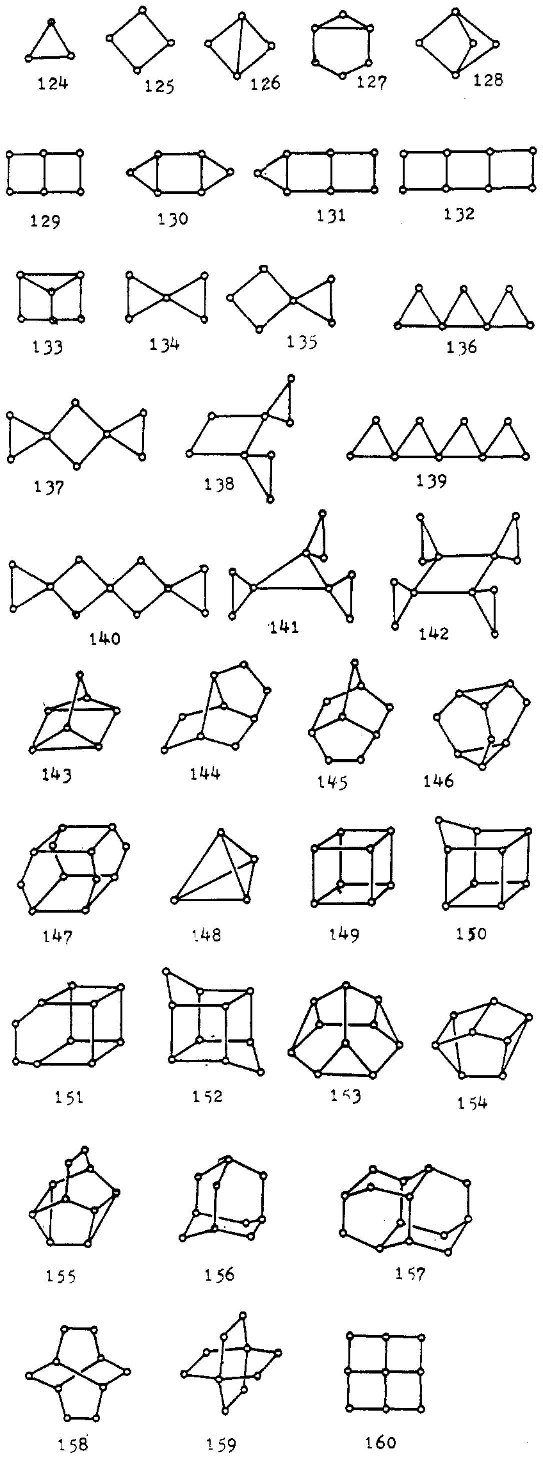


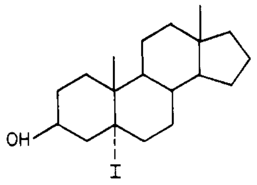
Figure 6. Skeletons of organic compounds showing different ring structure including unusual ring systems.

Table XI. Carbon Skeletons for Molecules Showing Different (and Unusual) Ring Structure[a]

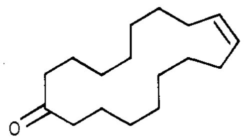
graph	I.D.	graph	I.D.
124	5.25	143	15.0076026
125	7.5	144	19.2052889
126	7.56060086	145	19.2127979
127	12.0020974	146	19.3869551
128	9.932653	147	26.8681152
129	12.1089994	148	7.77777778
130	12.1653	149	17.9917696
131	12.4543807	150	20.2633
132	16.7565365	151	22.4172175
133	12.4557854	152	22.5248849
134	9.49632035	153	22.4573998
135	11.735597	154	17.6762689
136	13.7850739	155	22.1002019
137	15.9613907	156	21.4170255
138	15.9616207	157	31.2107297
139	18.0975525	158	21.3924287
140	22.488961	159	21.1787608
141	18.0379058	160	19.4773775
142	24.3159362		

[a] Numbers showing less than nine digits end with zeros, which are not shown in their I.D. values.

minimize or exclude accidentally similar I.D. numbers. We see from Table XI that I.D. numbers are close to $2n$, twice the number of atoms in a structure. Variations around $2n$ indicate variations in complexity, cyclicity of individual structures having n atoms. Hence, I.D. numbers have an important property—they treat molecules of similar *size* similarly. Ordinary path numbers have shown considerable variation among structures of a similar size and were less suitable for comparison of structures that have different ring structure. Similarity among sterol I and macrocyclic civetone II that is limited to their shape and periphery would remain undetected if unweighted paths are used for comparison. With weighted paths, the difference in path numbers is drastically reduced, suggesting the way for modification of a comparison procedure that may reveal partial resemblance of the two structures. We require such capabilities because such compounds may possess similar properties. In fact for the two compounds, Prelog and Ružička have shown many years ago[43] suprisingly remarkable similarity in one of their properties: both showing a decidedly musklike odor.



I



II

Terpenes. In Table XII we show I.D. numbers for 30 sesquiterpenes that are ordered with increasing I.D. numbers. Diversity of structural forms (acyclic, monocyclic, bicyclic, and tricyclic) (Figure 7) is apparent, yet the I.D. numbers vary less than 10%, from 28.30 to 30.63. Besides the expected increase in I.D.'s with the number of cycles, we see that, among bicyclic structures, those having a "bridge" (e.g., trichothecane, cuparane, laurane) precede those with a spiro carbon atom (e.g., chamigrane, vetispirane, acorane). Here again, we can testify that similar structurs have similar I.D. numbers. The three bicyclic terpenes having a "bridge" have I.D. numbers 29.7377, 29.7708, and 29.8487, differing by less than 1%.

Miscellaneous Graphs. The present I.D. numbers, though called "molecular", are at this stage in fact graph I.D. numbers, because we have not yet developed extension of the scheme to account for the presence of heteroatoms and spatial architecture of molecules. Extension to structures having double bonds is not difficult; in fact, the present scheme is applicable to such situations. One of the problems to be resolved in such

Table XII. I.D. Numbers for Selected Terpenes

graph	name	I.D.	graph	name	I.D.
161	linear	28.3090241	176	acorane	29.9090846
162	elemane	28.8736353	177	pseudo-guaiane	29.9197456
163	humulane	29.2158421	178	carotane	29.9597003
164	bisabolane	29.2168136	179	eudesmane	29.9825304
165	germacrane	29.3102768	180	eremophilane	29.9833881
166	drimane	29.7310353	181	himachalane	29.9938107
167	trichothe-cane	29.7377264	182	guaiane	30.0213171
168	tutin group	29.75838	183	cadinane	30.0672085
169	cuparane	29.7708744	184	widdrane	30.1605561
170	santalane	29.7935876	185	aristolane	30.3678638
171	laurane	29.8487795	186	maaliane	30.3680133
172	caryophyl-lane	29.8641538	187	aromaden-drane	30.403248
173	chamigrane	29.8730547	188	khusane	30.5713067
174	vetispirane	29.8834234	189	cedrane	30.6231791
175	valerane	29.9004075	190	patchoulane	30.6394614

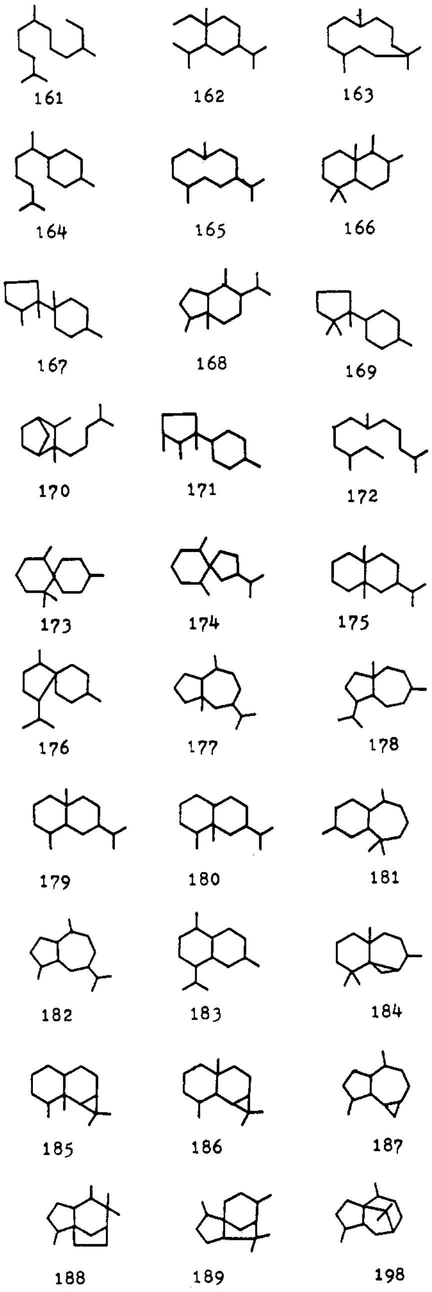


Figure 7. Carbon skeletons for selected terpenes.

Table XIII. I.D. Numbers for Potential Counterexamples to Uniqueness of the I.D. Codes[a]

graph	I.D.	graph	I.D.
191	27.665803	194	27.8691202
192	27.6792156	195	31.3665699
193	27.8726143	196	31.3689082

[a] (191–192) Graph having same total number of paths (reported by Quintas, Pace University, Department of Mathematics, New York); (193–194) another pair having the same total number of paths; (195–196) isospectral graphs of Baker (*J. Math. Phys.* **1966**, *7*, 2238) representing "drum shapes".

extensions is how to modify weights when multiple bonds appear. The most simple way is to the use multiplicity of a bond as an additional weighting factor. Preliminary results show that again we have apparant diversity of molecular I.D. numbers among such multigraphs (unsaturated compounds).[44] However, more work is required before one can with the same or similar confidence claim high selectivity for unsaturated systems. While thus incorporation of "chemical" features remains as the outstanding task, before such extensions are attempted one should critically examine the performance of the scheme for ordinary graphs. The most outstanding problem here is to determine the uniqueness of I.D. numbers. There is no simple way of showing that different graphs will *always* produce different I.D. numbers. By accident, by some systematic exploration of structural siblings, or, perhaps, by deliberate design, it is possible to arrive at structures having the same I.D. In fact, one can expect a "hunt" for such iso-I.D. structures in near future. Be as it may, unless immediate chemical application does not reveal structures having the same I.D., occurence of occasional degeneracy in I.D. need not to be very disturbing for practical chemical application. We are almost in a no-lose proposition: if no counterexample is found, the scheme may have wider applications (e.g., for isomorphism testing). If a counterexample is found, we would become aware of special structural conditions that underline such rather unusual situations, just as the occurrence of isospectral graphs has lead to recognition of isospectral points, which were later tied to equinumerocity of walks and unusual random walks.[45] With such general thought, it seems prudent to continue to explore structurally unusual graphs. Hence, we conclude this paper with Table XIII and Figure 8, which give results for selected special graphs. We have selected isospectral graphs having the same bond types and differing in positioning of inside bridges. The remaining graphs represent pairs that have the same (unweighted) path count and are regular graphs. Being *regular* means uniform weighting for all bonds, thus opening a possibility for coincidental path sums. However, as we see in both cases, we find different I.D. numbers. Graphs of Figure 8 and the results in Table XIII represent the beginning of the "hunt" for counterexamples by an ad hoc test for potentially iso-I.D structures. Each such attempt ought to be recorded, even if a counterexample is not found, particularly if close numerical values are found for I.D. numbers, as this can give us more insight into all the contributing factors.

FURTHER APPLICATIONS

Besides the already mentioned extensions of the scheme to incorporate "chemical" characteristics of a structure and continuation of the search for counterexamples to uniqueness of the I.D. numbers, there are other directions of exploration and application of the present approach. Alternative weighting schemes as well as alternative modes for contration of atomic information may be considered. There is considerable potential of atomic path sequences a_0, a_1, a_2, ... as well as molecular path sequences m_0, m_1, m_2, ... in study of structure–activity for characterization of atomic environments, for quantifying molecular similarity, and for systematic classification of compounds. Finally, path numbers can be used in cluster analysis,[26] and weighted path numbers can here have interesting applications. Use of I.D. numbers in structure–activity work is not preconditioned by either of the difficult problems involving the I.D. concept: (a) uniqness of I.D. numbers and (b) nonpolynomial character of the algorithm counting paths. Both are hardly relevant to such applications, because similar molecules frequently have similar properties; hence, having the same numerical characterization may even be desirable. In that respect as Dubois et al. have shown, topological indexes of very high selectivity do not perform as well in structure–property correltion as other indexes of lesser selectivity. On the other hand, the size of most biologically active substances is, relatively speaking, "limited" and lays well within the domain of practicality of the ALL-PATH program. However, from the position of chemical documentation *both* open problems are of profound significance. This, on its own, justifies a "hunt" for iso-I.D. graphs, but such a hunt should not halt when the first cases of such graphs are found. Rather, it should halt when we understand all structural factors determining these unusual conditions. In the same spirit, one can appreciate continued interest in isospectral graphs, interest that comes not because we want to catalog yet another case of special graphs but because we still do not fully understand all the structural conditions causing isospectrality. With respect to the size of the graph, clearly as n increases eventually a point of unpracticality arises. Dodecahedron with 11 fundamental rings took 45 min to generte a single atomic path sequence. Since this molecules has a high symmetry, that is all that is required. Probably a factor of 10 can be considered as the factor in speed of a program in BASIC and FORTRAN when run on a personal computer (or if BASIC is complied in binary before used). A similar factor is involved between different sorts of computers, but in short, we see that dodecahedron approaches the limits of practical use. Before dismissing hopes for application to even larger structures, one should, however, consider a possible upgrading of the algorithm used. Recently for example,[25] the count of paths in relatively large molecules (tricyclic derivatives of phenanthrenylcarbinol having more than 20 atoms in hydrogen-supressed graphs) was accom-

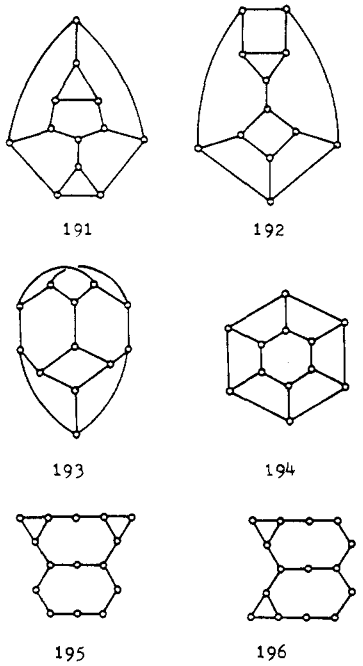


Figure 8. Initiation of a "hunt" for counterexampless to uniqueness of I.D. numbers: a pair of regular trivalent graphs having the same total number of (unweighted) paths (two examples) and a pair of isospectral graphs having additional structural properties.

plished by hand calculation. This was possible by suitably fragmenting the molecule, by enumerating the paths in the molecular fragments, and, then, by recombining the partial results. By implementation of this kind of approach in a computer program the amount of work in complex structures will be considerably reduced.

CONCLUDING REMARKS

The present scheme offers a simple topological (or, more correctly, graph theoretical) index—the total number of suitably weighted paths in a structure. As a single item, the numbers show a remarkable ability to differentiate among structüres, which justifies them to be referred to as molecular I.D. numbers, despite that the question on uniqueness remains open. The I.D. numbers have preserved some structural information and may provide guidance in the organization of large collections of data. They represent a contraction of information on total number of (weighted) paths for individual vertices, which in turn are the contraction of the count of (weighted) paths of different length for individual atoms. Alternatively, I.D. numbers can be viewed as contraction of data on molecular path numbers, i.e., the count of weighted paths of different length for a molecule as a whole. Thus, one can say that with consideration of molecular I.D. numbers we have seen only "the tip of an iceberg". Hence, even when two structures having the same "tips" are found, we have numerous other associated descriptors that are likely to suffice to differentiate such special cases. More importantly, the "reservoir" can be used for a more complete characterization of a structure, some fragments, or individual atomic environment—all in effort to contribute to structure–activity and structure–documentation analyses that would be as free as possible of arbitrary, empirical, contaminations devoid of rigorous structural content.

APPENDIX

The program ALL-PATH, the version capable of counting paths in graphs with multiple connections (see *Comput. Chem.* **1980**, *4*, 27) can be used without modification to derive I.D. numbers. The input *I*, *J*, *M* in which vertices *I* and *J* are neighbors with multiplicity *M* has to be interpreted as *I* and *J* as neighbors with *M* as the weight for the connection. A subroutine has been written that automatically evaluates the weighting factors when *I* and *J* are given as input. The program is available in BASIC, suitable for an Apple IIe personal computer.

ADDED IN PROOF

We have derived I.D. numbers for all 159 undecanes and have not found any duplicate case.

REFERENCES AND NOTES

(1) Ebe, T.; Zamora, A. "Wiswesser Line Notation at Chemical Abstracts Service". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 33. Dyson G. M. "A New Notation and Enumeration System for Organic Compounds"; Longmans: London 1949.

(2) Morgan, H. L. *J. Chem. Doc.* **1968**, *5*, 107.

(3) Goodson, A. L.; Lozac'h, N.; Powell, H. W. *Angew. Chem., Int. Ed. Engl.* **1979**, *18*, 887.

(4) Chemical Abstracts Service, Columbus, OH.

(5) Randić, M. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171.

(6) Randić, M.; Brissey, G. M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 52.

(7) Read, R. C. "The Coding of Various Kinds of Unlabelled Trees". In "Graph Theory and Computing"; Read, R. C., Ed.; Academic Press: New York, 1972; p 153.

(8) Read, R. C. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 135.

(9) Randić, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. *Comput. Chem.* **1979**, *3*, 5; **1980**, *4*, 27.

(10) Garey, M. R.; Johnson, D. S. "Computers and Intractability-A Guide to the Theory of NP-Completeness"; W. H. Freeman: San Francisco, 1979.

(11) Randić, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 36.

(12) Editorial correspondence, Feb 22, 1984.

(13) Trinajstić, N. "Chemical Graph Theory"; CRC Press: Boca Raton, FL, 1983; Vol. II, Chapter 4.

(14) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17; *J. Am. Chem. Soc.* **1947**, *69*, 2636. *J. Chem. Phys.* **1947**, *15*, 766; *J. Phys. Chem.* **1948**, *52*, 1082. Platt, J. R.; *J. Chem. Phys.* **1947**, *15*, 419; *J. Phys. Chem.* **1952**, *56*, 328.

(15) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332; *Theor. Chim. Acta* **1972**, *25*, 215; *Fibonacci Q.* **1973**, *11*, 255. Hosoya, H.; Kawsaki, K.; Mizutani, K. *Bull. Chem. Soc. Jpn.* **1972**, *45*, 3415. Mizutani, K.; Kawasaki, K.; Hosoya, H. *Natl. Sci. Rep. Ochanomizu Univ.* **1971**, *22*, 39. Hosoya, H. *J. Chem. Doc.* **1972**, *12*, 181.

(16) Randić, M. *J. Am. Chem. Soc.* **1972**, *97*, 6609.

(17) Kier, L. B.; Hall, L. H. "Molecular Connectivity in Chemistry and Drug Research"; Academic Press: New York, 1976.

(18) Balaban, A. T.; Motoc, I. *Math. Chem.* **1979**, *5*, 197. Bonchev, D.; Balaban, A. T.; Mekenyan, O. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 106. Balaban, A. T. *Theor. Chim. Acta* **1979**, *53*, 355.

(19) Bonchev, D.; Mekenyan, O.; Trinajstić, N. *J. Comput. Chem.* **1981**, *2*, 127.

(20) Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.

(21) Razinger, M.; Chrētien, J. R.; Dubois, J. E. "Structural Selectivity of Topological Indices in Alkane Series". Submitted for publication in *J. Chem. Inf. Comput. Sci.*

(22) Randić, M.; Wilkins, C. L. *Chem. Phys. Lett.* **1979**, *63*, 332.

(23) Randić, M.; Wilkins, C. L. *J. Phys. Chem.* **1979**, *83*, 1525.

(24) Randić, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31.

(25) Randić, M.; Wilkins, C. L. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1979**, *6*, 55. Wilkins, C. L.; Randić, M. *Theor. Chim. Acta* **1980**, *58*, 45. Wilkins, C. L.; Randić, M.; Schuster, S. M.; Markin, R. S.; Steiner, S.; Dorgan, L. *Anal. Chim. Acta* **1981**, *133*, 637. Randić, M.; Kraus, G.; Jerman-Blažič Džonova, B. "Proceedings of the Symposium on Chemical Applications of Topology and Graph Theory". *Phys. Theor. Chem.* **1983**, *28*, 192–205. Jerman-Blažič, B.; Randić, M. "Proceedings of the International AMSE Conference an Modelling and Simulation", Nice, France; AMSE Press: Tassin, France, 1983; Vol. 5, pp 161–174.

(26) Randić, M.; Jerman-Blažič Džonova, B., submitted for publication in *J. Am. Chem. Soc.*

(27) Randić, M. *J. Comput. Chem.* **1980**, *1*, 386.

(28) Seybold, P. G. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1983**, *10*, 95; **1983**, *10*, 103.

(29) Randić, M. *MATCH* **1979**, *7*, 5.

(30) Observe that when P_0 is included as part of the atomic and molecular sequences they contribute to m_0 simply by addition, while other path numbers give twice the number of molecular paths; hence, total of atomic paths minus *n* gives I.D. values.

(31) See Appendix, copy of the program is available upon request (for noncommercial use).

(32) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. *J. Pharm. Sci.* **1976**, *65*, 1226.

(33) See footnote 9. One should qualify "complexity" of a graph; Here we speak of path complexity; other graph invariants (walks, cycles, etc.) will result in a different measure for relative complexity of graphs.

(34) Trinajstić, N. "Chemical Graph Theory"; CRC Press: Boca Raton, FL, 1983; Vol. I, Chapter 7.

(35) Collatz, L.; Sinogowitz, U. *Abh. Math. Semin. Univ. Hamburg* **1967**, *21*, 63.

(36) Živković, T.; Trinajstić, N.; Randić, M. *Mol. Phys.* **1975**, *30*, 517 (preliminary report by T. Živković at Quantum Chemistry School, Leningrad, 1973). Herndon, W. C. *Tetrahedron Lett.* **1974**, 671.

(37) Slater, P. J. *J. Graph Theory* **1982**, *6*, 89.

(38) Randić, M. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 101.

(39) Shelly, C. A.; Trulson, M. Arizona State University, private communication, 1978.

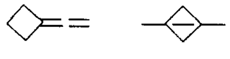
(40) Fratev, F.; Polansky, O. E.; Mehlhorn, A.; Monev, V. *J. Mol. Struct.* **1979**, *56*, 245. Carbo, R.; Leyda, L.; Arnau, M. *Int. J. Quantum Chem.* **1980**, *17*, 1185. Adamson, G. W.; Bush, J. A. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55. Knop, J. V.; Trinajstić, N. preprint, 1983. Of interest here are also reports on similarityof binary trees and dendograms (e.g., *J. Theor. Biol.* **1983**, *100*, 427; **1978**, *73*, 789).

(41) Patterson, A. M.; Capell, L. T.; Walker, D. F. "The Ring Index", 2nd ed.; American Chemical Society: Washington, DC, 1960.

(42) Hanack, M.; Subramanian, L. R.; Eymann, W. *Naturwissenschaften* **1977**, *64*, 397.

(43) Prelog, V.; Ružička, L. *Helv. Chim. Acta* **1944**, *27*, 61, 66.

(44) As pointed out by a referee, "extensions with symmetry, multiple bonds, etc. may produce real problems." If we use multiplicity of a bond (such as double bond) as another multiplicative bond factor, one can generate a number of I.D. for unsaturated carbon skeletons. Among 100 so produced new results, we find a full coincidental path count for one unsaturated structure and another structure that is saturated:



(45) Randić, M.; Woodworth, W. L.; Graovac, A. *Int. J. Quantum Chem.* **1983**, *24*, 435.