

TOPOLOGY OF MOLECULES

by

Joshua Lederberg  
Department of Genetics  
Stanford University School of Medicine  
Stanford, California 94305

Pages 37-51 from

# The Mathematical Sciences

*A Collection of Essays*

*Edited by the National Research Council's*

Committee on Support of Research in the Mathematical Sciences (COSRIMS)  
*with the collaboration of George A. W. Boehm*

Published for the  
National Academy of Sciences—National Research Council by

*The M.I.T. Press*

MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
CAMBRIDGE, MASSACHUSETTS, AND LONDON, ENGLAND

1969

The structures of organic molecules are sometimes so bewilderingly complex that chemists have difficulty describing, classifying, and even naming them. Graph theory, a special tool borrowed from topology, has now been used to reduce even quite complicated chemical structures to a chain of numbers so that a computer can analyze them. This attempt to make organic chemistry more systematic could make it much easier for students to learn basic principles and to solve vexatious problems of classifying chemical compounds so that computers could be more readily applied to retrieve chemical information. It may be a forerunner of similar mathematical simplifications that will be applied to chemical genetics and other much more complex fields.

## Topology of Molecules

*Joshua Lederberg*

The enterprise known as science rests on two pediments: the power and social utility of empirical knowledge, and the esthetic satisfaction that comes from an elegant restatement of principles. These views have been contrasted as the Baconian versus Newtonian justifications of science. Newton's name evokes a very apt image, his epochal contributions to the mathematical formulation of physics. Some esthetes judge how far a science has advanced in its development by the extent to which it has been mathematized — made into a deductive science by a set of axioms and rules for their manipulation.

The fruitfulness of pursuing such an aim is debatable for such fields as embryology, genetics, or psychology. Outside the rather special area of evolutionary theory, few examples of useful prediction are based upon any comprehensive mathematization of living behavior. On the other hand, for many special situations, models can be created that are sufficiently simplified to justify the application of some numerical mathematics or statistics. In his essay in this volume, Hirsh Cohen has discussed many examples of this kind of application of mathematics to biology and medicine.

With the rapidly growing speed, size, and availability of digital computers, the esthetic ideal of rationalizing a science acquires a new dimen-

sion of practical importance. If we could give biology sufficient formal structure, it might be possible to mechanize some of the processes of scientific thinking itself. Many of the most striking advances in modern biology have come about through the formulation of some spectacularly simple models of important processes, for example, virus growth, genetic replication, and protein synthesis. Could not the computer be of great assistance in the elaboration of novel and valid theories? We can dream of machines that would not only execute experiments in physical and chemical biology but also help design them, subject to the managerial control and ultimate wisdom of their human programmer.

This vision is so far beyond our present grasp, it makes what will be reported below seem quite trivial. These remarks may, however, give some notion of the reasons a geneticist took an interest in the formalization of organic chemistry. Chemical genetics embodies many statements in natural language, and its reasoning embodies an enormous range of expertise covering chemistry, geometry, and most of the natural sciences, as well as that most difficult realm, common sense. As a further complication, many quite fundamental discoveries are being reported almost daily. I wanted more experience with the mechanization of a simpler science before tackling chemical genetics. A scan across neighboring disciplines suggested that elementary organic chemistry might be a challenge that was more amenable yet had not been exhausted.

For various reasons, including the good fortune of my association with Professor Carl Djerassi of Stanford's Chemistry Department, the analysis of mass spectral data for the solution of structural problems in organic chemistry was taken as the focal process for which a formalization would be attempted. Equally fortunately, Professor E. A. Feigenbaum joined the faculty of Stanford's Computer Science Department, and the entire effort of translating the formalisms and developing the heuristics for implementation on the computer has been done in close collaboration with him.

We may now turn to a consideration of the application of some elementary nonnumerical mathematics, that is, graph theory, for the representation of organic molecules. The use of these representations for a computer mechanization of the concepts of organic structural analysis will be summarized briefly.

The mathematical tool for translating chemical structures into a form that a computer can handle digitally is a concept that topologists call a *graph*. This kind of graph has little relation to the curves and bar charts used to display data; rather, it is a formal diagram for analyzing connections among a number of entities, in this case the individual atoms that make up an organic molecule.

Graphs have two components: *nodes* (representing atoms) and *edges* (chemical bonds between atoms). Each edge is associated with exactly two nodes, each node with at least one edge. The lengths of the edges are irrelevant. Disconnected graphs are regarded as representing molecules

that are distinct, even if they are bound by diffuse chemical forces as in a crystal.

Our main approach is mapping: a rule of correspondence between a part of a chemical structure and a part of some abstract graph. Graphs lend themselves to canonical forms, that is, methodical choices among equivalent representations according to a precise rule, which eliminates ambiguity and redundancy. The objective is to represent each molecular structure by just one graph and, conversely, to have each graph represent just one structure. Chemistry will re-emerge after a few levels of abstraction.

The structural formula for an organic molecule is then a paragon of a topological graph, that is, the connectivity relations of a set of chemical atoms we take as the nodes of the graph. True, we recognize more than one type of connection — double, triple, and noncovalent bonds, as well as single bonds. From an electronic standpoint, however, the special bonds could just as well be denoted as special atoms. The structural graph does not specify the bond distances and bond angles of the molecule. In fact, these are known for only a small proportion of the enormous number of organic molecules whose structure is very well known from a topological standpoint.

Most of the syllabus of elementary organic chemistry, thus comprises a survey of the topological possibilities for the distinct ways in which sets of atoms may be connected, subject to the rules of chemical valence. The student then also learns rules that prohibit some configurations as unstable or unrealizable. (He may later earn his scientific reputation by justifying or overturning one of these rules.) But the field of organic chemistry has reached its present stature without many benefits from any general analysis of molecular topology. These benefits might arise in applications at two extremes of sophistication: teaching chemical principles to college undergraduates and teaching them to electronic computers. They may also apply to the vexatious problems of nomenclature and systematic methods of information retrieval.

Although the topological character of chemical graphs was recognized by the first topologists, very little work has been done on the explicit classification of graphs having the greatest chemical interest. Some difficult problems, e.g., the analytical enumeration of polyhedra, remain unsolved.

This article will, then, review some elementary features of graphs that may be used for a systematic outline of organic chemistry.

### All the Ways to Build a Molecule

A problem statement might be: Enumerate all the distinct structural isomers of a given elementary composition, say  $C_3H_7NO_2$ . That is to say, produce all the connected graphs that can be constructed from the atoms of the formula, linked to one another in all distinct ways, compatible with

the valence established for each element (4, 3, 2, 1 for C, N, O, H, respectively). For compactness, H can be omitted from the representations, being implied by every unused valence of the other atoms.

The first discrimination is between trees and cyclic graphs, the "aliphatic" versus the "ring" structures of organic chemistry. Trees are graphs that can be separated into two parts by cutting any one link. How may we establish a canonical form for a tree, after first noting its order (number of nodes)?

The first step might be to find some unique place to begin the description. A tree must have at least two terminals and may have many more if highly branched; these are, therefore, not suitable starting points. However, each tree has a unique center. In fact, in 1869 Jordan showed that any tree has two kinds of center, a mass center and a radius center. Each center has a unique place in any tree; the two may or may not coincide.

To find the radius center, the tree is pruned one level at a time, cut back one link from every terminal at each level. This will leave, finally, an ultimate node or node pair (in effect, edge) as the center, the radius not of a length but, rather, of levels of pruning needed to reach the center.

To identify the mass center of a tree, we must consider the two or more branches that join to each nonterminal node. The center is the node whose branches have the most evenly balanced allocation of the remaining mass (node count) of the tree. This is the same as saying that none of the pendant branches exceeds half of the total mass. If the structure is a union of equal halves, the center is the edge that joins them.

Each of the centers (Figure 1) is unique and so could solve our problem of defining a canonical starting point of a description. The center of mass is more pertinent to finding a list of isomers, which of course have the same mass. The radius center is ill-adapted for this but matches conventional nomenclature, which is based on finding the longest linear path, that is, a diameter.

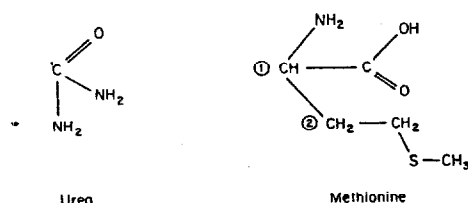
In chemical terms, the center divides the graph into two or more radicals. These radicals can be ordered by obvious compositional principles, giving rise to a canonical description of the whole graph in a linear code. Thus, methionine becomes  $(C(N)(C(O)(=O))(C-C-S-C))$  or, in a parenthesis-free notation the example should make obvious,  $C \cdot \cdot \cdot NC \cdot \cdot \cdot OOC \cdot C \cdot S \cdot C$ . This is more legible to the human reader, if the implied hydrogens are restored, as



Any linear code has an implicit numbering system: Each atom is numbered according to the place where it occurs in the string.

Some thirty years ago, Henze and Blair showed how Jordan's principle could be used for the enumeration of isomers of saturated hydrocarbons

and some simple derivatives of them. Here, the nodes are all carbon atoms, and the enumeration can proceed by working outward from smaller to larger complexes. For example, for the isomers of undecane,  $C_{11}H_{24}$ ,

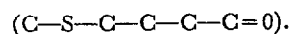


Urea

Methionine

FIGURE 1. *Chemical trees and their centers.*  
 In urea, the carbon atom is both the radius center and the mass center.

In methionine, carbon atom 1 is the mass center, according to the numerical partition 1 . . . 1 3 4. Carbon atom 2 is the radial center, on a diameter of 7, that is, the center of a largest string



For both analyses, we ignore hydrogen atoms.

one atom is designated as center, leaving 10 to be allocated among 2, 3, or 4 branches. Only the following partitions shown in Figure 1 satisfy the rules (leaving dissymmetry out of account):

Branches	Partitions	No. of partitions	
2	$  \begin{array}{c}  \square \\  \diagup \quad \diagdown \\  C \\  \diagdown \quad \diagup \\  \square  \end{array}  $	5 5	1
3	$  \begin{array}{c}  \square \\  \diagup \quad \diagdown \quad \diagup \quad \diagdown \\  C \\  \diagdown \quad \diagup \quad \diagdown \quad \diagup \\  \square \quad \square \quad \square  \end{array}  $	$  \begin{array}{c c c c}  1 & 2 & 2 & 3 \\  4 & 3 & 4 & 3 \\  5 & 5 & 4 & 4  \end{array}  $	4
4	$  \begin{array}{c}  \square \\  \diagup \quad \diagdown \quad \diagup \quad \diagdown \quad \diagup \quad \diagdown \\  C \\  \diagdown \quad \diagup \quad \diagdown \quad \diagup \quad \diagdown \quad \diagup \\  \square \quad \square \quad \square \quad \square \quad \square  \end{array}  $	$  \begin{array}{c c c c c c}  1 & 1 & 1 & 1 & 2 & 2 \\  1 & 2 & 1 & 2 & 2 & 2 \\  3 & 2 & 4 & 3 & 2 & 3 \\  5 & 5 & 4 & 4 & 4 & 3  \end{array}  $	6

No closed algebraic expression has been found for this enumeration. However, the recursive expansion was done manually by Henze and Blair with a few trivial errors later found by a computer check. No organic chemist will be surprised by the enormous scope of his field of study. There are, for instance, 366,319 isomeric icosanes,  $C_{20}H_{42}$ , and 5,622,109 icosanols,  $C_{20}H_{41}OH$  (Table 1).

Table 1. Counting the different arrangements of compounds of carbon and hydrogen containing no double or triple bonds and no rings. These have the general formula  $C_nH_{2n+2}$ .

Number of Carbon Atoms		Number of Possible Isomers	
		11	159
1	1	12	355
2	1	13	802
3	1	14	1858
4	2	15	4347
5	3	16	10359
6	5	17	24894
7	9	18	60523
8	18	19	148284
9	35	20	366319
10	75		

The total range of acyclic compounds containing atoms other than that of the hydrocarbons C or H is, of course, very much larger than these subsets. To generate them, an allocation of nodes to constituent radicals takes account of the kind as well as number of remaining atoms. A complete enumeration of structural isomers of a given composition, for example of alanine,  $C_3H_7NO_2$ , can thus be made. We find 216 such isomers if we apply only these simple topological principles, compared with just 5 isomers of  $C_6H_{14}$ .

### Graphs of Ring Compounds

Cyclic graphs are less tractable than trees. A linear representation is difficult because every path may return to a specific node already defined. The symmetries of cyclic graphs complicate the problem of defining a unique center on morphological criteria. These taxonomic difficulties are reflected by the existence and popularity of the American Chemical Society's Ring Index, which displays the "11524 rings known to chemistry" together with a profusion of synonyms and arbitrary numbering systems. Many more rings are discovered every day.

Molecules may also contain both acyclic and cyclic parts. However, if a strictly cyclic part is once defined, it can be regarded as a single node in a tree.

We now consider the strictly cyclic graphs, wherein at least two (sometimes more) links must be cut to separate the graph. First we produce a set of strictly trivalent cyclic graphs. Then these are related to the chemical graphs by ignoring the bivalent nodes of the latter. That is, the trivalent vertices are preserved to describe an abstract, basic graph and each linear path between vertices maps onto an edge of the basic graph. The degenerate case of zero vertices, the circle, must be included in the set

since the simple ring is the most important cyclic structure of organic chemistry. A double ring can be generated in only one way, mapping onto a two-vertex trivalent graph: the molecule naphthalene maps onto the hosohedron. Figure 2 gives some of the more familiar cyclic hydrocarbons to illustrate these correspondences.

Some organic molecules have one or more quadrivalent vertices. This contingency can be met head-on by enumerating the full set of corresponding tri- quadrivalent graphs. It is expedient to convert these, when needed, to trivalent graphs by any of a number of tricks. For example, map a quadrivalent node onto a pair of connected trivalent nodes.

We now proceed to enumerate the trivalent graphs, each with an associated canonical representation and an implied numbering of nodes and edges for mapping the molecule.

### Once Around the Network

A practical key to the solution of this problem, as to many other network problems, takes advantage of the Hamilton circuits found in most of the abstract cyclic graphs having chemical interest. A Hamilton circuit is a round trip through the graph that traverses each node just once. It therefore uses  $n$  edges, leaving out  $n/2$  edges of a trivalent graph. Figure 3 is Hamilton's own example, the dodecahedron, proposed by him as a parlor game, each node representing a city that the round-the-world traveler would wish to visit once but not more often.

A convenient representation of an HC maps the nodes and edges of the circuit as vertices and bounding edges of a regular polygon. The remaining  $n/2$  edges then form chords, each node being one of the two termini of one chord. A description of the graph then needs only some notation for the  $n/2$  chords. First, we should canonicate the orientation of the polygon, having chosen to initialize the HC arbitrarily among  $n$  nodes and 2 directions (the rotational and reflectional symmetries of the polygon). Each node is joined by some chord having a certain span. The span list can be put in cyclic order, where it is immaterial which node is selected as starting point. The effect of reflection is also easily computed. If the span list is regarded as a number, its minimum value under rotation or reflection becomes the canonical form. For example, an 8-node graph might be represented (Figure 4) by any one of the span lists 17522663, 31752266, and so on, or the reflections 75226631, and so on. Of these, one quickly finds that 17522663 is the lowest-valued, hence the canonical form.

The same procedure establishes a canonical ordering of the nodes and edges. For the latter, we take the HC sequence (the polygon) first, then each chord in order of first reference.






























The span list has  $n$  terms. Only  $n/2$  are necessary since each chord is referred to twice in the span list. For an abbreviated code, simply omit the




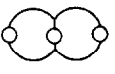
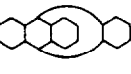
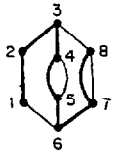
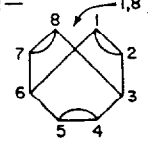
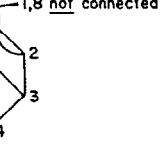
POLYGONAL REPRESENTATION	POLYHEDRAL FORM	PLANAR MESH DIAGRAM	EXAMPLE
 A			
 BB			
 AA			
 BCB			
 AAA			
 ABB			
 ACA			
 CCC	Gauche		No example
 BCCB	}		
 BDDB			
 CEDC		 Cubane	

FIGURE 2. The cyclic trivalent graphs with 8 or fewer nodes. Up to 6 nodes, these all have Hamilton circuits but may also be represented in other ways. In a few examples, the circuits are drawn with emphasis on planar map representations. Complete tables of chord lists like those shown under the circuit (polygonal) representations have been published for up to 12 nodes, virtually exhausting graphs of chemical interest.

The chemical examples are, wherever possible, hexacyclic hydrocarbons. Each vertex stands for a carbon atom.

POLYGONAL REPRESENTATION	POLYHEDRAL FORM	EXAMPLE
 AAAA		
 AABB		
 AACA		
 ABCE		
 ABOA		
 ACDB		
 ADDA		
 AEBC		
 AECA		
 BBBB		

Nonpolygonal graph with known chemical examples

CODE	MAPPING ON UNDERLYING GRAPH	POLYHEDRAL FORM	CHEMICAL EXAMPLE
(8A:1,8:ACA)			
(*(AE)EAA)			

A Hamiltonian path where a circuit is lacking

The final example has no Hamilton circuit. It can be computed either as a predicted union of two circuits (A with ACA, edge 1 with edge 8), in canonical form, or as a Hamiltonian path (\*(AE)EAA), the asterisk signifying that the polygon cannot be closed, and (AE) that two chords, A and E, both issue from the same, initial, node.

As explained in the text, each chord of the polygonal representation is coded by one character for its span the first time it is encountered in a serial circuit of nodes.

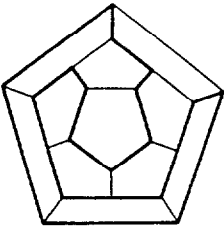


FIGURE 3. Hamilton's own Hamilton circuit. The abstract dodecahedron, represented as a planar map of 20 nodes.

second reference to any chord. Thus 17522663 becomes 1522 to encode the graph in a canonical form (Figure 4). Since we need more than 10 numbers, we use the alphabet, character by character. Thus 1522 becomes

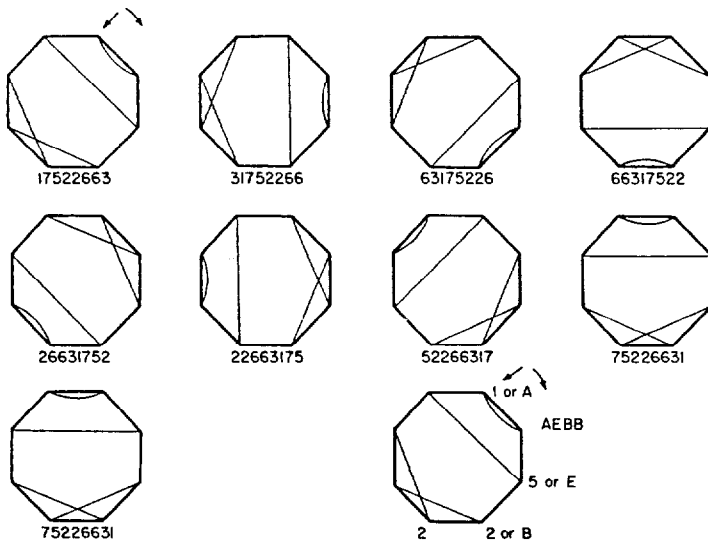


FIGURE 4. Symmetries and encoding of a cyclic trivalent graph with 8 nodes. There are 16 symmetry operations (8 rotational  $\times$  2 reflection). Shown are 8 rotations, and a reflection that could be combined with each of these. With each figure is also a span list; the canonical choice of the 16 (not all distinct) is the lowest-valued span list, 17522663, calculated with the upper rightmost node as the initial. This can then be reduced to the code AEBB.

AEBB. Furthermore, we can reconstruct the graph from the code by retracing the steps just recited. *Caution:* Unlike span lists, the abbreviated chord lists cannot be freely rotated.

Having a systematic, linear code, we are now in a position to compute all possible Hamilton circuits. Any span list is a string of numbers; therefore, the complete set of circuits can be sieved by a computer program from the series of integers. A great deal of fruitless computation can be saved by incorporating some of the canons of preferred representations into the generating algorithm. For example, no later digit can be smaller

than the leading digit; else a simple rotation of the span list, which is an obvious isomorphism, would give a smaller, preferred code number.

In this manner, exhaustive lists of Hamilton circuits for  $n \leq 12$  have been computed. They are illustrated here up to  $n = 8$  (Figure 2). Some planar, trivalent graphs lack a Hamilton circuit. The simplest has 8 nodes (Figure 2, last item) and, as it happens, it does underlie the mapping of a known compound. Obviously, these graphs will not be anticipated by a computer program that generates Hamilton circuits. However, it is not difficult to describe these figures as unions of circuits or else, for every practical case, as Hamilton paths. Furthermore, at each level of graph-building, it is possible to anticipate combinations of cut edges that will yield circuit-free graphs upon union with other partial graphs. A complete set of trivalent graphs is, therefore, computable.

The special case of the smallest, circuit-free trivalent polyhedron has been a challenge to mathematicians for some time. A polyhedron is here defined as a 3-connected trivalent planar graph, that is, one that cannot be separated with less than three cuts. Tait had conjectured that a Hamilton circuit always existed, but this was refuted by Tutte with a 46-node counterexample. Subsequently a 38-node case was built which lacks a Hamilton circuit (Figure 5). So far as is known, this is the smallest;

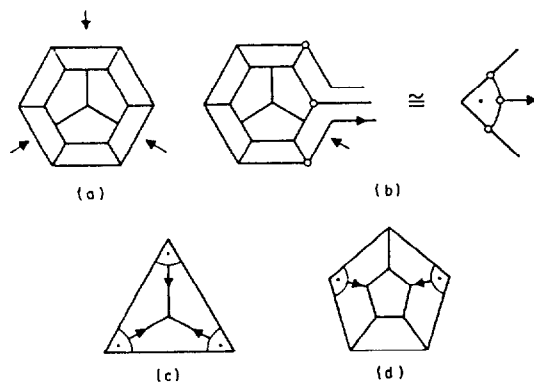


FIGURE 5. A graph with special edges and two HC-free polyhedra. (a) has 16 nodes. The marked edges are included in any HC of the graph. Hence the 3-cut (b), with 15 nodes, obligates the marked edge as part of an HC of any graph in which (b) is inserted. This leads to a contradiction, that is, no Hamilton circuit in (c) Tutte's graph, with 46 nodes and (d) with 38 nodes.

however, there is no proof of it. All the trivalent polyhedra with up to 18 and 20 nodes have been scrutinized or anticipated, and all have Hamilton circuits.

No incisive theory yet deals with these curiosities of empirical mathematics, in the same sense that we have no systematic generator for pro-

ducing the  $n$ th prime number. However, if the elegance of the theory of polyhedra is marred by such empiricism, it is no impediment to putting the chemistry of real molecules on the computer.

Nonplanar graphs are theoretically important possibilities. The corresponding molecules (Figure 6) should be difficult but not impossible to synthesize. So far, none has been reported.

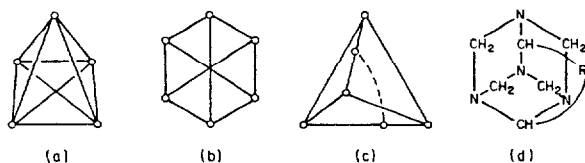


FIGURE 6. Nonplanar graphs. (a) and (b) are Kuratowski's fundamental forms, 4-valent and 3-valent, respectively. At least one of these must be included in any nonplanar graph. (c) is a projection of (b) as a tetrahedron with an additional internal chord, and (d) is a hypothetical molecular structure that maps on to (c).

## Mapping and Symmetry

Having explored the trivalent graphs, we now return to mapping chemical atoms on their nodes and bonds or linear chains on their edges. Many graphs have substantial symmetry, and the correspondingly redundant operations must be considered to decide on a canonical representation. Here, again, the HC's are helpful. If an HC is present, it can also be projected on the same graph after any symmetry operation. Therefore, the whole set of symmetry operations is included within the list of the HC's, giving both remarkable economy of computational effort to the search for the symmetries and a straightforward expression of the operators. To describe a molecular structure, we can map it on an arbitrary choice of form and then subject the result to the symmetry operators. The canonical representation satisfies some rule, say the highest-order listing, of the mapped elements. Thus, for the morphine nucleus, we would have to choose among the 4 symmetries of its underlying graph (Figure 7), and we can then encode the morphinan molecule as

(8BDDB 4\*0031301000 NC3,C3,O,C3,C).

The first two words define the basic map, "\*" standing for a fused edge, and the digits for the lengths of the paths between vertices. The last clause maps the atomic strings on to the nonempty edges.

Besides the linear paths of the cyclic structure, the mapping may also include specifications for fused edges (quadrivalent centers), heteroatom replacements of vertices, and specifications of stereosymmetry of vertices. The details are inevitably fussy, but the computer handles all the fuss once the program is worked out. After the mapping, each atom is numbered in the order of its reference.

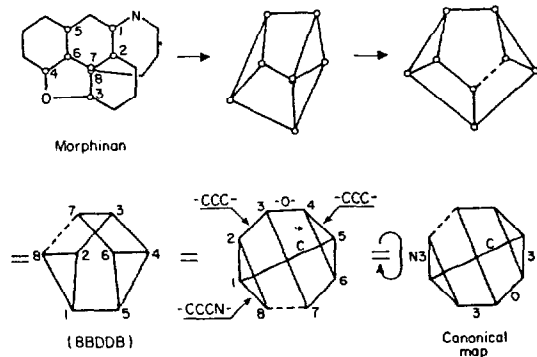


FIGURE 7. Mapping a complex ring: morphinan.

## Applications

This development was needed for a continuing effort to program the automatic computation of structural hypotheses to be matched against various sets of analytical data, especially mass spectra. The growing sophistication of instrumental methods has already begun to outdo the chemist's capacity to interpret the results. Since mass spectrometers now commercially available can generate 10,000 spectra per second, the need for computational assistance to make full use of this speed is self-evident. Such devices are also being considered for the automated exploration of the planets, which puts even heavier demands on the local intelligence available to the system.

These applications relate primarily to the possibility of anticipating hypothetical structures. The language also provides a format for expressing synthetic insights, that is, the elementary reactions by which functional groups can be altered or exchanged. We might then expect the ultimate development of computer programs that have been taught a few thousand unit processes (and their limitations) and could be challenged to anticipate a synthetic route from given precursors to a given end product. Such programs might at least assist the chemist by reminding him of a few among myriad possibilities of combining the unit processes learned from the same chemist or, better, from a diverse school. For the moment, we do not consider the empirical testing in the computer's own laboratory of a few thousand routes chosen on its own initiative.

The nomenclatural utility of a system of canonical forms is self-evident. We are very nearly at the point where linear notation may again be dispensable for human use since the computer should be able to interpret structural graphs as such. However, a mathematically complete system of classification of structures is still important, regardless of the notation in which the structures are expressed.

There are, of course, many alternative approaches to notation, reviewed by a National Academy of Sciences Committee (1964) and appearing

from time to time in the *Journal of Chemical Documentation*. As far as I know, none of them has been addressed to the exhaustive prediction of canonical forms, and most of them are too complicated to be easily adaptable to this end.

### Computer Implementation

The notation of the computer language called DENDRAL is the foundation of some current efforts at mechanized induction in organic chemistry. A program to generate all isomers of tree structures has been fully implemented in the LISP programming language and is routinely run on a time-shared PDP-6 computer at Stanford University. Most of this program was developed on the Q-32 computer of the System Development Corporation at Santa Monica, California, using remote teletype consoles located at various homes and offices at Stanford, 400 miles away.

The kernel of the program is a "topologist" embodying the principles of the first part of this paper. It is, however, restrained by some common-sense chemistry to eliminate many inappropriate constructions. For example, the chemist knows that enolic structures like  $\cdot\text{CH}=\text{CH}\cdot\text{OH}$  are unstable, rapidly reverting to a tautomeric equivalent (aldehydes),  $\cdot\text{CH}_2\cdot\text{CH}=\text{O}$ , and this information is embodied in the higher-level program. Also included is a model of the process of molecular fragmentation in the mass spectrometer, leading to a deduction of the mass spectrum expected from a hypothetical structure. The program uses the input data to guide its induction of candidate hypotheses, then tests these hypotheses deductively against the data, in an emulation of the traditional scientific method.

Much to our surprise, the program already works with real data, sometimes giving correct solutions. Not so surprising, the program greatly outdoes human chemists in problems like generating all the isomers of a given composition. Most of us founder on the isomorphisms.

Students encountering organic chemistry for the first time are often frustrated because they are challenged with graph-theoretic concepts, implicitly, without being told that this is their problem. For example, a student is expected to use his intuition to discover that there are only two isomers of  $\text{C}_2\text{H}_6\text{O}$  (in our notation,  $\text{CH}_2\cdot\cdot\text{CH}_3\text{OH}$ , ethanol, and  $\text{O}\cdot\cdot\text{CH}_3\text{CH}_3$ , dimethyl ether), but this intuition is achievable only with extensive practice. And even an experienced chemist will be hard-put to describe, irredundantly, all the isomers of slightly more complicated molecules, say  $\text{C}_6\text{H}_{14}\text{O}$ . Many problems in elementary chemistry are solved by excluding all but one of a list of possible isomers, implying that the whole list is deducible. The concept of the center of a tree and the algorithms for systematic generation of isomers should be of substantial value in teaching this subject, quite apart from the implementation of the algorithms on the computer. The same consideration should also

apply to the ways in which rings can be built and to positional isomerism of substituted rings.

#### BIBLIOGRAPHY

- E. F. Beckenbach, *Applied Combinatorial Mathematics*, Wiley, New York, 1964.
- O. Ore, *Graphs and Their Uses*, New Mathematics Library, Random House, New York, 1963.
- J. Lederberg, "Topological Mapping of Organic Molecules," *Proc. Nat. Acad. Sci. U.S.*, *53*, 134-139, 1965.
- J. Lederberg and E. A. Feigenbaum, "Mechanization of Inductive Inference in Organic Chemistry," in *Formal Representation of Human Judgment*, B. Kleinmetz, Ed., Wiley, New York, 1968.
- V. Klee, "Long Paths and Circuits on Polytopes," in *Convex Polytopes*, B. Grunbaum, Wiley, New York, 1967.
- J. Lederberg, "Hamilton Circuits of Convex Trivalent Polyhedra (Up to 18 Vertices)," *Amer. Math. Monthly*, *74*, 522-527, 1967.

#### BIOGRAPHICAL NOTE

Joshua Lederberg is Professor of Genetics at Stanford University School of Medicine, Palo Alto, California. He interrupted his medical studies at Columbia University in 1946 for what was intended to be a brief research leave with Professor E. L. Tatum at Yale University. This work on the genetics of bacteria became too fruitful to be dropped, and indeed was the basis of the Nobel Prize in Medicine that was awarded to him twelve years later.

Professor Lederberg's best-known research accomplishment has been the discovery of mechanisms of genetic exchange in bacteria, which underlie much contemporary research in molecular biology and which have helped to unify modern concepts of life on earth. He has been involved in the search for extraterrestrial life that has become part of the space-exploration program. As indicated by this essay, he is also interested in scientific methodology as a problem in itself, with a view to augmenting human intelligence by the use of machines.

He has been broadly concerned about the implications of new biological knowledge for mankind — for example, in his role since 1961 as Director of the Lt. Joseph P. Kennedy, Jr. Laboratories for Molecular Medicine (dedicated to the study of mental retardation). He has served on President Kennedy's Panel on Mental Retardation, and is currently a member of the National Advisory Council of the National Institute of Mental Health. He also writes a column on "Science and Man," which appears weekly in the *Washington Post* and is syndicated to a few other newspapers on four continents.