match no. 7 pp. 5 - 64 1979

Plenary Lecture

CHARACTERIZATION OF ATOMS, MOLECULES AND CLASSES OF MOLECULES BASED ON PATHS ENUMERATIONS+

MILAN RANDIĆ

Ames Laboratory, U. S. Department of Energy,
IOWA State University, Ames, Iowa 50011

Abstract

Paths of different lengths are taken as a basis for characterization of atomic environment in a molecule, for characterization of molecules, and for characterization of classes of molecules. Enumeration of paths in a graph is briefly outlined and some properties of derived sequences of path numbers are given. It is conjectured that the list of path numbers is unique to the graph. A support for the conjecture is indicated by a reconstruction, though only acyclic cases have been considered. Use of truncated atom codes is briefly indicated. Construction of molecular codes by summation of atomic path sequences , term by term, is outlined and the properties of

the molecular codes discussed. Examples are given of molecular codes in some polycyclic structures. Generally similar structures will have similar molecular codes. This property is used for characterization of classes of structurally related molecules.

+Dedicated to Professor Božo Težak

Introduction

The study of chemistry is primarily concerned with problems of forming and breaking bonds and synthetizing, isolating, and identifying unknown substances and determing their molecular structures. The number of registered compounds is now over four million and is still increasing. Various molecular forms (constitutions , configurations, isomers, tautomers, enantiomers, etc.) clearly demonstrate the enormous number of the combinatorial possibilities in nature, even when one restricts attention to only a few kinds of atoms. Yet, it is surprising how combinatorial chemistry, initiated over one hundred years ago, [1] and recognized before the turn of the century as a new and distinctive branch of chemistry,[2] still has not firmly established itself. During past years there have been numerous contributions to the field of chemical combinatorics and chemical topology, although these have not been necessarily recognized as having a graph-theoretical origin. As an example, we may mention the Coulson and Rushbrooke theorem on the pairing of orbital energies in simple MO calculations[3] for molecules whose graphs are classified as bipartite. The current revival of interest in chemical application of graph theory[4] has already lead to numerous new results. Particularly application of graph theory helped viewing some old

problems from a different standpoint and contributed to clarification of some ambiguities concerning such concepts as aromaticity, molecular resonance energy, molecular additivities, etc. However, chemical graph theory as a subject is still in the early phase of development. Complexity of some problems that are currently important to chemistry contributes to an interest in graph theory and promises further evolution and expansion of chemical graph theory as a prominent branch of theoretical chemistry. Already now graph theory has been found valuable in the field of chemical documentation, computer use in chemistry (artificial intelligence, pattern recognition, computer assisted synthesis), chemical kinetics, and statistical mechanics. In such applications primarily graph theory provides help with book-keeping of numerous combinatorial possibilities. Similarly in studies of chemical transformations and even in some ab initio calculations graphs are used to visualize complex interrelations. However, graphs are not merely a diagramatic representation of relations, they can be altered, augmented or fragmented in a meaningful way and corresponding to certain mathematical or chemical operations. Such applications although not yet numerous justify claims to the fundamental character of graph theory in chemistry, which is parallel to the role of group theory, statistcs, and quantum mechanics, each of which characterizes distinctive property of nature. Group theory is concerned with symmetry, statistics with distributions, quantum mechanics with interactions, and graph theory with relations (binary relations - to be precise). Since relations can be depicted as graphs we may equally consider connectivity as the basic quality of graphs. Relationship to chemistry is now obvious: molecules can be viewed as graphs, chemical bonds corresponding to binary relations. Consider now the problem of enumeration of all isomers of given molecule. Clearly the problem does not belong to quantum mechanics, nor statistics, nor group-theory. Combinatorial aspect of the problem arises in having multiple possibilities of linking bonds, the topological aspect is reflected in the fact that the result does not depend on details of graphical representation. Non-triviality of the problem, which is deceptively simple, is well known to all who have

been confronted with the isomorphism problem.[5] The above is historically the oldest graph-theoretical problem in chemistry and despite an important advance made by Polya who derived his counting theorem[6] the enumeration of structures of certain type as a topic still remains of interest and poses problems.

There is no doubt that chemical graph theory is on the rise and that it will earn its deserving place in chemistry, though in this respect, chemistry is today behind physics and biology where its role is fully recognized. Historically, chemistry was the first among natural sciences to find use for a graph-theoretical approach. Equally, in return, chemistry contributed to developments of some aspects of mathematical graph theory. A visible trace of past close relationship between chemistry and graph theory is adoption of some chemical terms in the mathematical graph theory literature. However, since chemical graph theory has not been so firmly established, and is still ignored in some chemical circles, one may doubt its role in chemistry, or at best agree its importance in some very special problems of peripheral interest. Some explanation for the lack of past impact of graph theory on the main stream of developments in chemistry would be desirable. Most likely, the reason for delayed recognition of graph theory among a majority of chemists is the past concern in chemistry with problems of limited combinatorial content. Just as group theory, which had not been well known in the early 1930's, obtained enormous impetus by the development of molecular spectroscopy and interest in highly symmetrical systems, so will graph theory surge to unsuspected heights when current interest in complex structures and complex problems further expands and makes the use of graph theory rather obvious. Just as one may proceed in problems of limited symmetry without the use of the otherwise powerful tool of group theory and resolve his questions intuitively, so in problems of limited combinatorial contents one may also proceed intuitively, not realising that he operates with some graph-theoretical concepts. It is with problems in which the number of combinatorial possibilities "explodes" that one can either (1) continue

investigating the problem and consequences using available graph-theoretical methods or devising new approaches, or (2) abandon the problem as impossible. Unfortunately, in the past, the limiting computational facilities have frequently forced one to look for a simpler problem with subsequent detrimental consequences for development of chemical graph theory.

Graph theory is of interest in chemistry because it is primarily concerned with structures and manipulations with structures. One can compare structures, combining or dissecting them, or looking for particular fragments in them. Such operations require an ability to recognize a structure, distinguish among non-isomorphic structures, order or classify structures. Normally one examines a structure and selects suitable graph invariants for characterization of the structure. Frequently this requires enumerations of some components or alternatively use of standard mathematical methods applies (e.g., evaluation of the determinant of the adjacency matrix, graph spectra etc.). Analysis of large molecules, although straightforward can frequently be tedious and the use of computers is preferred. This requires a method for encoding for a structure suitable for computer manipulations and retrieval. Depending on application, one may be interested in reconstruction, fragment search, ordering (which may be only partial), or similarity among structures and their relevant properties. It is clear that these various aspects appear in many chemical problems. Table 1 summarizes some of these operations on graphs of interest to chemistry. The brief descriptions only serve to clarify the operations qualitatively, not to define them.

This contribution deals with the concept of atomic and molecular codes based on enumeration of paths as the fundamental graph invariants. Such codes are mathematically given by a sequence of integer numbers, and, as will be seen, provide useful identification of atomic molecular environments. Hence, one can anticipate use of such atom codes in chemistry. In fact, the notion of atomic neighbours and paths is not so novel in chemistry, and has been frequently implied in discussions of molecular

properties. Nevertheless, it appears that the concept of atomic paths has not had complete formal mathematical development. Our emphasis is on this formalism, and as will be seen, such an attitude has lead to various useful developments. Atomic codes have been found useful in discerning regularities in available atomic and molecular data,[7] lead to sequencing of structures that show trends for the relative magnitudes among apparently unrelated data,[8] and produced a scheme for a quantitative measure of the degree of similarity among molecules[9] with subsequent applications in structure-activity correlations and drug design type work.[10]

Table 1

A Selection of Operations on Graphs of Interest in Chemistry

AUTOMORPHISM	Graph symmetry. The problem of finding labels for a graph that will not alter a form of the adjacency matrix
CHARACTERIZATION	Assigning a set of parameters (or single index) that reflects some structural features, to a structure
CLASSIFICATION	Grouping structures according to a set of preselected parameters into smaller lots
COMPARISON	Watching for differences in selected parameters among two or more structures
COMPOSITION	The process of combining constituents (such as fragments) into a system
CONSTRUCTION	Building of a structure or a graph according to prescribed rules or of prescribed form
DISCRIMINATION	Selection of a singular parameter or a limited number of parameters for structure differentiation
DISSECTION	Dismembering of a structure in smaller fragments, preferably parts which can not be fragmented further

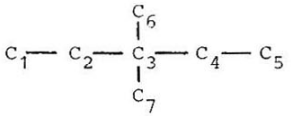
ENCODING	Assigning to a structure or graph a set of parameters, usually derived by counting the selected graph invariants
ENUMERATION	Counting structures of fragments of a particular form, size, or shape
FRAGMENTATION	Breaking a structure into smaller parts
FRAGMENT SEARCH	Subgraph isomorphism. The problem of identifying the occurrence of a smaller structure in a larger system
ISOMORPHISM	Verifying whether two graphs have identical connectivity, and therefore represent the same system
ORDERING	Sequencing structures according to a selected set of parameters
RECOGNITION	Testing a structure or graph against known structures or a list of known properties
RECONSTRUCTION	Recovery of a structure from a collection of incomplete data or collection of fragmentary data
RETRIEVAL	The process of finding items of particular interest among large quantity of data
SIMILARITY	Appraise the degree of agreement or disagreement among given structures with respect to a preselected feature

Paths and Their Enumeration

A path is defined in graph theory as a sequence of vertices (or edges) in a graph which are connected, with no vertex appearing more than once in the list. The length of the path is the number of edges involved. The enumeration of paths in large graphs is difficult because of their rapid proliferation as the size of the graph increases and especially as more rings are introduced. Although many molecules of chemical interest will fall into a class for which enumeration of paths is possible and practical, one should be aware of the exponential growth of the number of paths so that one can recognize

intrinsic limitations of a method using paths with no restrictions. As an illustration of the almost hopeless situation, consider a 10 x 10 grid graph and the task of finding the number of ways one can go from one corner of the grid to the opposite corner.[11] Problems of this nature appear in the physics and chemistry of very large molecules.[12] Currently such problems can only be tackled using probability calculations. (The above example is estimated to have 10^{24} possible paths.) As will be seen later, molecules of medium size and medium complexity(e.g., having 20-30 atoms and 6-7 rings) have only 10^3 - 10^4 possible paths and are fully within the reach of detailed examination.

To illustrate the concept of paths, enumeration, and derived codes, consider a relatively simple molecule of 3,3-dimethylpentane:



In Table 2 we list all paths starting from atom 1. In this case there is only one path of length one and one path of length two, there are three paths of length three and finally one path of length four. The number of paths of length one, two three, four, etc. can be listed as a sequence p_1, p_2, p_3, p_4, ... which represents a code for the considered atom. For atom 1 of 3,3-dimethylpentane we obtain the sequence 1, 1, 3, 1. In general other atoms will have different sequences signifying their participation in paths of different length.

Table 2

Path length	1	2	3	4
Paths	1-2	1-2-3	1-2-3-4 1-2-3-6 1-2-3-7	1-2-3-4-5
Number of paths	1	1	3	1

Between pairs of atoms in acyclic structures there is a unique path so that the number of paths of a given length corresponds to the number of neighbours of a given distance. As a further illustration in Table 3 we list atomic codes for all heptane isomers. The numbering of atoms is shown below:

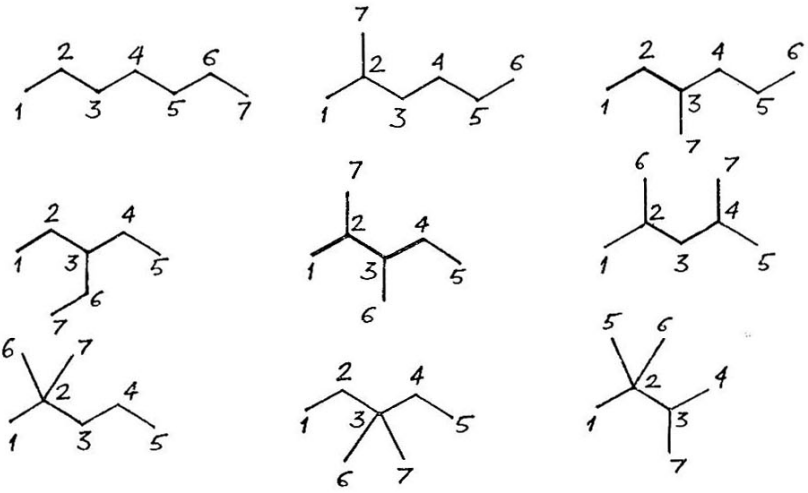


Table 3

Atomic Codes for Heptane Isomers

n-heptane			2-methylhexane			3-methylhexane		
1	*	1,1,1,1,1,1	1	*	1,2,1,1,1	1	*	1,1,2,1,1
2	*	2,1,1,1,1	2	*	3,1,1,1	2		2,2,1,1
3		2,2,1,1	3		2,3,1	3		3,2,1
4		2,2,2	4		2,2,2	4		2,3,1
5		2,2,1,1	5	*	2,1,1,2	5	*	2,1,2,1
6	*	2,1,1,1,1	6	*	1,1,1,1,2	6	*	1,1,1,2,1
7	*	1,1,1,1,1,1	7	*	1,2,1,1,1	7		1,2,2,1

3-ethylpentane			2,3-dimethylpentane			2,4-dimethylpentane		
1		1,1,2,2	1		1,2,2,1	1	*	1,2,1,2
2		2,2,2	2		3,2,1	2	*	3,1,2
3		3,3	3		3,3	3		2,4
4		2,2,2	4		2,2,2	4	*	3,1,2
5		1,1,2,2	5		1,1,2,2	5	*	1,2,1,2
6		2,2,2	6		1,2,3	6	*	1,2,1,2
7		1,1,2,2	7		1,2,2,1	7	*	1,2,1,2

2,2-dimethylpentane			3,3-dimethylpentane			2,2,3-trimethylbutane		
1	*	1,3,1,1	1	*	1,1,3,1	1		1,3,2
2	*	4,1,1	2		2,3,1	2		4,2
3		2,4	3		4,2	3		3,3
4	*	2,1,3	4		2,3,1	4		1,2,3
5	*	1,1,1,3	5	*	1,1,3,1	5		1,3,2
6	*	1,3,1,1	6		1,3,2	6		1,3,2
7	*	1,3,1,1	7		1,3,2	7		1,2,3

* indicates unique codes

Previous work

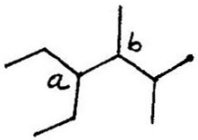
Enumeration of neighbours a certain distance away is neither such a novel nor unfamiliar concept and must have been considered in various problems in one form or another. The notion of a "shell of neighbours" was considered attractive in the field of chemical documentation.[13] It is the basis for a systematic enumeration of all alkanes and leads to a particular hierarchy of structures of interest in the study of the various additivity schemes.[14] A rather comprehensive system for encoding atomic environments by Dubois[15] is essentially

based on neighbour and path concepts. The approach was used in structure-property correlations as well as providing a basis for a logical scheme of chemical nomenclature.[16] Simple enumeration of neighbours already in such regular structures as benzenoid conjugated hydrocarbons will characterize atomic environments and provide basis for a classification of chemical shifts.[7]

The number of neighbours a given distance away is the prime feature of the wellknown additivity scheme of Grant and Paul[17] which predicts magnitudes for carbon chemical shifts in paraffins. Truncated atomic codes in this case provide the basis for classification of carbon atoms and subsequent illumination of the trends for the magnitudes of the chemical shifts.[18] Atomic codes have also been found useful in searches for fragments in acyclic and polycyclic structures.[19] Finally, atomic codes lead to construction of molecular codes (as will be outlined here) and the latter have found important applications in recognizing regularities in available molecular data.

Properties of Atomic Path Codes

An examination of Table 3 reveals some properties of atomic path codes. The same code may appear in different molecules, yet may describe somewhat different environment. For example, the code 2,2,2 describes different environments in n-heptane, 2-methylhexane, 3-ethylpentane, and 2,3-dimethylpentane. In each case the distribution of the next-to-nearest neighbours is different, although the number of neighbours a certain distance away is always the same. Hence, it should not be surprising to find that even within the same molecule two atoms which are non-equivalent may have the same code. For example, in 2,3-dimethyl-4-ethylhexane atoms a and b both have the same code 3,4,2 but are clearly non-equivalent.



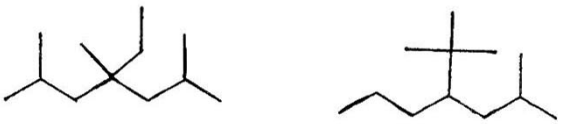
It is important to consider how much of the information on a structure is preserved in the list of atomic codes. If we can reconstruct a graph from the list if atomic codes than atomic codes have preserved all the original connectivity information. In some cases already a single atomic code does preserve all the information. Clearly such is the case with the code 1,1,1,1,1,1 of n-heptane, since in the course of reconstruction each time only one neighbour has to be added to the already existing fragment. Such singular codes which suffice for a reconstruction of a grpah necessarily have to be unique to the structure, other codes may appear in different structures. In Table 3 we have indicated these unique codes with asteriks (*). They can be recognized quickly by observing that they cannot contain entries larger than one in succession. Only 3-ethylpentane, 2,3-dimethylpentane and 2,3,3-trimethylbutane of the heptane isomers have no such unique code. Their reconstruction will require use of several codes. On the other hand, the existance of the unique codes also indicates that the list of atomic paths is, in some cases, redundant.

Another question which needs attention before considering reconstruction is the legitimacy of the atomic code list. A list is legitimate if it has been obtained from a graph, and therefore corresponds to the graph. Some necessary conditions on atomic codes to ensure legitimacy are obvious. The codes (for acyclic graphs) must represent a partition of the same number. Since acyclic graphs are bipartite, one can further deduce that the sum of alternate entries in a code is the same for atoms of the same class. In addition, the sum of the products of the first entry in a code times nth entry for all

atoms equals the sum of the nth and nth + 1 entries.[20] Finding whether a given partition is graphical (i.e., corresponds to a graph when terms represent valencies of vertices) has been solved for graphs and directed graphs.[21] The present problem is more comprehensive as it also requires conditions for next nearest neighbours and neighbours beyond.

Atom Codes and Graph Isomorphism

Comparing graphs and verifying if two graphs are the same can be conveniently performed by examining the corresponding lists of atomic path codes. Although we have no proof, it seems likely that two different graphs should have different lists of atomic paths. We have examined all graphs with fewer than 12 vertices and have found no case that contradicts this assertion. In the collection we considered, we have already several pairs of isospectral graphs[22] among which the characteristic polynomial cannot discriminate. In addition, we have examined numerous isospectral graphs of graphs with 12 vertices and more and have not found one identical list of atomic paths.[23] Another, even more restrictive group of graphs, not only have the same spectra but also have the same other matrix functions (e.g., permanent). Such a pair is:[24]



Again we find different lists of atomic paths. All this seems to justify the following

CONJECTURE: Two graphs are isomorphic if and only if they have identical lists of atomic paths

Recently Shelly and Trulson[25] verified that the hypothesis is correct for all acyclic alkanes with up to fourteen carbon atoms.

So far we have not considered cyclic and polycyclic graphs, but, as will be seen later, the above statement applies to them also. The primary difficulty involving cyclic graphs is finding all the paths, since their number is increasing exponentially with the size of the graph and the number of cycles. To refute the conjecture, two graphs would have to be non-isomorphic and have identical lists of atomic codes. Such graphs should not permit a reconstruction (i.e., in the process of reconstruction one should arrive at an ambigous step with alternative possibilities, each possibility producing a different graph). Hence, each successful reconstruction provides additional strength to the conjecture.

Reconstruction of Acyclic Graphs

We will illustrate the reconstruction procedure on one of the nonane isomers. In Table 4 we give the list of atomic codes as derived by inspection of the molecular skeleton. Atoms have been arbitrarily numbered from 1 to 9:

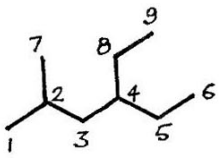


Table 4

Atomic Codes for 2-methyl-4-ethylhexane

1	1,2,1,2,2	a -	1,1,2,2,2
2	3,1,2,2	b -	1,1,2,2,2
3	2,4,2	c -	1,2,1,2,2
4	3,3,2	d -	1,2,1,2,2
5	2,2,2,2	e -	2,2,2,2

6	1,1,2,2,2	f - 2,2,2,2
7	1,2,1,2,2	g - 2,4,2
8	2,2,2,2	h - 3,1,2,2
9	1,1,2,2,2	i - 3,3,2

In addition we make an ordered list of atomic codes, starting with the codes with the smallest initial entries. When there are several such codes, one compares the next entry in the sequence, and if these are again identical, the next path numbers are tested, and so on. Labels (a) to (i) are introduced only for the purpose of reference. One should realize that the list of atomic codes (whether ordered lexicographically as shown above or in some other arbitrary way) represent the registering of a structure without the use of atomic labels. The collection of atomic codes

1,1,2,2,2	1,1,2,2,2	1,2,1,2,2
1,2,1,2,2	2,2,2,2	2,2,2,2
2,4,2	3,1,2,2	3,3,2

represents molecule 2-methyl-4-ethylhexane. If the conjecture that list of atomic codes is unique holds we would have simple label-free encoding of molecular skeletal forms. This particular property of atomic codes will be of considerable interest in applications involving the comparisons of different files which are based on different internal labelling system or the comparisons of different nomenclature systems.

The reconstruction is accomplished when we establish which codes in a list of atomic codes belong to adjacent atoms. By knowing adjacency relationships we know the graph. Because

terminal atoms have only one neighbour, and are easy to recognize we can start the reconstruction with them. The first entry for a terminal atom is necessarily 1, hence codes (a) to (d) belong to terminal atoms. Atoms adjacent to terminal atoms will have the same neighbours as the terminal atoms but each path will be one unit of bond distance smaller. In addition, they will have one more nearest neighbour, the terminal atom. From the codes of the terminal atoms we can find the codes of adjacent atoms by shifting the code to the left by one place and erasing the first entry, then adding to the new first entry +1. For the two non-equivalent terminal codes of the graph of Table 4 we have:

Terminal codes:	1,1,2,2,2	1,2,1,2,2
Shift to the left:	1,1,2,2,2	1,2,1,2,2
Erasure:	1,2,2,2	2,1,2,2
Add +1	+1	+1
New code	2,2,2,2	3,1,2,2

The new codes can now be identified as (e) or (f) and (h) respectively. In this way we arrive at the following connectivities:

(a)-(e) and possibly (b)-(e) or (b)-(f)
(c)-(h) and (d)-(h)

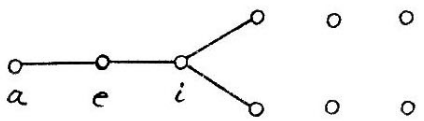
Since there are two vertices with the code 2,2,2,2 we cannot, without further verification, be certain if the both terminal vertices (a) and (b) are connected to the same or two different atoms. No such an ambiguity arises with the terminal vertices (c) and (d), because both have to connect to the same and only atom with the required code. We continue the process with atom (e) in a similar fashion:

2,2,2,2	Code of atom (e)
2,2,2	Shift + erasure: this accounts for all neighbours of (e) which have not yet been assigned label

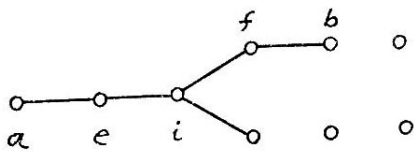
+ 1,1	Adding neighbours (paths) which have been assigned (i.e., for which we know connections)
3,3,2	New code, to be identified as (i) in Table 4

It is helpful to accompany such search for adjacency by depicting vertices and inserting edges as the process of reconstruction advances. For the code (a) we first place the required number of vertices at the intervals corresponding to an increasing distance:

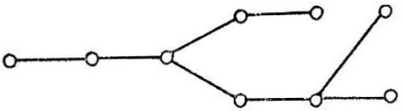
Because we consider connected graphs we can immediately insert several edges:



This already show that atom (e) can have only one terminal neighbour, hence (b) is connected to (f). Since (b) and (f) have respectively same codes as (a) and (e) it follows that vertex (i) is also connected to (f):



Because (b) is terminal one can complete the reconstruction by making the missing links:



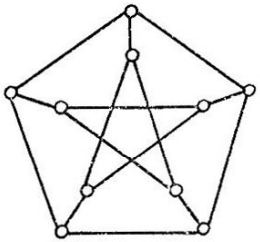
From the above identification of unassigned vertices is straightforward. However we will pretend that we dont know the result and will continue to describe the searching algorithm which is suited for full computer processing. The basic element of the search is inclusion of found information which then allow one to determine the part in the construction of the new code corresponding to vertices with assigned labels. With the appearance of the branching vertex (i), one must prior to the shifting operation, subtract from the code known (assigned) neighbours:

3,3,2	Code for the branching vertex (i)
-2,2	Substract known (assigned) neighbours
1,1,2	The result: the neighbours (paths) of (i) which have not yet been identified
1,2	Shift for one place: unidentified neighbours of the vertex investigated (adjacent to (i))
+1,2,2	Add known neighbours (already assigned)
2,4,2	New code, identified as (g) from Table 4

In the next step one likewise finds that (g) and (h) is connected, thus completing the assignment of codes and connectivities.

Path in Polycyclic Structures

The concept of the path is the same whether we consider acyclic or polycyclic structure. However, no longer a parallelism between the count of neighbours and the count of paths is valid, since one can arrive at a same atom generally by using different paths. The concept of path appears more fundamental and no longer we will refer to enumeration of neighbours, except for special applications. The enumeration of paths in polycyclic structures is rather involved. We will illustrate the complexity of the search for paths on the Petersen's graph[26]



In Table 5 we have summarized the number of rings of different size and the number of paths of different length for the Petersen's graph. In all there are 57 rings of different size and 2750 paths of various length. The perception of rings has been considered in the literature and is not a trivial task. Finding paths may be a simpler routine, but the proliferation of paths makes the task equally difficult.

Table 5

Rings of different size and paths of different length for the Petersen's graph

Ring size	5		6		7		8		9
Ring number	12		10		-		15		20
Path length	1	2	3	4	5	6	7	8	9
Path number	15	30	60	120	180	240	300	300	120

Monocyclic systems:

If there is only one ring present in a structure, paths can be enumerated by inspecting the molecular graph. For a cycle C_n (single ring without pending bonds) the result is almost trivial; the number of paths of any length is two, since there are two directions to circle a ring. The maximal path length is, of course, (n-1), where n is the number of atoms in the ring. With pending bonds the situation becomes somewhat composite but still relatively simple. In Table 6 we listed atomic codes for a selection of monocyclic eight-atom systems to illustrate the kind of variations found. In addition to the path sequences, we also include for each atom the corresponding total number of paths obtained by summing all entries in the sequence. This number has a diagnostic value since its value can be derived in advance and is 2(n-1) - k, where k is the number of pending atoms at the site of the substitution on the ring. All exocyclic atoms and atoms of the ring at the site of the substitution have the same path sum indicated above. This regularity helps to check the accuracy of the derived codes.

Polycyclic systems

A systematic enumeration of paths in polycyclic systems requires well-organized bookkeeping of the multiple possibilities. The task is not profound and can be accomplished with the help of path graphs, the acyclic graphs which represent the searching tree in the enumeration of paths. Let's illustrate the process of a systematic enumeration of paths on the skeleton of norbornane:

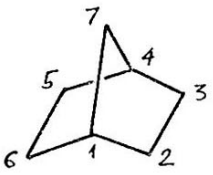
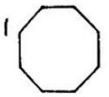
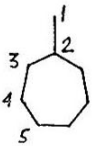


Table 6

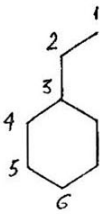
Atomic Codes for a Selection of Monocyclic Eight-Atom Molecular Skeletons



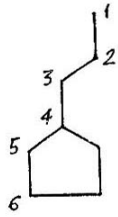
1	2,2,2,2,2,2,2,2	14



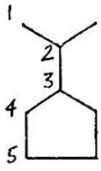
1	1,2,2,2,2,2,2,2	13
2	3,2,2,2,2,2,2	13
3	2,3,2,2,2,2,2,1	14
4	2,2,3,2,2,2,3	14
5	2,2,2,3,3,2	14



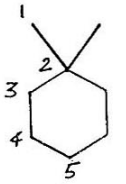
1	1,1,2,2,2,2,2,2	12
2	2,2,2,2,2,2,2	12
3	3,3,2,2,2	12
4	2,3,3,2,2,1,1	14
5	2,2,3,3,3,1	14
6	2,2,2,4,4	14



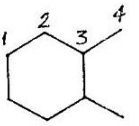
1	1,1,1,2,2,2,2,2	11
2	2,1,2,2,2,2,2	11
3	2,3,2,2,2,2	11
4	3,3,3,2	11
5	2,3,3,3,1,1,1	14
6	2,2,3,4,2,1	14



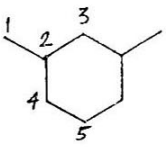
1	1,2,2,2,2,2,2	11
2	3,2,2,2,2,2	11
3	3,4,2,2,2	11
4	2,3,4,2,1,2	14
5	2,2,3,5,2	14



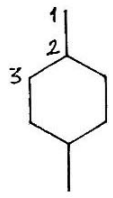
1	1,3,2,2,2,2,2	12
2	4,2,2,2,2	12
3	2,4,2,2,2,2,2	14
4	2,2,4,2,4	14
5	2,2,2,6,2	14



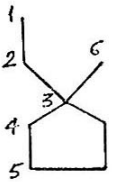
1	2,2,3,4,3	14
2	2,3,3,2,3,1	14
3	3,3,2,2,2,1	13
4	1,2,3,2,2,2,1	13



1	1,2,2,3,2,3	13
2	3,2,3,2,3	13
3	2,4,2,2,2,2	14
4	2,3,2,4,2,1	14
5	2,2,4,2,4	14

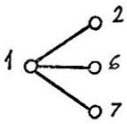


1	1,2,2,2,4,2	13
2	3,2,2,4,2	13
3	2,3,3,2,3,1	14

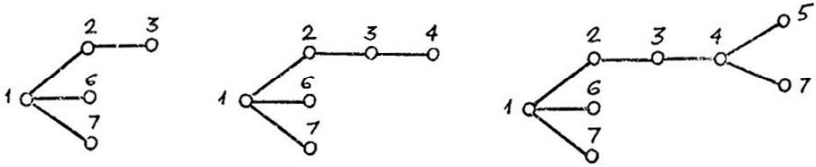


1	1,1,3,2,2,2	11
2	2,3,2,2,2	11
3	4,3,2,2	11
4	2,4,3,2,2,1	14
5	2,2,4,5,1	14
6	1,3,3,2,2	11

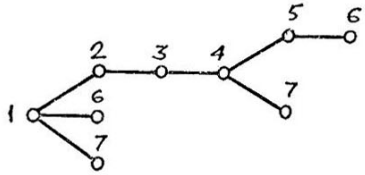
Consider atom 1. It has three nearest neighbours, 2,6, and 7, which can be diagramatically represented by



The three paths are 1-2, 1-6, and 1-7. We return to the first path and continue to search for its extensions. Atom 2 has only a single nearest neighbour in the direction of the exploration, atom 3, which increases our path to 1-2-3. From atom 3 we come to atom 4. Here we find a branching. Each step of the search can be followed by developping the searching graph as illustrated below:



Again, continuing with the first choice, smaller label, we arrive at atom 6, the end of this particular search trail:



Therefore, if we start from atom 1 and always select the atom with the smaller label, when a choice is present, we find the following paths:

1-2, 1-2-3, 1-2-3-4, 1-2-3-4-5, and 1-2-3-4-5-6.

Now we backtrack to the previous branching site, atom 4, in order to explore the remaining possibilities associated with

this branching point. Here, however, the trail ends with atom 1, giving only one additional path originating at atom 1,

1-2-3-4-7.

The process is continued by retreating to another earlier branching site and continuing until all possibilities have been fully explored. In Fig. 1 we give the corresponding path graphs for the three non-equivalent vertices of norbornane graph. The corresponding atomic codes are derived by counting atoms in path graphs for each separation (distance):

1	3, 3, 4, 6, 2	18
2	2, 3, 4, 5, 5, 2	21
7	2, 4, 4, 4, 4, 4	22

The atomic path sums are of some interest of themselves, as they seem to reflect some salient structural features. In Fig. 2 we give these numbers for carbon atoms in a selection of norbonane derivatives. By substituting a single atom, we change the number of paths for the site of substitution by one. If two atoms are added, the path sum is increased by two, etc.

Proliferation of paths in polycyclic structures increases rapidly as the size of the graph and the number of edges increase . This is illustrated in Table 7 for the complete graphs, K_n. Here the regular increase in the number of paths of different length can be even summarized in a compact expression.

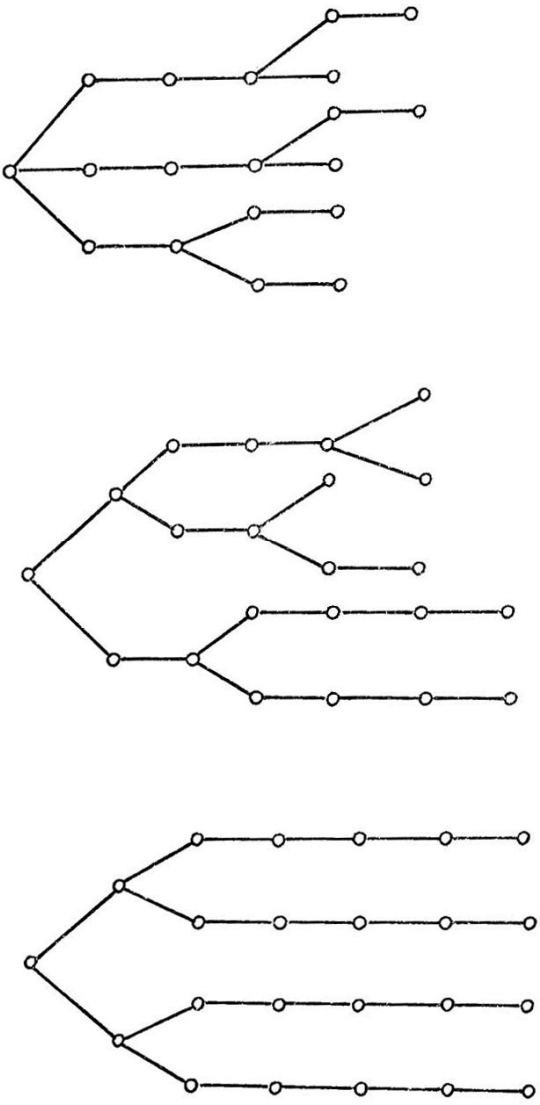
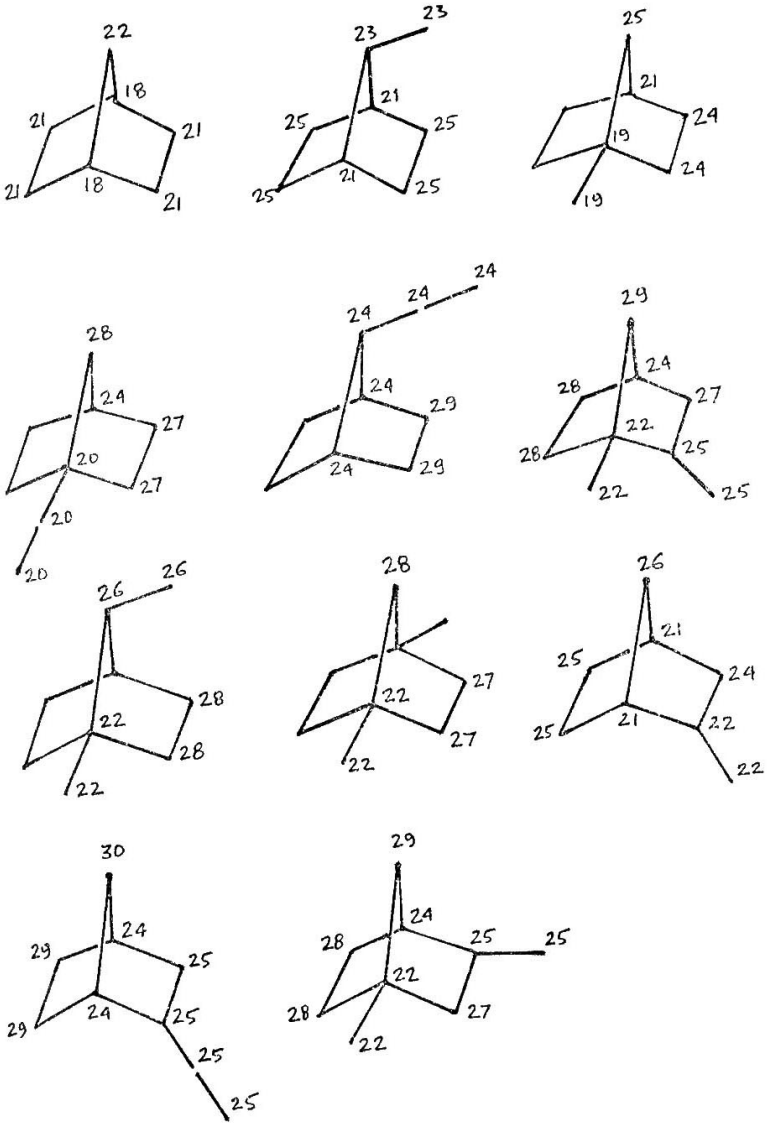


Fig. 1


22
18
21
21
21
18
21
23
23
21
25
25
25
21
25
25
25
21
24
19
24
19
28
24
27
20
27
20
20
24
24
24
29
24
29
29
24
24
29
28
24
27
28
22
25
22
25
26
26
28
22
28
22
28
22
27
27
22
26
21
25
24
25
21
22
22

30
24
29
25
29
24
25
25
25

29
24
28
25
25
22
28
27
22

Fig. 2

Table 7

The number of paths of different length for the complete graphs K_n

Graph	Number of paths of length: 1	2	3	4	5	Total number of paths
K_2	1					1
K_3	3	3				6
K_4	6	12	12			30
K_5	10	30	60	60		160
K_6	15	60	160	360	360	975
....	.....					.....
K_n	$\frac{n!}{2(n-1)!}$	$\frac{n!}{2(n-2)!}$	$\frac{n!}{2(n-3)!}$	....	$\frac{n!}{2!}$	$\frac{n!}{1!}$ $\frac{n!}{0!}$

It has been suggested[27] that the number of paths per vertex be considered a quantitative parameter characterizing the complexity of a graph. Such a definition of complexity measure applies equally to cyclic, polycyclic, and even acyclic structures. The latter would be excluded if the concept of complexity is restricted by considering the count of rings alone. From illustrations given it is clear that acyclic graphs have generally low complexity. This agrees with the common experience. However, a large acyclic graph, such as may result from an extensive search project, may be nevertheless more complex for detailed analysis than a simple unicyclic or bicyclic graph. The enumeration of path can properly reflect such differences. The fact that the complete graphs have simple compact expressions for paths of different length which allows predictions of path numbers without searching makes them useful as reference points of the "complexity" scale.

Conveniently the scale can be converted to a linear form by using logarithmic values of the total number of paths, instead the numbers themselves.

As another illustration of the growth of the number of paths with the size of the molecule and the number of rings in that molecule or structure, we list the number of path for a single vertex in regular solids in Table 8.
We see that an octahedron and a cube may be considered of a similar complexity and an order of magnitude more complex than a tetrahedron. Similarly a dodecahedron and an icosahedron are of approximately similar complexity with the same order of magnitude of the total number of paths and several orders of magnitude more complex than a cube, an octahedron, or a tetrahedron.

Table 8

Number of paths in regular solids

Polyhedron	Number of paths of increasing length					Total
Tetrahedron	3	3	6			60
Octahedron	4	12	28	48	40	729
Cube	3	6	12	18	30	
	24	18				888
Dodecahedron	3	6	12	24	42	
	78	144	240	408	654	
	936	1272	1626	1818	1806	
	1614	1140	552	162		250740
Icosahedron	5	20	70	230	710	
	1980	4740	9160	13360	13000	
	6320					595140

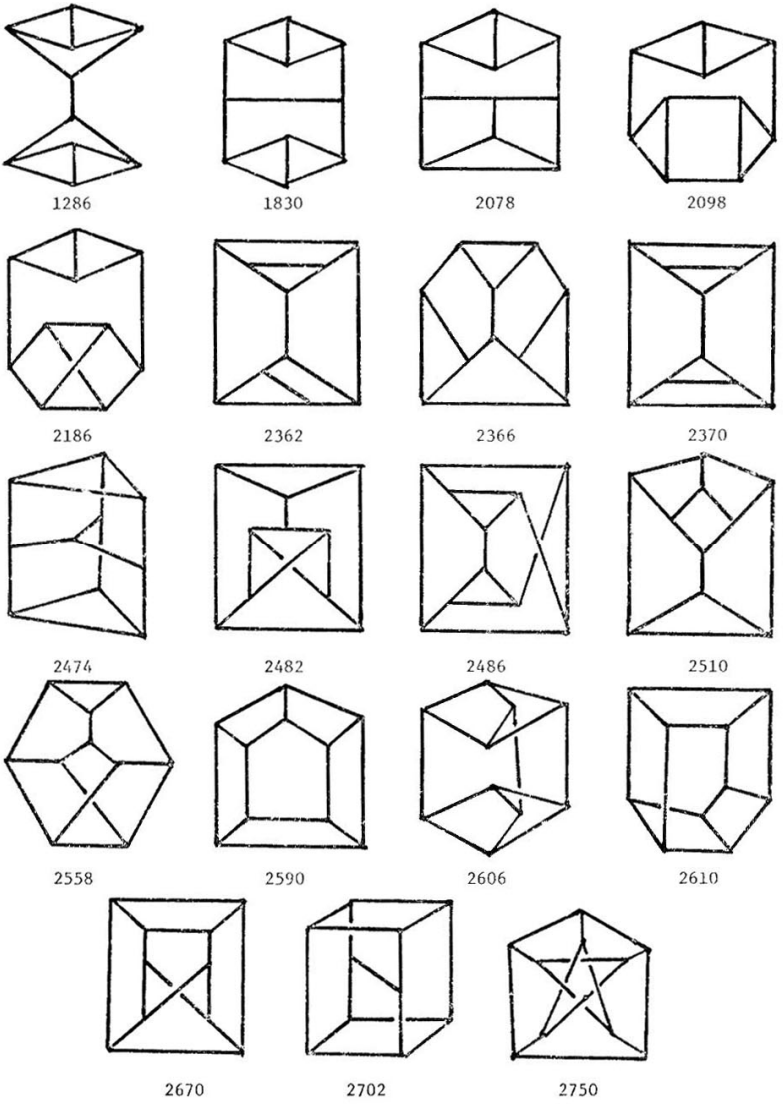

1286
1830
2078
2098
2186
2362
2366
2370
2474
2482
2486
2510
2558
2590
2606
2610
2670
2702
2750

Fig. 3

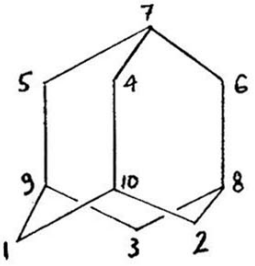
The total number of paths may also be of interest when one compares graphs of the same size. In Fig. 3 we show all the trivalent graphs with 10 vertices and the corresponding number of possible paths. As expected 1-connected and 2-connected graphs show the smallest number of paths, but for other graphs it is not apparent how the relative ordering based on the total number of paths will proceed. It is of some interest that the Petersen's graph appears the last in the list with the largest number of paths among trivalent regular graphs, hence, it is the most complex graph in the class, in the sense of the previously suggested measure of the complexity.

Molecular Path Numbers

When considering a molecule as a whole, one will seldom be interested in all the details of the atomic environments for individual atoms, and some contraction of the data is desirable. The most natural approach is to consider paths of different lengths for the molecule as a whole, rather than for separate atoms. Formally, one can derive the number of paths for a molecule by summing the corresponding entries of the atomic codes. In Table 9 we give the results for the adamantane carbon skeleton as derived using the available computer program.[27] The first entry for each atom is its label and the number one, which formally represents the number of paths of length zero. The first entry in the last line which represent the total, is the number of atoms in the molecule. There are only two non-equivalent sets of carbon atoms in adamantane, (1) - (6) and (7) - (10). Therefore, the first six rows are equal among themselves and the remaining four rows will also be alike. The last row, other than the first entry, is obtained by summing the previous rows (columnwise) and dividing the result by 2, since each atomic path has been encountered twice, once for each end atom. The summation resembles averaging, so that the molecular code becomes essentially equivalent to an average vertex code in a graph. For adamantane the average atomic code

CONNECTION MATRIX

0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	1	1	0
0	0	0	0	0	0	1	0	0	1
0	0	0	0	0	0	1	0	1	0
0	0	0	0	0	0	1	1	0	0
0	0	0	1	1	1	0	0	0	0
0	1	1	0	0	1	0	0	0	0
1	0	1	0	1	0	0	0	0	0
1	1	0	1	0	0	0	0	0	0



1									
1	2	4	4	8	8	12	8	12	0
2									
1	2	4	4	8	8	12	8	12	0
3									
1	2	4	4	8	8	12	8	12	0
4									
1	2	4	4	8	8	12	8	12	0
5									
1	2	4	4	8	8	12	8	12	0
6									
1	2	4	4	8	8	12	8	12	0
7									
1	3	3	6	6	12	6	12	0	0
8									
1	3	3	6	6	12	6	12	0	0
9									
1	3	3	6	6	12	6	12	0	0
10									
1	3	3	6	6	12	6	12	0	0
10	12	18	24	36	48	48	48	36	0

TOTAL PATHS 560

Table 9

is

1 2.4 3.6 4.8 7.2 9.6 9.6 9.6 7.2

To an atom one can now assign a parameter which would indicate how much the actual code departs from the average. The collection of such data for all atoms would be a measure of non-uniformity (heterogeneity) of the system.

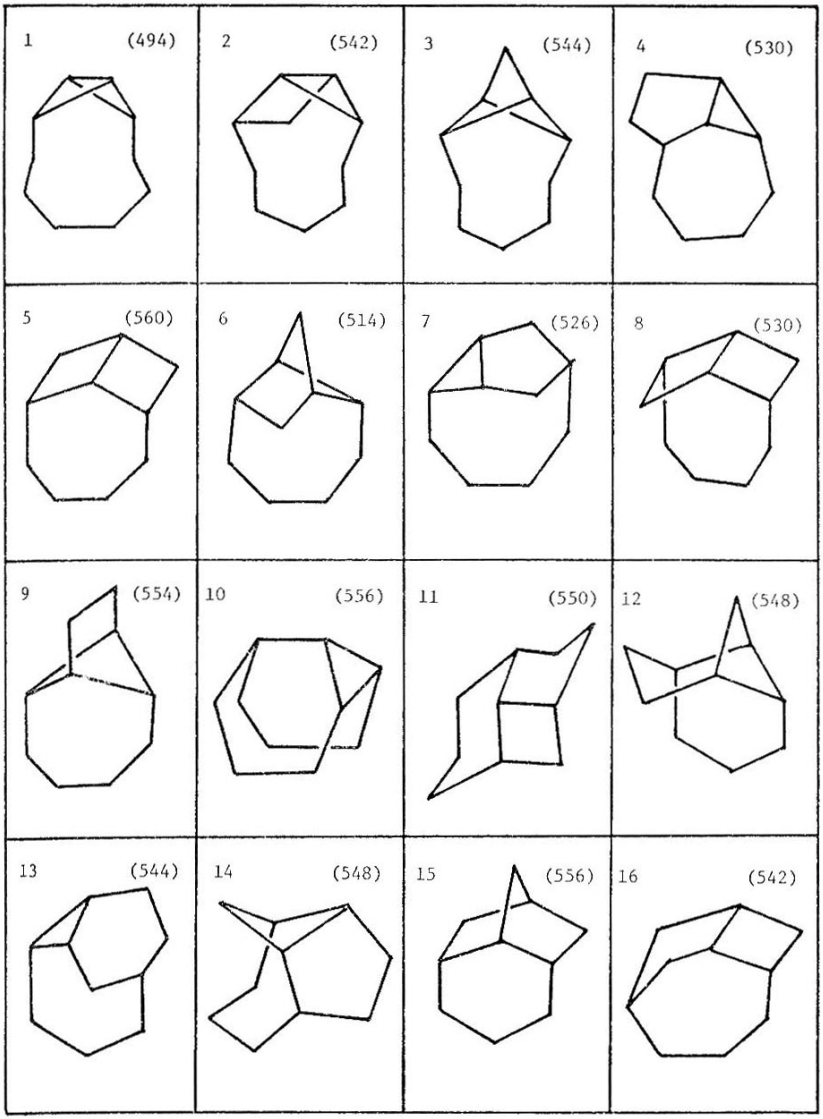
Molecular codes have useful applications, since they can now permit quantitative comparison among molecules. First of all, we find by application that different graphs generally produce different molecular path sequences. In Table 10 we list molecular codes for all tricyclic ten atomic saturated hydrocarbons having a same topology.[28] We have ordered the molecular codes lexicographically, that is according to the increasing values for paths. Since all isomers have the same number of atoms (p_0), bonds (p_1) and the pairs of adjacent bonds (p_2) they have been discriminated by the number of paths of length three (p_3).

Table 10

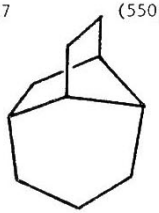
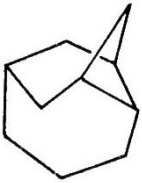
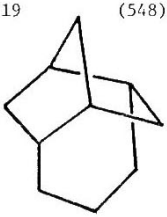
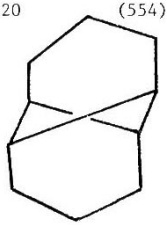
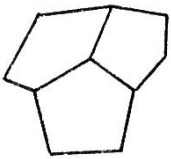
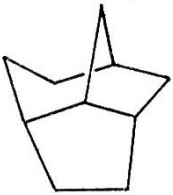
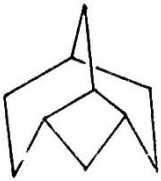
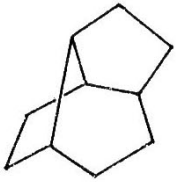
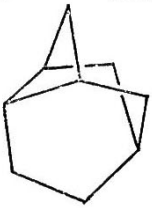
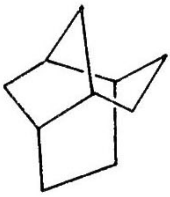
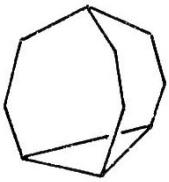
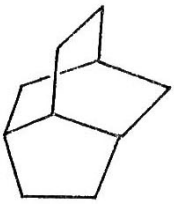
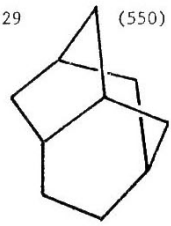
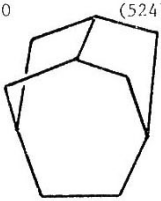
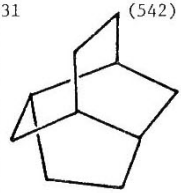
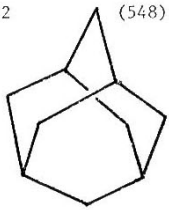
Molecular Codes for Tricyclic $C_{10}H_{16}$ Hydrocarbons

Molecule	Paths of Length									
	0	1	2	3	4	5	6	7	8	9
1	10	12	18	23	26	29	32	35	38	24
2	10	12	18	24	30	39	54	48	24	12
3	10	12	18	24	31	41	48	44	29	12
4	10	12	18	24	32	39	40	44	34	12
5	10	12	18	24	36	48	48	48	36	0
6	10	12	18	25	31	33	36	39	35	18
7	10	12	18	25	34	39	40	37	30	18
8	10	12	18	25	34	43	42	33	30	18
9	10	12	18	25	36	42	50	50	28	6
10	10	12	18	25	36	46	49	45	31	6

11	10	12	18	26	33	43	52	43	28	10
12	10	12	18	26	33	44	45	52	24	10
13	10	12	18	26	34	40	49	43	30	10
14	10	12	18	26	34	44	48	38	34	10
15	10	12	18	26	36	52	44	42	24	14
16	10	12	18	26	37	43	46	45	28	10
17	10	12	18	26	37	44	49	38	27	14
18	10	12	18	26	37	47	45	40	32	10
19	10	12	18	26	38	44	42	40	34	10
20	10	12	18	26	36	47	52	38	24	14
21	10	12	18	27	32	40	40	44	26	17
22	10	12	18	27	33	39	40	37	37	12
23	10	12	18	27	35	41	49	37	25	17
24	10	12	18	27	35	46	41	41	25	17
25	10	12	18	27	36	39	41	38	31	17
26	10	12	18	27	36	41	42	39	32	12
27	10	12	18	27	38	43	48	38	24	17
28	10	12	18	27	39	39	42	45	27	12
29	10	12	18	27	39	45	42	35	30	17
30	10	12	18	28	31	34	37	40	28	24
31	10	12	18	28	38	39	42	32	28	24
32	10	12	18	28	38	48	36	32	28	54



The total number of paths is shown in brackets.

17 (550) 	18 (554) 	19 (548) 	20 (554) 
21 (532) 	22 (530) 	23 (542) 	24 (544) 
25 (538) 	26 (538) 	27 (550) 	28 (542) 
29 (550) 	30 (524) 	31 (542) 	32 (548) 

When several isomers have the same p_3 they are ordered according to the increasing values of paths of length four, p_4, etc. Observe that the lexicographic ordering of the isomers does not coincide with an ordering that would be based on the total number of paths. The two schemes reflects different structural elements. In several instances, molecules have identical numbers of paths p_4 and p_5, and a comparison of the corresponding molecular skeletons reveal apparent similarity. Although each structure has been found to have a different sequence of molecular path numbers, the sequences among themselves differ to varying degrees. Intuitively similar skeletal forms appear to have sequences that are much alike. Hence the molecular codes based on path numbers offer a basis for a quantitative discussion of skeletal similarities. This approach has already been found useful[9] and has lead to the recognition that naturally occuring products, such as various terpenes, are generally more similar among themselves than to hypothetical terpenes generated by a computer. Similarity is of interest in discussions of structure-property and structure-activity correlations, for it provides a quantitative measure of the degree of differences between a structure-candidate and a standard.

As an illustration of the notion of the similarity among the molecular codes and molecular skeletons, let's consider a selection of pentacyclic molecules shown in Fig. 4. In Table 11 we list the corresponding molecular codes. Visual inspection of the table reveals that molecular codes of the structures A and B are more similar among themselves and somewhat different from the codes of the remaining molecules, since the numbers in the two sequences for paths of a same length are similar in magnitude.

Table 11

Molecular Path Sequences for Pentacyclic Molecules of Figure 4

Molecule	Path lengths											
1	2	3	4	5	6	7	8	9	10	11	12	13
A	18	30	50	82	113	146	180	212	223	184	137	94	38
B	18	30	50	82	112	146	182	214	218	182	140	98	34
C	18	30	49	80	109	141	178	219	221	187	154	88	18
D	18	30	48	78	106	136	176	218	218	212	166	64	8
E	18	30	48	77	104	140	176	218	242	203	140	72	14
F	18	30	49	79	107	144	185	216	225	198	136	80	29
G	18	30	48	76	102	138	188	226	234	208	130	64	24

The degree of similarity (or dissimilarity) can now be quantitatively derived. The individual entries in the above path-number sequences are viewed as coordinates of a point in n-dimensional space. The distance between points in such a space is a measure of similarity and can be obtained as the square root of the sum for the individual differences in path numbers squared. The results for the seven pentacyclic isomers considered here are shown in Table 12.

Table 12

Distances Among Pentacyclic Structures (A) - (G)

(A,B) = 8.89	(A,C) = 28.93	(A,D) = 60.59	(A,E) = 44.60	(A,F) = 23.85
(A,G) = 48.24	(B,C) = 26.36	(B,D) = 60,13	(B,E) = 47.69	(B,F) = 26.87
(B,G) = 51.62	(C,D) = 38.70	(C,E) = 34.73	(C,F) = 26.80	(C,G) = 44.97
(D,E) = 38.13	(D,F) = 44.65	(D,G) = 45.30	(E,F) = 27.09	(E,G) = 23.79
(F,G) = 26.04				

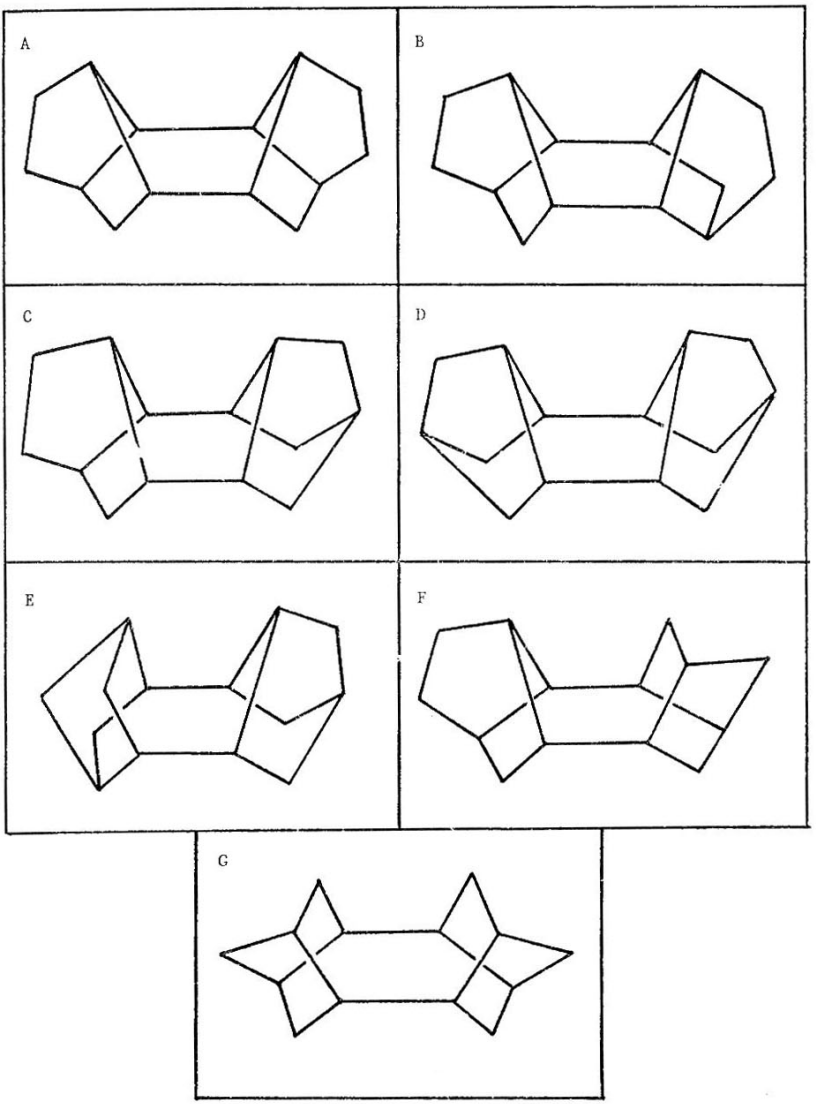


Fig. 4

From the Table 12 we see that the pair of structures A and B are indeed the most similar, which also agree with our intuitive judgement derived from a close inspection of the Fig. 4. Several pairs of structures would correspond to the next class of the most similar structures: (A,F), (E,G), and others found to have the distance below 30 units. The least similar among the collection of structures are the pairs (A,D) and (B,D), but even here the difference is not excessive.

Clearly the molecular path sequences contain considerable structural information in a very convenient form. Sequences are easy to compare and manipulate, yet the particular entries have simple structural interpretation. In acyclic systems the path sequences correspond to enumeration of neighbours at different distances, and one is not suprised to find that paths, as graph invariants, have found use in analysis of molecular data[29]. In a work on isomeric variations of thermodynamic properties of alkanes in 1947, Platt recognized the influence of neighbour bonds on additive bond properties[30]. Few years later he tabulated path numbers for the alkanes through octane, and suggested that these path numbers, which differ by a trivial factor of 2 from molecular codes reported here, "seem likely to be useful in future analysis of other properties of hydrocarbons". It seems appropriate, therefore, to refer to sequences of path numbers, at least in the case of acyclic graphs and in the discussions of molecular properties, as Platt's molecular codes. The importance of purely topological parameters has also been recognized by Wiener,[31] who felt compelled to include, in addition to the total number of paths in a molecule, the number of paths of length three, p_3, in his discussion of isomeric variations in molecular properties. Several other additivity schemes, to a greater or a lesser degree, also involve essentially the graph theoretical characterizations, although this is not explicitly indicated.[32]

Of the various uses of the molecular path codes, we will outline as an illustration the ordering of structures. From a topological point of view, the ordering of structures is of fundamental significance, though in practical applications it

may be of limited scope. Although we speak of ordering of structures, in fact we consider sequences and their ordering. This subject has received some attention in the mathematical literature. Muirhead,[33] at the beginning of this century prescribed a scheme which can resolve ambiguities in the problem of structure comparison. According to Muirhead, from a given sequence, one can construct a sequence of partial sums and uses the sequence of partial sums for the comparison. If all entries in one such sequence are greater than or equal to the corresponding entries in the other sequence the two sequences are comparable, and the structures can be ordered. If some entries of one sequence are greater while other partial sums are smaller, the sequences are incomparable, and the corresponding structures cannot be ordered in the above manner. This results in partial order. Karamata[34] has generalized the above considerations, from sequences on integers, to real numbers and has been able to prove a theorem that ensures that properties which can be expressed as continous and convex functions of the parameters in the sequence will follow the same (complete or partial as it may be the case) ordering. Hence the ordering of structures becomes immediately important to the study of regularities of molecular properties. This shows the significance of the ordering of structures, but for different chemical applications the rules for ordering have yet to be established.

We will outline possible orderings of alkanes that may reflect trends and regularities in their selected properties. First we have to resolve the problem of the rules for ordering the sequences. Let's consider the nine isomers of heptane. The atomic codes have been listed in Table 3. From the atomic codes given, one can derive molecular codes, from which it is not difficult to construct the sequences of partial sums. The results are summarized in Table 13.

Table 13

Molecular Codes and the Sequences of Partial Sums of Heptane Isomers

Molecule		Molecular codes	Partial sums
A	n-heptane	6, 5, 4, 3, 2, 1	6, 11, 15, 18, 20, 21
B	2-methylhexane	6, 6, 4, 3, 2, 0	6, 12, 16, 19, 21, 21
C	3-methylhexane	6, 6, 5, 3, 1, 0	6, 12, 17, 20, 21, 21
D	3-ethylpentane	6, 6, 6, 3, 0, 0	6, 12, 18, 21, 21, 21
E	2,3-dimethylpentane	6, 7, 6, 2, 0, 0	6, 13, 19, 21, 21, 21
F	2,4-dimethylpentane	6, 7, 4, 4, 0, 0	6, 13, 17, 21, 21, 21
G	2,2-dimethylpentane	6, 8, 4, 3, 0, 0	6, 14, 18, 21, 21, 21
H	3,3-dimethylpentane	6, 8, 6, 1, 0, 0	6, 14, 20, 21, 21, 21
I	2,2,3-trimethylbutane	6, 9, 6, 0, 0, 0	6, 15, 21, 21, 21, 21

We see that the last structure I, according to Muirhead's comparability scheme, dominates all other structures because every partial sum in the last row of the Table 13 is greater than or equal to the corresponding partial sum for any other structure. Similarly H dominates G and other structure above it. Again G dominates F, but not E, since the third partial sum is larger for E than for G. By examining all pairs of structures we arrive at the partial orders

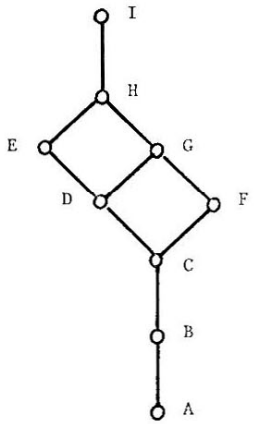
$$I > H > E > D > C > B > A$$

$$I > H > G > D > C > B > A$$

$$I > H > G > F > C > B > A$$

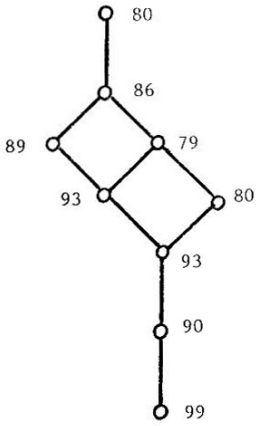
The hierarchical relationships implied in the above inequalities can be graphically displayed: We place at the top the structure that dominates others, if two structures cannot be compared they are placed in separate branches of the graph. For the seven isomers of heptane we obtain:

Fig. 5



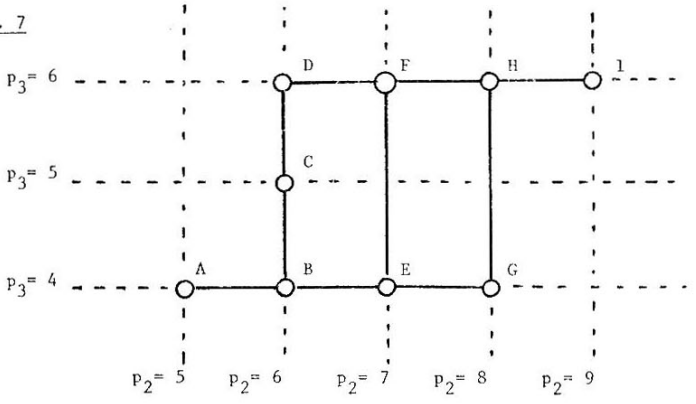
Following the Karamata theorem one can now replace each structure in the hierarchical diagram with a selected property, and if the particular ordering rule applies, one should obtain the same hierarchical relationship for that molecular property. In Fig. 6 we show the diagram for the boiling points of the heptane isomers A through I.

Fig. 6



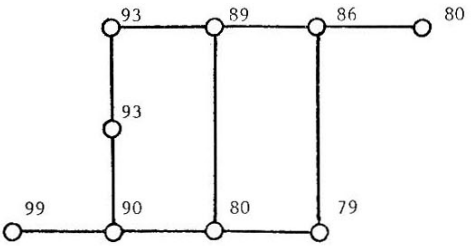
The result is not very impressive. In particular the relative magnitude of H with respect to G and F is discordant. Perhaps a different rule for the ordering should be applied. Since we expect that path numbers in a molecule will play an important role, let us focus our attention on the differences between path numbers among isomers. All heptane have $p_1=6$ (six CC bonds), but become different in p_2,p_3, etc. If we condider p_2 and p_3 as coordinates of a structure in a reduced two-dimensional "structure-space", we can represent pictorially each of the molecules as a point on a grid, where different structures are generally represented by different points. Fig. 7 shows the grid with the isomers A to I as points. The coordinate lines that connect various isomers have been emphasized.

Fig. 7



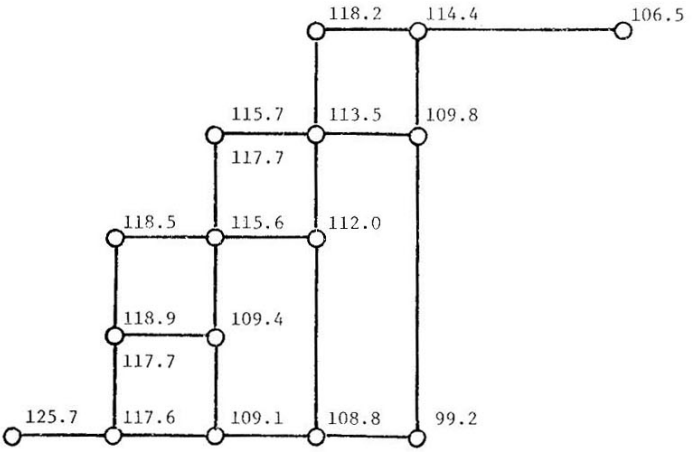
In Fig. 8 we replace each structure by the value for the boiling point of that isomer. We observe a better behaviour. The relative magnitudes for the selected property appear to show a regular trends and follow the hierarchy derived for the structures. For instance, as we move from left to right or from top to bottom along the coordinate lines the values of the boiling temperatures of heptane isomers decrease. The sample is not very large and the agreement may be fortuitous, but one can verify this ordering rule on larger alkanes.

Fig. 8



In Fig. 9 we show the corresponding grid graph for the 18 octane isomers with the boiling points replacing the molecules.[35]

Fig. 9



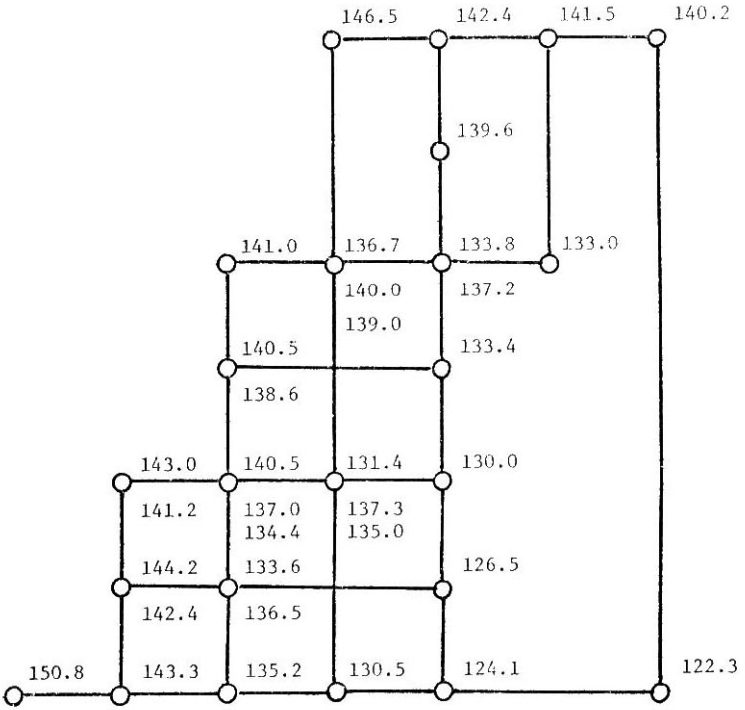
That the magnitudes of the boiling points show the same trends and regular change as we move along the coordinate lines, is now even more clearly seen.

We do have, however, in two instances molecules with the same number of paths p_2 and p_3, hence they occupy the same site on the grid. These are:

(p_2,p_3)	Molecules
(7,6)	3-methylheptane; 4-methylheptane
(8,8)	2-methyl-3-ethylhexane; 3,4-dimethylhexane

As we go to nonane isomers, with 35 molecules, we again find the same regular behaviour of the isomer boiling points when ordered as prescribed[36], that is by replacing the structure which has fixed position on the coordinate grid with paths of length two and three functioning as the coordinate axes.

Fig. 10



In Fig. 10 we illustrate the grid graph derived for nonane isomers. Now two and even three molecules can occupy the same

grid site:

(p_2,p_3)	Molecules
(8,7)	3-M; 4-M
(9,7)	2,4-MM; 2,5-MM; 3,5-MM
(8,8)	3-E; 4-E
(9,8)	2,3-MM; 3,5-MM; 2,4-ME
(10,8)	2,3,5-MMM; 4,4-MM; 3,3-MM
(9,9)	2,3-ME; 3,4-MM
(10,10)	2,3,4-MEM; 2,3,4-MMM; 3,3-ME
(10,11)	2,3,3-MMM; 2,2,3-MME

where M = methyl and E = ethyl, with the letters corresponding to the preceeding sequence of atomic numbers. It is clear that as the size of a molecule increases, we will have even more cases of identical coordinates. This, however, is not necessarily a disadvantage. Relatively speaking, as the size of a molecule increases, the variation becomes less and less important, at least among some of the molecules. Notice that the variation of the boiling points for nonane is almost $30^{o}C$, while isomers having the same (p_2,p_3) usually differ in their boiling points by less than $3^{o}C$. Of course, the enumeration of paths does not encode all structural features, and the variations within the group of isomers having the same p_2 and p_3 may have the origin in other molecular characteristics. In fact, it is suprising how much of the regularities in the observed data can be attributed to these simple graph-theoretical invariants.

Among nonane isomers we find a pair of structures which have all path numbers equal. These are 2,3,4-trimethylhexane and 3,3-methylethylhexane with the code 8, 10, 10, 6, 2. These molecules can be refered to as isocodal, and are expected to show considerable similarity in selected molecular properties.

Such molecules are of special interest since differences among their properties clearly point to contributions originating from factors which are not accounted in path enumerations such as non-bonded interactions, stereospecifity, etc.

In the case of 75 isomers of decane we find the grid graph with 39 points[36], hence already some crowded sites, which makes a reference to the individual structures on the grid problematic. The grid and the coordinates (p_2,p_3) should be used for classification of the large number of structures in smaller lots which would show some similarity among themselves. Attempts to classify (i.e., group together in smaller lots) such a sizable set of molecules on the basis of some experimental data is conceptually doubtful and can be quite misleading, since structurally different molecules may have similar measured properties. For instance, frequently the largest boiling points, besides characterizing n-alkane, a linear chain, also belong to several highly branched isomers. Our approach to classification is of structural origin, conceptually simple and devoid of ambiguities.

Among the 75 isomers of decane we find one isocodal pair: 2,4-dimethyl-4-ethylhexane and 2,2-dimethyl-3-ethylhexane with the code 9, 12, 11, 9, 4. As the number of carbon atoms increases, one may also expect that such coincidental sequences of path numbers would increase. Among the 159 undecanes, $C_{11}H_{24}$, there are ten pairs and one triplet of isocodal structures (listed in Table 14), indicating limitations of the molecular path number concept to discriminate among molecules. However, in all cases where molecular codes are the same, the molecules differ in the complete list of atomic codes. This supports the conjecture that the list of atom codes may be unique to the structure.

For nonanes, decanes, and undecanes the ratio between the number of isomers and the number of distinctive grid points is 1.45, 1.92, and 3.18 respectively. These numbers correspond to the average number of molecules in a class. For larger alkanes the average size of classes will increase further. Not all classes, however, will be numerous, some still contain a single

molecule, while, for instance for undecanes several classes having eight and nine molecules can be found. In Fig. 11 we illustrate members of several classes of undecane isomers, from which one sees some common structural elements in some cases.

Fig. 11

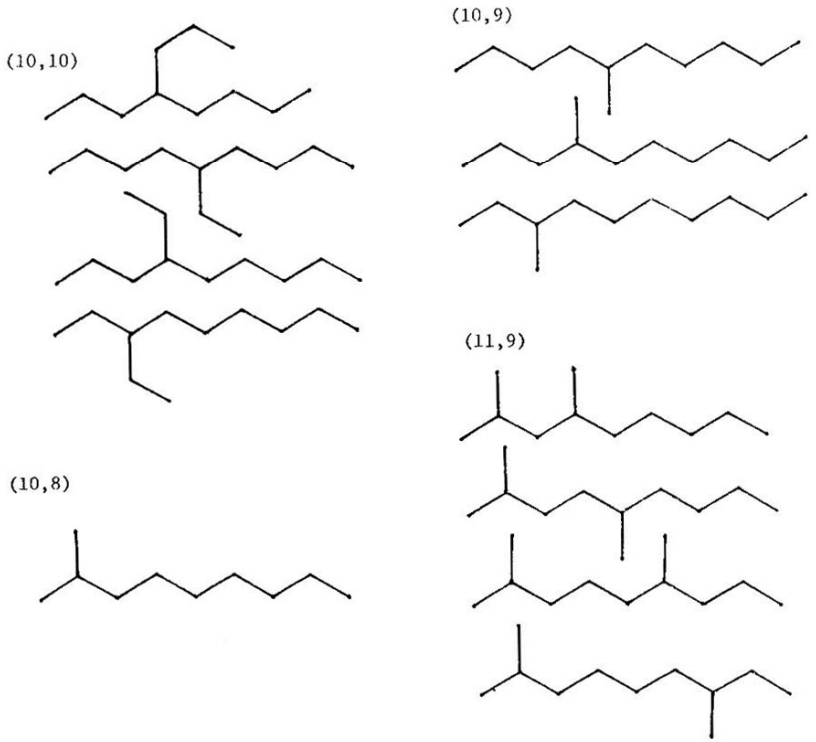


Table 14

Code	Isocodal structures
10, 11, 11, 10, 8, 4, 1	4-isopropyloctane
	3-methyl-5-ethyloctane
10, 12, 14, 12, 6, 1	4,4-diethylheptane
	3,5-dimethyl-4-ethylheptane
10, 13, 12, 10, 6, 4	2,2-dimethyl-3-ethylheptane
	2,3,4,6-tetramethylheptane
10, 13, 16, 12, 4	2-methyl-3,3-diethylhexane
	3-methyl-3,4-diethylhexane
10, 13, 14, 11, 6, 1	4-methyl-4-isopropylheptane
	3,3-dimethyl-4-ethylheptane
10, 13, 14, 12, 6	2-methyl-4,4-diethylhexane
	2,3,5-trimethyl-4-ethylhexane
10, 14, 13, 9, 7, 2	2,4,5,5-tetramethylheptane
	2,3,3,5-tetramethylheptane

10, 14, 16, 11, 4	2,3-dimethyl-3-isopropylhexane
	2,3,3-trimethyl-4-ethylhexane
	2,3,4-trimethyl-4-ethylhexane
10, 14, 14, 11, 6	2,3,5-trimethyl-3-ethylhexane
	2,2,3-trimethyl-4-ethylhexane
10, 15, 18, 12	2,2-dimethyl-3,3-diethylpentane
	2,3,4-trimethyl-3-isopropylpentane
10, 15, 16, 10, 4	2,2,3-trimethyl-3-ethylhexane
	2,3,3,4,5-pentamethylhexane

On the Classification of Molecules

Full characterization of a class of molecules remains to be examined. Here we define the class as the collection of molecules (already preselected, such as all isomers of a given size, cyclization etc) having same selected path numbers. In particular we will confine the discussion to classification based on p_2 and p_3 and will continue with examination of undecanes. For a representative characterization of a class one can consider an average class code, which can be derived from path numbers for the members of a class. In Table 15 we have listed for several classes of undecanes the corresponding members (by simply giving their molecular path sequences, repeating the sequence if more than one member has a same code) and also derived the average path numbers for each class. Just as before, when we introduced the molecular path sequences by summing the atomic contributions, so now class codes are derived by summing (and averaging) the contributions from the individual members of the class. One may expect that such average-class codes may play similar role when one compares distinctive classes and their properties as molecular codes served for comparisons of molecules and molecular properties. The example of undecane isomers serve to illustrate the concept of class characterization, but the first question to be considered in other applications is that of defining the class. In case of alkanes we have been lead by successful ordering of structures to recognize paths p_2 and p_3 as the dominant contributions to molecular additive properties. In other situations molecules that form a class or a family may be selected on different grounds. Nevertheless, once molecules have been selected on some basis, if that involves structural considerations, it is likely that similarity in molecular codes for the members of the class will be preserved. The validity of any such classification and its characterization will be reflected in the variations among the individual codes and the average class code. Having this simple possibility to check whether the selected structures can be

Class: (10,10)

10,	10,	10,	8,	6,	5,	4,	2
10,	10,	10,	9,	7,	5,	3,	1
10,	10,	10,	9,	8,	5,	2,	1
10,	10,	10,	10,	8,	5,	2	
10,	10,	10,	9,	7.25	5,	2.75	1

Class: (11,9)

10,	11,	9,	7,	6,	5,	5,	2
10,	11,	9,	8,	6,	6,	3,	2
10,	11,	9,	8,	8,	4,	3,	2
10,	11,	9,	9,	6,	5,	3,	2
10,	11,	9,	8,	6.5	5,	3.5	2

Class: (11,10)

10,	11,	10,	7,	6,	5,	4,	2
10,	11,	10,	7,	6,	6,	4,	1
10,	11,	10,	8,	6,	6,	4	
10,	11,	10,	8,	7,	5,	3,	1
10,	11,	10,	9,	7,	4,	3,	1
10,	11,	10,	9,	8,	5,	2	
10,	11,	10,	10,	6,	5,	2,	1
10,	11,	10,	10,	8,	4,	2	
10,	11,	10,	11,	8,	5		
10,	11,	10,	8.78	6.89	5,	2.67	0.67

Class (11,11)

10,	11,	11,	8,	6,	5,	3,	1
10,	11,	11,	8,	7,	6,	2	
10,	11,	11,	9,	6,	5,	3	
10,	11,	11,	9,	7,	4,	2,	1
10,	11,	11,	10,	7,	4,	2	
10,	11,	11,	10,	8,	4,	1	
10,	11,	11,	10,	8,	4,	1	
10,	11,	11,	9.14	7,	4.57	2	0.29

Class (11,12)

10,	11,	12,	10,	6,	4,	2	
10,	11,	12,	10,	7,	4,	1	
10,	11,	12,	10,	8,	4		
10,	11,	12,	11,	7,	3,	1	
10,	11,	12,	11,	8,	3		
10,	11,	12,	10.4	7.2	3.6	0.8	

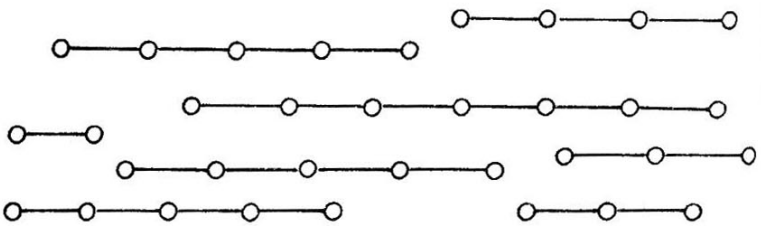
Class (12,10)

10,	12,	10,	7,	6,	5,	4,	1
10,	12,	10,	7,	6,	6,	4	
10,	12,	10,	9,	6,	5,	2,	1
10,	12,	10,	9,	8,	3,	2,	1
10,	12,	10,	10,	6,	5,	2	
10,	12,	10,	11,	8,	4		
10,	12,	10,	8.83	6.67	4.67	2.33	0.5

grouped together and considered structurally related allows one to apply the concept of class-codes to arbitrary collection of structures, which can then be pruned from undesirable members by setting the level of tolerance for the class.

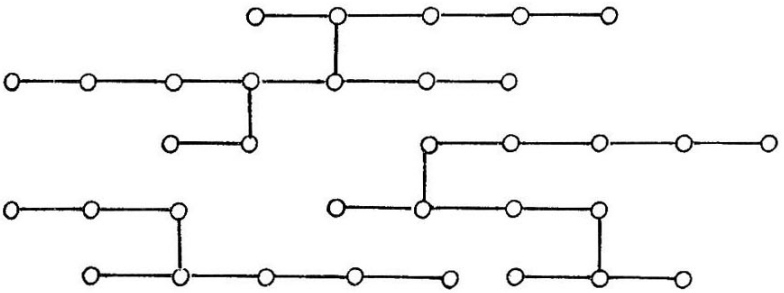
The concept of the average code can also be viewed as means for characterization of composite systems of varying degree of inhomogenuity. If we have a collection of N identical molecules as a model of a liquid or gas, the average system code would be that of the molecules from which the system is composed. A minor "impurity", that is the occasional presence of a molecule of a different kind, will hardly affect the system-code. However, if the number of different molecules represents significant fraction and we have a mixture, rather than homogenous system, the change of the character of the system will manifest itself in the average code obtained by taking into account contributions from all molecules present. One can see that the gradual change in the composition will similarly gradually alter the average path numbers. Hence, the average codes can serve for characterization of inhomogenous systems as well. In order to illustrate the use of such average codes, let's consider a simple model system for cross-linking in polymers. We have selected a sample of 34 units (which can stand for atoms or some monomer units) distributed in entities having from 2 to 7 units in a linear chain.[37] One such arrangement is shown in Fig. 12.

Fig. 12



We consider now various arbitrary cross-linkings that can appear in the system shown in Fig. 12, allowing the individual components to change their relative positions. Because of the convenience of use of small computer we have limited the size of cross-linked structure to 25 units, but this restriction can be easily lifted, and considered such resulting structures that follow from introduction of five cross links at random. Fig. 13 illustrates one such simulated cross-linking of polymers.

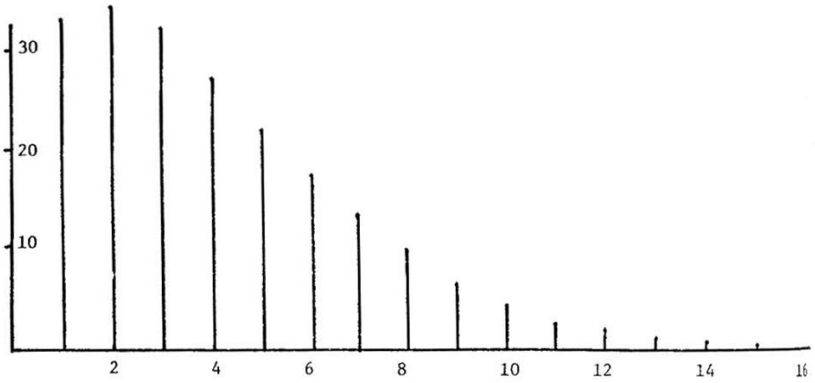
Fig. 13



In this way, one generates random-like polymers, and can study their characterization. The variables at our disposal, within the model considered, are the sizes of the initial components and the distribution of various sizes. The initial structures may all be of the same form, and of course, need not be always linear chains. One may consider various cyclic structures with or without pendant bonds. Also at our disposal is the assumption on the number of cross-links and their functionality (valency). In characterizations of polymers usually the singular most important parameter is the molecular weight, the average molecular weight, to be precise. Then, at least in simulated model calculations as the above described system, one can study the distribution of various sizes. If the sample is sufficiently large and allow valid statistical analysis one may

take the size distribution as useful characterization of the system. By simulating some 20 random cross-linking systems we find polymer sizes of all sizes, from 2 to 25. However the size of the sample (20 trials) is by far too small to even discern which, if any, of the sizes may be preferred. Perhaps a sample of some 200 trials may provide some insight into the distribution of polymer sizes. In contrast, the 20 simulations appear to be sufficiently large sample for deriving the average code for selected model polymerization. The reason for this is that the number of paths is several orders of magnitude larger than the number of polymer units, hence a rather regularly shaped distribution of paths of different lengths can be derived from relatively small sample. The resulting distribution is pictorially represented in Fig. 14.

Fig. 14



One can see that already a sample of only 10 simulated polymerizations produces an average code for the model which hardly differs from that based upon 20 such simulations. This assures that the average codes for otherwise individually appreciably

different systems will give useful characterization. The average system code, which represents a distribution of paths of different length within the system, is not only useful by being derived from a smaller sampling. It contains more information than a seemingly similar distribution of sizes. It is this ability of molecular codes based on enumeration of paths of different length, to summarize some structural features in a simple and interpretative way which indicates that the use of molecular codes may find widespread application. With this note of optimism one may anticipate the opening of yet another area of graph-theoretical applications to structural chemistry.

Acknowledgements

This work was supported in part by U. S. Department of Energy, Division of Basic Energy Sciences, administrated by Ames Laboratory. I would like to thank Professor R. S. Hansen for his interest in this work, in particular for helpful discussions concerning model polymerization. Dr. L.J. Dorgan kindly examined the manuscript and suggested numerous improvement in the presentation. Last but not least I would like to thank The University of Bremen and Professor P. J. Plath for the financial assistance which made my attendance to the Symposium possible and all the participants for memorable time.

References

1 A. Cayley, Phil.Mag. 47, 444 (1874)
F. Flavitzky, J.Russ.Chem.Soc., 160 (1871)

2 S.M. Lozanić, Rad Jogoslav.Akad. (Zagreb), Math.Sci.class 26 (1897)
S.M. Lozanić, Chem.Berichte, 30, 1917 (1897)

3 C.A. Coulson and G.S. Rushbrooke, Proc.Cambridge Phil.Soc. 36, 193 (1940)

4 A.T. Balaban (ed.), Chemical Applications of Graph Theory, Academic Press, London (1976)
A. Graovac, I. Gutman and N. Trinajstić, Topological Approach to Chemistry of Conjugated Molecules, Springer-Verlag (Lecture Notes in Chemistry, #4), Berlin (1977)

5 R.C. Read and D.G. Corneil, J. Graph Theory, 1, 339 (1977)

6 G. Pólya, Acta Math., 68, 145 (1938)

7 M. Randić, Chem.Phys.Lett., 58, 181 (1978)

8 M. Randić and C.L. Wilkins, Chem.Phys.Lett., (in press)

9 M. Randić and C.L. Wilkins, J.Chem.Inf.Computer Sci., (in press)

10 M. Randić and C.L. Wilkins, J.Amer.Chem.Soc., (submitted)

11 D.E. Knuth, Science, 194, 1235 (1976)

12 M.N. Barber and B.H. Ninham, Random and Restricted Walks, Gordon and Breach, New York (1970), chapter 7

13 R.H. Penny, J.Chem. Doc., 5, 113 (1965)

14 M. Gordon and J.W. Kennedy, J.C.S. Faraday II, 69, 484 (1973)

15 J.E. Dubois, Ordered Chromatic Graph and Limited Environment Concept, a chapter in Chemical Applications of Graph Theory (ref 4)

16 J.E. Dubois, D. Laurent and A. Aranda, J.Chim.Phys., 70, 1608 (1973)
J.E. Dubois and J. Chretien, J.Chromatogr.Sci., 12, 811 (1974)

17 D.M. Grant and E.G. Paul, J.Amer.Chem.Soc., 86, 2984 (1964)

18 M. Randić, J.C.S. Faraday II, (submitted)

19 M. Randić, J.Chem.Inf.Computer Sci., 18, 101 (1978)
M. Randić and C.L. Wilkins, J.Chem.Inf.Computer Sci., (in press)

20 F.E. Harris, private communication (1978)

21 P. Erdös and T. Gallai, Mat.Lapok, 11, 264 (1960) (in Hungarian)
V. Havel,Casopis Pest. Mat., 80, 477 (1955)(in Czech)
S.L. Hakimi, Jour.Frank.Inst., 279, 290 (1965)
W.K. Chen, Jour.Frank.Inst., 281, 406 (1966)

22 A.T. Balaban and F. Harary, J.Chem.Doc., 11, 258 (1971)
T. Zivkovic, N. Trinajstić, and M. Randić, Mol.Phys., 30 517 (1975)
W.C. Herndon, Tetrahedron Lett., 671 (1974)

23 The list of connectivities originates with P.A. Morris (Trinidad, West Indies) who tabulated characteristic polynomials of trees on up to 14 vertices

24 J. Turner, SIAM, J.Appl.Math., 16, 520 (1968)

25 C.A. Shelley and M. Trulson, private communication (1978) used the molecule assembler in program CASE (Computer-Assisted Structure Elucidation). The number of isomers verified is: $C_{11}H_{24}$ - 159 isomers; $C_{12}H_{26}$ - 355 isomers, $C_{13}H_{28}$ - 802 isomers and $C_{14}H_{30}$ - 1858 isomers

26 The Petersen's graph is of considerable interest in chemistry, because it depicts relationships among different penta-coordinated complexes having five different ligands. C.f.: J.D. Dunitz and V. Prelog, Angew.Chem., 80, 700 (1968)

27 M. Randić, G.M. Brissey, R.B. Spencer, and C.L. Wilkins, Chem.and Computers (in press)

28 N. Tanaka, T. Iizika , and T. Kan, Chem.Lett. (Japan), 539 (1974)

29 For different applications see:
K. Altenburg, Kolloid Zeit., 178, 112 (1961)
M. Randić, Chem.Phys.Lett., 53, 602 (1978)

30 J.R. Platt, J.Chem.Phys., 15, 419 (1947)

J.R. Platt, J.Phys.Chem., 56, 328 (1952)

31 H. Wiener, J.Amer.Chem.Soc., 69, 17, 2636, (1947)

32 W.J. Taylor, J.M. Pignocco, and F.D. Rossini, J. Research Natl. Bur. Standard, 34, 413 (1945)

33 R.F. Muirhead, Proc.Edinburg Math.Soc., 36, 21 (1903), c.f: G.H. Hardy, J.E. Littlewood, and G. Polya, Inequalities, Cambridge Univ. Press, London (1934) p. 44

34 J. Karamata, Publ.Math.Univ. Belgrade, 1, 145 (1932), c.f: E.F. Beckenbach and R. Bellman, Inequalities, Springer-Verlag, Berlin (1961)

35 We have limited our example to the boiling points, however many other thermodynamic properties show similiar regularities. These include: heats of atomization, densities, indices of refraction, critical density and critical pressure, molar magnetic susceptibilities, specific dispersion, molar volumes, surface tension, heat of combustion, critical temperature and critical volume, etc. For details see ref. 36

36 M. Randić, and C.L. Wilkins, J.Phys.Chem., (submitted)

37 M. Randić, work in progress