Interim Report to the
National Aeronautics and Space Administration
Grant NsG 81-60

NASA STAR No. N66-14075
NASA CR-68899

Instrumentation Research Laboratory
Technical Report No. 1040

# SYSTEMATICS OF ORGANIC MOLECULES, GRAPH TOPOLOGY AND HAMILTON CIRCUITS

## A General Outline of the DENDRAL System

submitted by

Joshua Lederberg
Professor of Genetics
School of Medicine
Stanford University
Palo Alto, California

January 12, 1966

Studies related to this report have been supported by research grants from the National Aeronautics and Space Administration (NsG 81-60), National Science Foundation (NSF G-6411), and National Institutes of Health (NB-04270, AI-5160, and FR-00151).

FOREWORD. This contribution is intended as an introductory survey of the topological concepts that underlie the DENDRAL system for chemical structure notation. The main purpose of the system is to provide a language in which a computer program can frame hypotheses of organic chemistry. For example, a program to generate all the isomers of a given formula has already been implemented.

This introduction is especially intended for users who wish only a general outline of DENDRAL rather than its full details of syntax. Some notation is necessarily used. This resembles the definitive DENDRAL forms, but the complete manual should be used as a definitive statement of the language.

# SYSTEMATICS OF ORGANIC MOLECULES, GRAPH TOPOLOGY AND HAMILTON CIRCUITS

Joshua Lederberg
Genetics Department
Stanford University School of Medicine
Palo Alto, California

The structural formula for an organic molecule is a paragon of a topological graph, that is, the connectivity relations of a set of atoms. True, we recognize more than one type of connection, double, triple, and non-covalent bonds, as well as single bonds. However, from an electronic standpoint the special bonds could just as well be denoted as special atoms. The structural graph does not specify the geometry, that is, the bond distances and bond angles of the molecule. In fact, this is 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 valence. The student then also learns rules which prohibit some configurations as unstable or unrealizable (and may later earn his scientific reputation by justifying or overturning one of these rules). The field of organic chemistry has, however, 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: the teaching of chemical principles to college undergraduates, and 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 the graphs having the most chemical interest. Some difficult problems, e.g., the enumeration of polyhedra, remain unsolved. However, the main obstacle may be the seeming triviality of the problems, many topologists being quite unsatisfied with systems restricted to 2- or 3-dimensional space.

This article will review some elementary features of graphs that may be used for a systematic outline of organic chemistry. The same theory has the broader significance of classifying the possible nets of relationships among the members of a set of objects. For present purposes, our graphs will be undirected, that is, any connections are reciprocal and unpolarized. Furthermore, our atoms have a maximum valence of 4. When we come to cyclic structures we shall have occasion to study an even more restricted set of graphs, those in which every node has a valence of 3.

A problem statement might be: enumerate all the distinct structural isomers of a given elementary composition, say $C_3H_7NO_2$. This is tantamount to producing 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, and 2 for C, N, O, respectively). For compactness, H can be left implicit, being later restored at every unused valence.

Our main approach throughout this article is <u>mapping</u>, a rule of correspondence between a part of the chemical structure and a part of some abstract graph. Thus, each <u>atom</u> may be mapped on to a <u>node</u>: each <u>bond</u> to an <u>edge</u> or link of the graph. For further analysis, however, it will be important to map from <u>complexes</u> of the structure to <u>elements</u> of a graph. The abstract graphs lend themselves to canonical forms, i.e., a choice among equivalent representations

according to precise rule. Since the root problem is generally not that of producing all possible combinations of atoms, but recognizing which forms are unique, this is of utmost importance. Chemistry will re-emerge after a few levels of abstraction.

These principles have been elaborated in a computer-oriented language "Dendral-64" which is described more fully elsewhere for the purpose of possible implementation in programming systems (Lederberg, 1964).

Trees are 1-connected graphs, i.e., can be separated into two parts by cutting any link. They correspond to the acyclic structures of organic chemistry. 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 very suitable. However, each tree has a unique center. In fact Jordan (1869) 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, being 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 of the graph is the number of levels of pruning needed to reach the center.

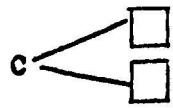
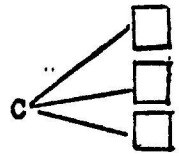
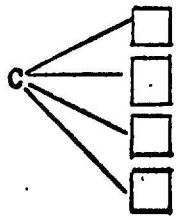
To identify the mass-center of a tree, we must consider the two or more branches that join to each non-terminal node. The center is the node whose branches have the most evenly balanced allocation of the remaining mass (node-count) of the tree. This is the same as to say that none of the pendant branches exceed half the total mass. A mass of even number allows the possibility of the center being a node pair or edge which joins equal halves.

Either of the centers (Fig. 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 enjoy the same mass. The radius-center is ill-adapted for this, but matches conventional nomenclature, which is based on finding the longest linear path, i.e., a diameter. The diameter is not necessarily unique. For example, urea has three diameters, N - $\overset{''}{C}$ - N and N - $\overset{'}{C}$ = O (twice), but just one radius-center, the C atom. The problem of generating isomers is the main justification for adopting the mass-center over the radius-center to work out canonical forms.

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 arginine becomes (C-C-N-C(N)-N C-C(N)-C(O)-O) or, in a parenthesis-free notation with some abbreviations .2.N.C.:NN 2..NC.:OO . Any linear code has an implicit number system: each atom is numbered according to when it is denoted in the string.

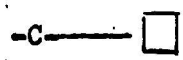
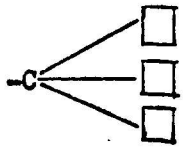
Some thirty years ago, Henze and Blair (1931) showed how Jordan's principle could be used for the enumeration of isomers of saturated hydrocarbons and some simple derivatives of them. Here, the nodes are all the same (carbon atoms) and the enumeration can proceed by recursion from smaller to larger complexes. For example, for the isomers of undecane, $C_{11}H_{24}$, one atom is designated as center, leaving 10 to be allocated among 2, 3 or 4 branches. Only the following partitions satisfy the rules (leaving dissymmetry out of account):

BRANCHES

| | | |
|---|---|---|
| 2 |  | 5,5 |
| 3 |  | 1,4,5<br>2,3,5<br>2,4,4<br>3,3,4 |
| 4 |  | 1,1,3,5<br>1,2,2,5<br>1,1,4,4<br>1,2,3,4<br>2,2,2,4<br>2,2,3,3 |

To complete the solution, one must have calculated the number of alkyl radicals $-C_5$, $-C_4$, etc. To illustrate with $C_5$:

The radical must have an apical atom, leaving the rest to be partitioned in all distinct ways among 1, 2 or 3 pendant branches, the radicals of the next level. Thus we have:

| | |
|---|---|
|  | 4 |
|  | 1,3 or<br>2,2 |
|  | 1,1,2 |

The count of $-C_n$ radicals is thus derived from the table for $-C_i$, taking i from 1 to n - 1, and the process may be iterated as far as needed, i.e., until partitions into units, $C_1$ , prevail. No deep mathematical insight is needed to

verify that the first steps of the alkyl series $C_1$, $C_2$, $C_3$, $C_4$ have 1,1,2,4 forms respectively.

No closed algebraic expression has been found for this enumeration. However, the recursive expansion was done by hand (Henze and Blair, 1931) with a few trivial errors found by a computer check; no organic chemist will be surprised by the enormous scope of his field. (Table 1).

The total range of acyclic compounds is of course very much larger than these subsets. At each step, instead of partitioning a mere number of nodes, an allocation to constituent radicals takes account of the kind as well as number of unused atoms. However, the specification of a hierarchy of ordering, which may be done almost arbitrarily to suit computational convenience, permits the same principles to be applied to a complete enumeration of structural isomers of a given composition, for example of alanine, $C_3H_7NO_2$. (Table 2.)

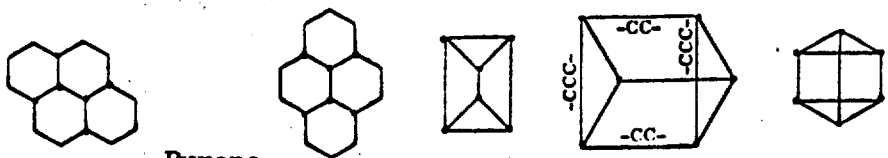
## Cyclic Structures

Cyclic graphs are much less tractable, since every path will return back to the complex, and a center is less easily defined. Sufficient reminder of the taxonomic difficulties posed by rings is the popularity of the Ring Index (1964) wherein the "11524 rings known to chemistry" are laid out, together with a profusion of synomyous and alternative numbering systems to map them as nodes. For example, naphthoyl pyridine would ultimately form a tree, $R_1 - \overset{\overset{O}{\|}}{C} - R_2$, $R_1$ and $R_2$.

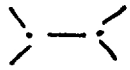
We now consider the domain of strictly cyclic structures. These are 2-connected graphs, since at least 2 (sometimes more) links must be cut in order to separate the graph.

For further analysis, we distinguish the trivalent vertices of the structure atoms that join 3 paths, or branch points. We can then construct the full set of

abstract, trivalent graphs. Define a path as a link or an unbranched chain of links and atoms. The paths between vertices of the structure can then be mapped onto the edges of an abstract graph which is regularly trivalent or trihedral. To illustrate, observe how pyrene is mapped onto an abstract graph of 6 vertices, indeed, the abstract prism.




Pyrene (a) (b) (c) (d)

Some vertices are 4-valent, in so-called spiro forms, but these graphs can be mappedoonto 3-valent graphs by expanding each 4-valent node into a pair of 3-valent nodes. That is,  becomes  . There is an obvious relationship between the number of vertices and the number of rings conventionally ascribed to a structure. We start with, say, benzene, 0 vertices, and 1 ring. Then naphthalene, 2 vertices and 2 rings. Each additional ring entails 2 more vertices. Hence, for r rings and n vertices

$$r = 1 + n/2 ,$$

and for these trivalent graphs, n must be an even integer. Recalling that a 4-valent vertex maps into 2 3-valent nodes, we can write

$$r = 1 + n/2 + q$$

for q 4-valent vertices. This calculation agrees with the Ring Index rule which counts rings as the number of cuts needed to convert a ring structure into a tree.

As each edge joins 2 nodes, a trivalent graph of order n will have 3n/2 edges.

Enumerating the trivalent graphs. A trivalent graph may have several

representations, and some effort may be required to relate them to one another, and to decide which form is to be regarded as a canonical reference for mapping purposes. Thus, the graphs of Figure 2 are all topologically equivalent or isomorphic. This is to say, they all represent the same connections of node to (three) nodes. A meaningful enumeration must unify these isomorphisms. Furthermore, it should relate to a convenient code by which to refer to each graph, better still, to embody a reconstruction. Finally, it should generate an obvious numbering of the nodes and edges.

Hamilton circuits. 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 graphs having chemical interest. A Hamilton circuit (HC) is a round trip through the graph that traverses each node just once. It therefore uses n edges, leaving out n/2 edges. 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 traveller would not wish to revisit. The utility of HC representations will become evident.

Finding all HC's of a graph may be a challenging game, but it is reduced to a merely tedious algorithm on the computer. Start from an arbitrary node. Trace a path as through a maze, each node presenting a binary choice of different edges. If the chosen path reverts to a node already visited, backtrack one step. A successful path has n correct choices. Thus, at most $2^n$ search steps will exhaust all possible paths; in practice, closer to $1/n$ times this number will be needed to identify all the HC's. Even for n up to 20 this is a modest task. And if the work has been done once, finding any HC, at perhaps n-fold less effort, will enable a given graph to be related to the

previously established set.

A typical problem in graph manipulation is to establish whether two complicated graphs are isomorphic. In the long run, this might require testing all possible permutations of nodes, with a scope of Factorial (n). At n = 20, this number is an utterly uncomputable $2.4 \times 10^{18}$ steps. On the other hand, if two graphs are isomorphic, they must have the same HC's, found with at most $2^{20} = 10^6$ steps.[1]

A convenient representation of a 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 invariant under rotation; i.e., 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/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, etc., or the reflections 75226631, etc. Of these, one quickly finds that 17522663 is the lowest-valued, hence the canonical form. Similarly, when other HC's are found for the same graph, they can be compared, and the lowest-valued of them chosen as the reference graph.

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 second reference. 17522663 becomes 1522. Indeed, one less character still suffices, the last chord being completely determined by the ones previously built. The chord list (152), or an alphabetic equivalent (8AEB) whose leading numeral merely reminds us of the order of the graph, then encodes the graph in a canonical form (Figure 4). Furthermore, the graph can be reconstructed from the code by retracing the steps just recited. Caution: Unlike span lists, the abbreviated chord lists cannot be freely rotated.

Chord lists can be computed by an obvious combinatorial procedure, with the help of a few tricks to save some fruitless effort. Most arbitrary lists become internally inconsistent after a limited number of initial characters; the number of combinations that must be tested is therefore considerably less than may appear. Additional restrictions can also be put on prospectively. In this way, exhaustive lists of trivalent graphs have been computed -- Table 3 (taken from the DENDRAL report) shows their scope. To unify isomorphisms, the complete list of HC's is computed for each chord list.

Apart from the rotation of the polygon, two or more incongruent HC's may be present in a graph. No general principle is known, except that graphs with high symmetry tend to have the fewest incongruent HC's. Tutte (1946) proved that any edge of a polyhedron must be involved in an even number (not excluding 0) of HC's, and that if a polyhedron admits one HC, it must admit at least three.

Classification of trivalent graphs. Two important, independent criteria

of abstract graphs are (1) planarity, and (2) level of connectedness.

A planar graph is one that can be represented on the plane without edges crossing over one another. The graph need not be drawn as an HC-polygon, which rarely lacks crossing chords: Figure 3 is certainly planar. Kuratowski has shown that any trivalent non-planar graph must contain 6CC (Figure 5 b). Fortunately, this condition is easily recognized in the building of span lists. As the surface of a polyhedron can be mapped onto the plane, planarity is a necessary condition for an abstract polyhedron.

In practice, nonplanar graphs are so far unknown in organic chemistry (barring coordination complexes); however, they might in principle be realized, e.g., by the hypothetical Figure 5d.

Connectedness is the least number of cuts that will anywhere separate the graph. The 3-connected planar graphs are the abstract convex polyhedra. Intuitively, it is obvious that a region bounded only by 2 edges would be unable to enclose a volume. Steinitz (see Lyusternik, 1963) showed that every 3-connected planar trivalent graph could be realized as a polyhedron. These graphs have, naturally, attracted some interest as a meeting point of topology and classic Greek geometry. Nevertheless, a complete enumeration is still unknown. In 1901, Brückner published figures of the trivalent polyhedra for $n \leq 16$; in an abstract and unpublished manuscript (1928) he also showed 1250 for $n = 18$. This work, done by hand over several decades, was repeated on the computer by Grace (1965) who found some errors in Brückner's

listings, and found 1249. However, even this census admits some possibility of being incomplete, though this is remote. Grace generated the polyhedra by induction as all possible slicings of the faces of smaller polyhedra. This produces many isomorphisms which must be unified; for this, Grace used a criterion, "equisurroundedness", which is already known to be too weak, albeit for much larger graphs. Therefore, it cannot be rigorously shown that the list of 1249 has not excluded additional forms, equisurrounded, but not isomorphic with the stated set. The analysis of HC's could afford an independent avenue of corroboration at relatively low cost.

The polyhedra play an important role in the classification of cyclic graphs but have no remarkable chemical significance except that they represent the most tightly caged polycyclic structures.[12] Note that many unfamiliar isomorphisms are generated by portraying a polyhedron as a planar mesh, i.e., as projected within an arbitrarily chosen face, called the base. The projection can be visualized as the view of the polyhedron from a point just outside the place of the face chosen as base (Figure 2).

HC-free graphs. These are promptly encountered in the 2-connected series, starting with $n_8$ (8(AC:8,1:A) Figure 6). An analysis of the conditions for no-HC illuminates some of the combinatorial processes involved in building graphs. Since all the graphs for $n \geq 6$ have HC's, an HC-free graph is generated by a particular mode of union of HC's of lower order. The simplest mode is bilineal, one edge is cut on each of two smaller graphs and reunited. If either of the edges involved is barred from any HC of its graph, the bilineal union will be HC-free. This follows, since the union introduced nodes which must be traversed by a path known to be forbidden.

In general, an HC-free graph can be canonicated by dissecting it into the largest cirbuits it contains. The dissertions are first completed across the bilineal (2-connecting) unions. If any resulting subgraphs are still HC-free, we must consider HC-free polyhedra as a mathematical, if not a pragmatic chemical possibility.

HC-free polyhedra. Tait believed that all convex trihedral polyhedra contained HC's and his conjecture was indeed unchallenged for over 60 years. However, Tutte (1946) refuted the conjecture with an example ingeniously proven to be HC-free, though with 46 vertices it would defy exhaustive search.[3] Chemical graphs of this order (24 rings) are out of range of systematic prediction, but the argument gives further insight into the combinatoric of abstract graphs.

We deal here with the process of trilineal union. This can be done in all possible ways by extracting one node from any source polyhedron, leaving 3 cut edges. This 3-cut graph can then replace one node of another graph. However, to influence the possibility of forming an HC, the edges must be subject to some restrictions distinguishing the 3-cut complex from a single node. The node poses no restrictions. That is, its 3 edges are available in any pairwise combination, thus any one of 3 ways. If the corresponding edges of the source graph have the same property, i.e., none of the 3 edges is either compulsory or forbidden, then the 3-cut graph will not influence the occurrence of an HC. By induction, the lower order polyhedra that already contain some 3-connected regions can be passed over in looking for special graphs. A systematic survey of the few 4-connected, - i.e., 4-connected except for the isolated nodes which are, of course, 3-connected, - graphs (Table 4) shows the polyhedron (16CGDIGDF), the smallest with a special edge, namely that the

ones marked are obligatory in any HC of the polyhedron (Figure 7). Tutte then replaced 3 nodes of a tetrahedron with a 3-cut graph from (16CGDIGDF) leading to the contradiction that all three edges from one node must be included in any HC; hence there can be no HC in this graph of $46 = 4 + 3(14)$ nodes. The cut graph can also be planted at two mutually-exclusive edges of the pentagonal prism to give an HC-free polyhedron of $38 = 10 + 2(14)$ edges.[4] This is clearly the smallest HC-free polyhedron with two 3-connected regions.

A smaller HC-free polyhedron may yet be found by analogous studies of 4-lineal[5] and 5-lineal unions, and if so, is just within the bounds of reasonable computational effort.

If Grace's list of polyhedra is correct, every one through $n_{18}$ has an HC. This conclusion is corroborated by a detailed consideration of the properties of the graphs $n_{16}$ of table 3. By the inductive argument, forms with any triangular face -- indeed, any 3-connected region -- could be passed over, greatly reducing the computational effort. Of course, from the smallest HC-free polyhedron, larger ones can be generated by replacing a node with a triangle or larger 3-connected region.

The HC-free polyhedra can be classified by the same principles used for bilineal unions, as complexes of the largest circuits united over the least levels of connectedness.

While distant from chemical graphs of any reasonable size, these studies do furnish a clearer indication of the sufficiency of HC representations, and of the sources of exceptions.

Recapitulation: the scope of anticipation and recognition. There is no perceptible limit except the computation of HC's and of alternative dissections to restrict the encoding of abstract graphs either as HC's or as canonicated unions of HC's. These assignments also facilitate the recognition of isomorphisms between given graphs.

The anticipation of all possibilities poses a greater burden. However, all the graphs up to $n_{12}$ (7 rings) have been tabulated together with their isomorphisms and symmetries. The series expands so rapidly that further extension would tax the output-printer, and before long the computer itself.

Mapping and symmetry. Having explored the trihedral 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 corresponding by 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.[6] Therefore, the whole set of symmetry operations is included within the list of the HC's, giving remarkable economy of computational effort to the search for the symmetries, as well as a straightforward expression of the operators. To describe a molecular structure, it can be mapped on an arbitrary choice of form, and the result then subjected 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 8).

Since this choice is readily computable, the human user may be relieved of the burden to make these tedious calculations.

Besides the linear paths of the cyclic structure, the mapping may also include specifications for fused edges (4-hedral centers), heteroatom replacements of vertices, and specifications of stereoasymmetry of vertices. The details are inevitably fussy and are given elsewhere. After the mapping, each atom is numbered in the order of its reference.

Merging cycles and trees. Each cyclic structure is now fully defined, with rules for a canonical code and numbering of every atom. The structure can then be handled as a node in a tree, the numbering system allowing precise reference for the point(s) of connection.

## 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 chemists capacity to interpret the results. Since mass spectrometers are now commercially

available that can generate 10,000 spectra per second, the need for computational assistance to make full use of such devices is self-evident. (Biemann & McMurray 1965; Lederberg 1964b) 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, i.e., the elementary reactions by which functional groups can be altered or exchanged. We might then expect the ultimate development of computer programs which have been taught a few thousand unit processes, and their limitations, and could be challenged to anticipate a synthetic route from given precursors or to a given end product. Such programs might at least assist the chemist by reminding 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 leave out of consideration the empirical testing in its own laboratory of a few thousand routes chosen on the computer's own initiative.

The nomenclatural applications of any system of canonical forms are also self-evident. We are very nearly at the point where linear notation may again be dispensable, 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.

The simple graph-theoretical ideas of DENDRAL could be implemented with a number of possible notations. The one adopted for DENDRAL - 64 aims to emulate

traditional notation for all linear chains, only the most obvious abbreviations, like "3." for "C.C.C.", and a "repeat" symbol, arbitrarily "/", being laid on. The user must of course understand the principles and notation for the abstract cyclic graphs. However, it would be quite reasonable to produce an abridged version of the Ring Index which would list the carbocyclic equivalents of expected forms, and allow the most unskilled assistant to transcribe structural data in a form readily matched to DENDRAL.

Some examples of structural codes the isomers of alanine, Table 2 are appended as a challenge to puzzle-minded readers. Hopefully the tedious manual of detailed specifications (Lederberg 1964a) is not required reading for pragmatic understanding of the system.

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 cananical forms and most of them are too complicated to be easily adaptable to this end.

Syntax and induction. One of the motives for this study was to uncover the kinds of problems that would be encountered in computer-emulation of the process of scientific induction from experimental data. A necessary step is a means of generating a set of relevant hypotheses. I have been impressed with both the difficulty and the utility of establishing a precise syntactical framework for the range of hypotheses even in a field as well structured as organic chemistry.

Some years ago, Woodger (1937) attempted to axiomatize developmental and genetic biology. His efforts were perhaps too remote from the experimental

data now available. However, he may have pointed the way to a more feasible enterprise, to establish a precise syntax for hypothetical statements in biology. This is a more modest aim, since it does not purport to deduce which statements are correct. However, there is every good reason why computers should compete very successfully in the exercises of model-building that preoccupy many biologists today, and with advantage to the rigor with which they are put together.

SYSTEMATICS OF ORGANIC MOLECULES, GRAPH TOPOLOGY AND HAMILTON CIRCUITS

Footnote p. 9.

[1]While this paper was being revised, another algorithm requiring only about $10\ n^2$ steps was discovered and programmed for routine use. It depends on (1) growing a subgraph, adding one node at a time, (2) defining the list of possible circuits at each level by recursion from the list of previous level, and (3) looking ahead some steps to choose nodes which close facets of the graph so as to minimize the size of the list that must be maintained.

Footnote to p. 12.

[2]The speculative "polyhedrqnes" have been discussed by Schultz, H.P.: Topological Organic Chemistry. Polyhedranes and Prismanes. J. Org. Chem. 30, 1361 (1965).

Footnote to p. 13.

[3]This is no longer true. With a new algorithm[1], Tutte's graph was exhausted in 29 seconds of 7090 time. The same algorithm is also very apt for finding the largest circuits and for forbidden edges.

Footnote to p. 14.

[4]This had already been found by other workers as disclosed in private communications: D. Barnett, University of Washington and J. Bosak, Bratislava.

Footnote to p. 14.

[5]Tutte (1960) quotes an example with 224 nodes! If any HC-free polyhedron has fewer than 38 nodes it probably has one 3-connected region. My own investigations leave no encouragement for such an example at less than $n_{36}$.

Footnote to p. 15.

[6]I note the following conjecture, that the symmetries of any abstract convex trihedral polyhedron can be realized in a geometrical polyhedron in 3-space with reflection, i.e. can be assigned to a point group. However, this conjecture is not a premise of the method indicated for finding the symmetries. The conjecture is plainly inapplicable to 2-connected or to non-planar graphs. I would be grateful for any refutation, or a formal proof, new or otherwise.

## REFERENCES


1. J. Lederberg, DENDRAL-64, A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures (NASA CR 57029). (1964).

2. C. Jordan, Sur les assemblages de lignes. Journal für die reine und angewandte Mathematik 70, 185 (1869).

3. H. R. Henze and C. M. Blair. The number of isomeric hydrocarbons of the methane series. J. Am. Chem. Soc. 53, 3077 (1931).

4. A. M. Patterson, L. T. Capell, D. F. Walker. The Ring Index, 2nd Edition, American Chemical Society, 1960. Supplement I, 1963 . Supplement II, 1964.

5. W. T. Tutte. On Hamiltonian circuits. J. London Math. Soc. 21, 98 (1946).

6. L. A. Lyusternik. Convex Figures and Polyhedra. Dover, New York, 1963.

7. M. Brückner. Vielecke und Vielfläche. Teubner, Leipzig, 1900.

8. D. W. Grace. Computer Search for Non-Isomorphic Convex Polyhedra. Stanford Computation Center Technical Report No. CS15 (1965).

9. P. G. Tait. Listings Topologie. Phil. Mag. (5), 17, 30-46 (1884); Scientific Papers, Vol. II, 85-98.

10. K. Biemann and W. McMurray. Computer-aided interpretation of high resolution mass spectra. Tetrahedron Letters No. 11, pp. 647-653 (1965).

11. J. Lederberg. Computation of Molecular Formulas for Mass Spectrometry. San Francisco, Holden-Day Inc., 1964.

12. National Academy of Sciences - National Research Council: Survey of Chemical Notations Systems. Publication 1150. 1964.

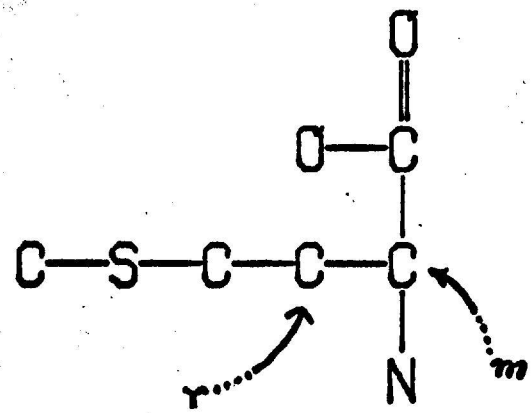
13. J. H. Woodger. The Axiomatic Method in Biology. Cambridge, The University Press. 1937.

14. D. Perry. The number of structural isomers of certain homologs of methane and methanol. J. Am. Chem. Soc. 54, 2918 (1932).

15. W. T. Tutte. A non-Hamiltonian planar graph, Acta Math. Acad. Sci. 11, 371 (1960).

C...NY2.S.C

A.




C...NY2./C

B.

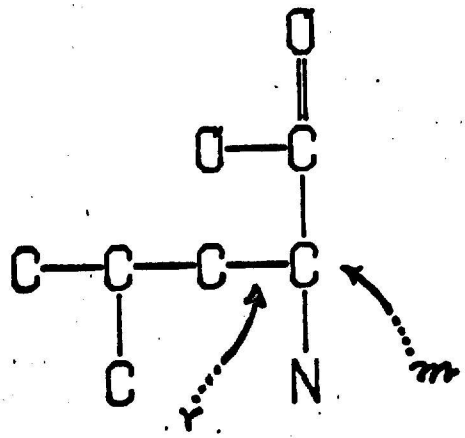



Fig. 1. Centers of trees: r (radius-center), and m (mass-center). Two examples, A., methionine, and B., leucine. The diagrams were plotted by a computer program from punch cards coded for each structure as indicated.

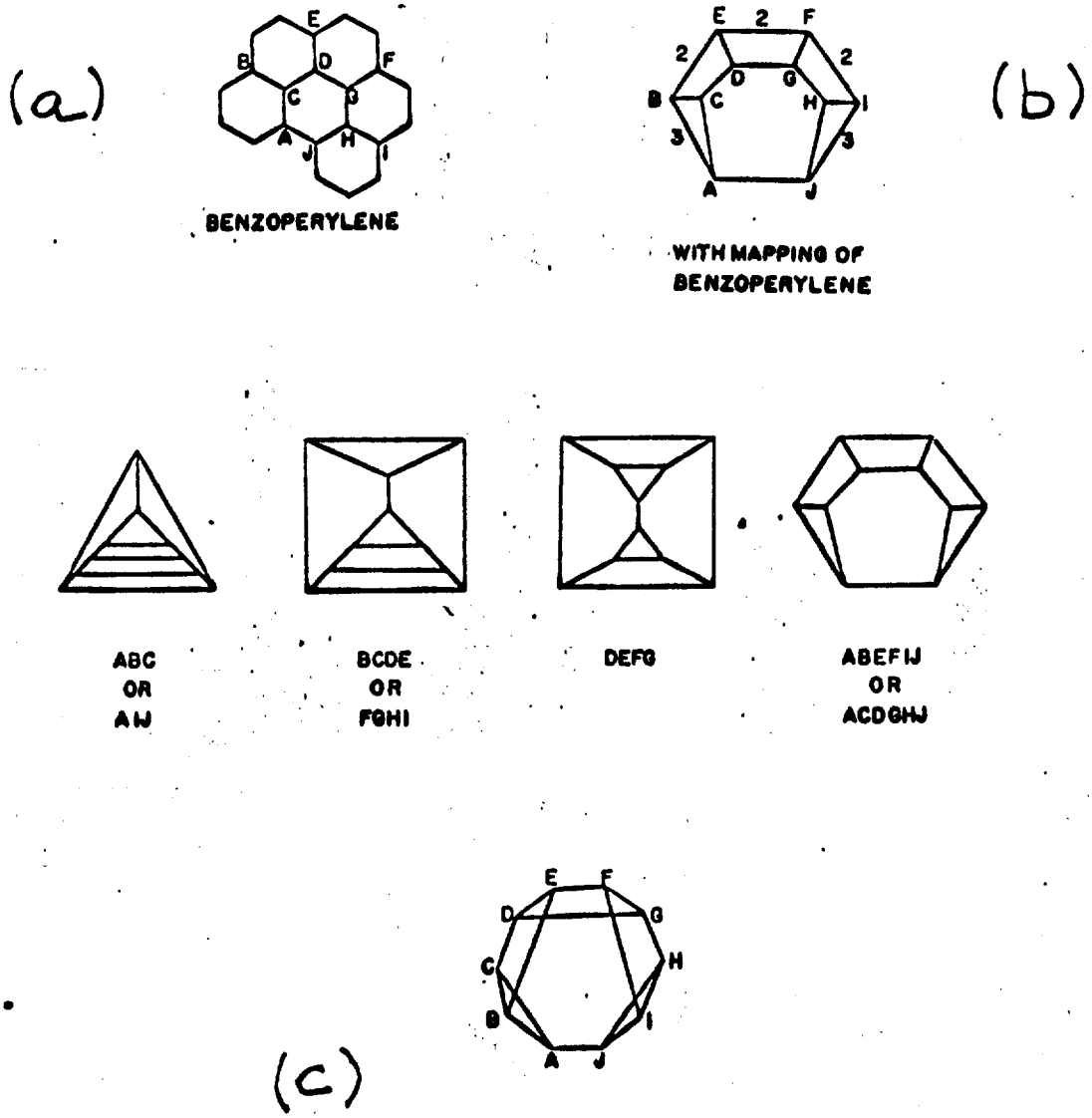



Fig. 2. (a) Benzoperylene and its mapping on a polyhedron (b) which has four isomorphic planar meshes, i.e. four kinds of faces, as labelled. (c) is the equivalent Hamilton circuit. Do not confuse the lettered labels of the nodes with abbreviated code for this graph whith is 10BCC. The reader may enjoy satisfying himself that these graphs are indeed isomorphic (equi-connected).

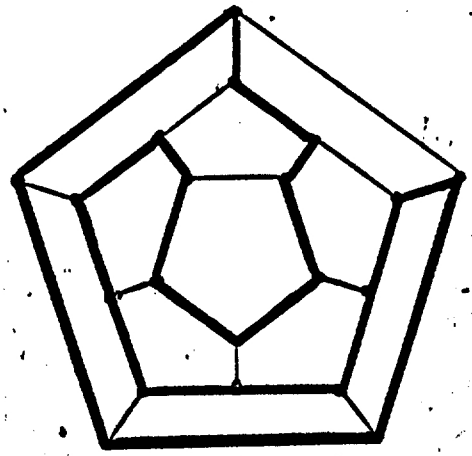
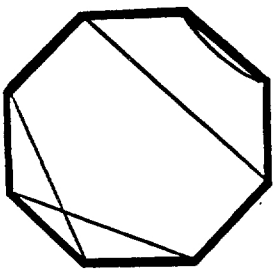
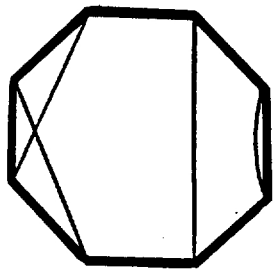


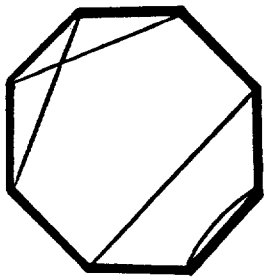
Fig. 3. Hamilton's Hamilton circuit. The abstract dodecahedron, represented as a planar map of 20 nodes.



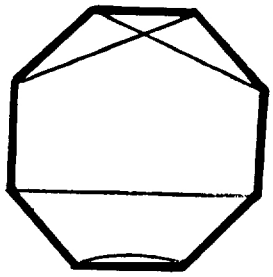
17522663



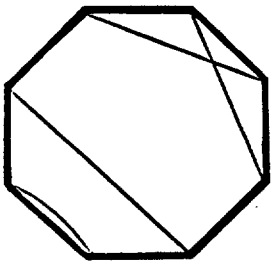
31752266



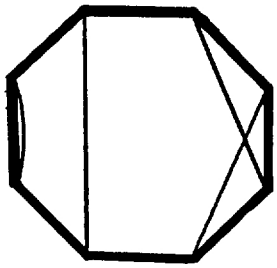
63175226



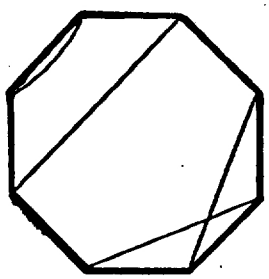
66317522



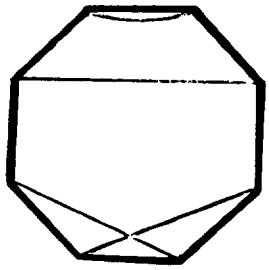
26631752



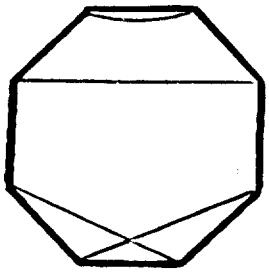
22663175



52266317



75226631



75226631

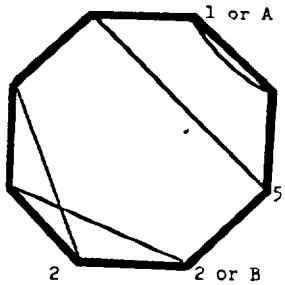
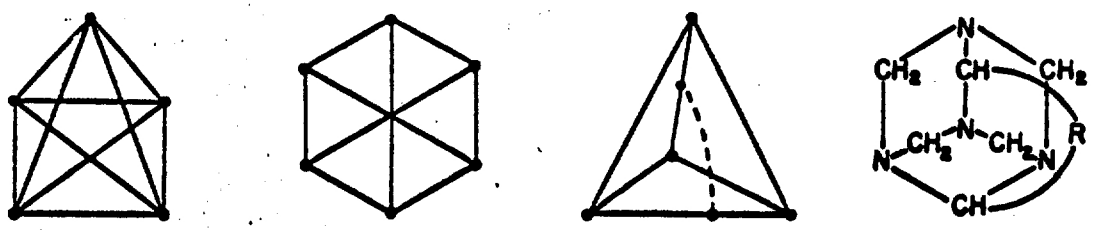



Figure 4
Caption follows

Fig. 4 Symmetries and encoding of a cyclic trivalent graph with 8 nodes. There are 16 symmetry operation (8 rotational X 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, or even more economically AEB, as outlined in the text.




(a) (b) (c) (d)

Fig. 5. Non planar 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).

Figure 6
Caption follows

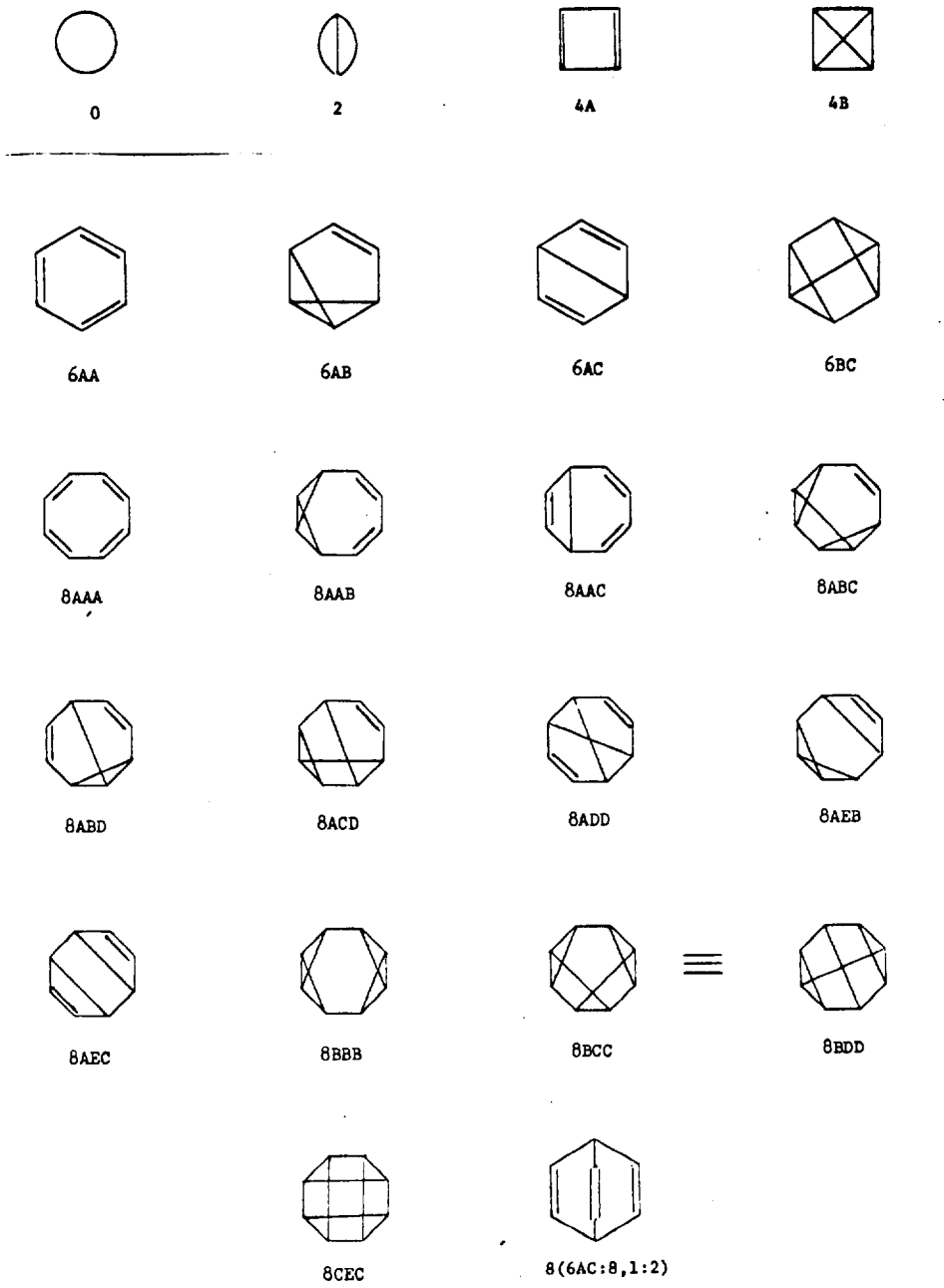



Fig. 6 The cyclic, trivalent planar graphs with 8 or fewer nodes. Where possible, these are represented as Hamilton circuits, the nodes of the graph being projected as vertices of a polygon which constitutes the circuit, the remaining edges shown as chords. Each of these figures can also be drawn as a planar map. The codes are abbreviated forms from which the graph can be reconstructed. Note that 8BCC and 8BDD are isomorphic despite the incongruence of the Hamilton circuits. The abstract polyhedra of this list include two degenerate forms (-, circle; 2, hosohedron) and 4B, tetrahedron; 6BC, prism; 8 CEC, cube; 8BCC ≡ 8BDD, pentagonal wedge. One of these graphs, 8(6C:8,1:2) has no Hamilton circuit, and is classified as a union which splices the 8'th edge of graph 6AC with the 1'st edge of graph 2. Complete lists of the graphs through 12 nodes are presented in Lederberg (1965).

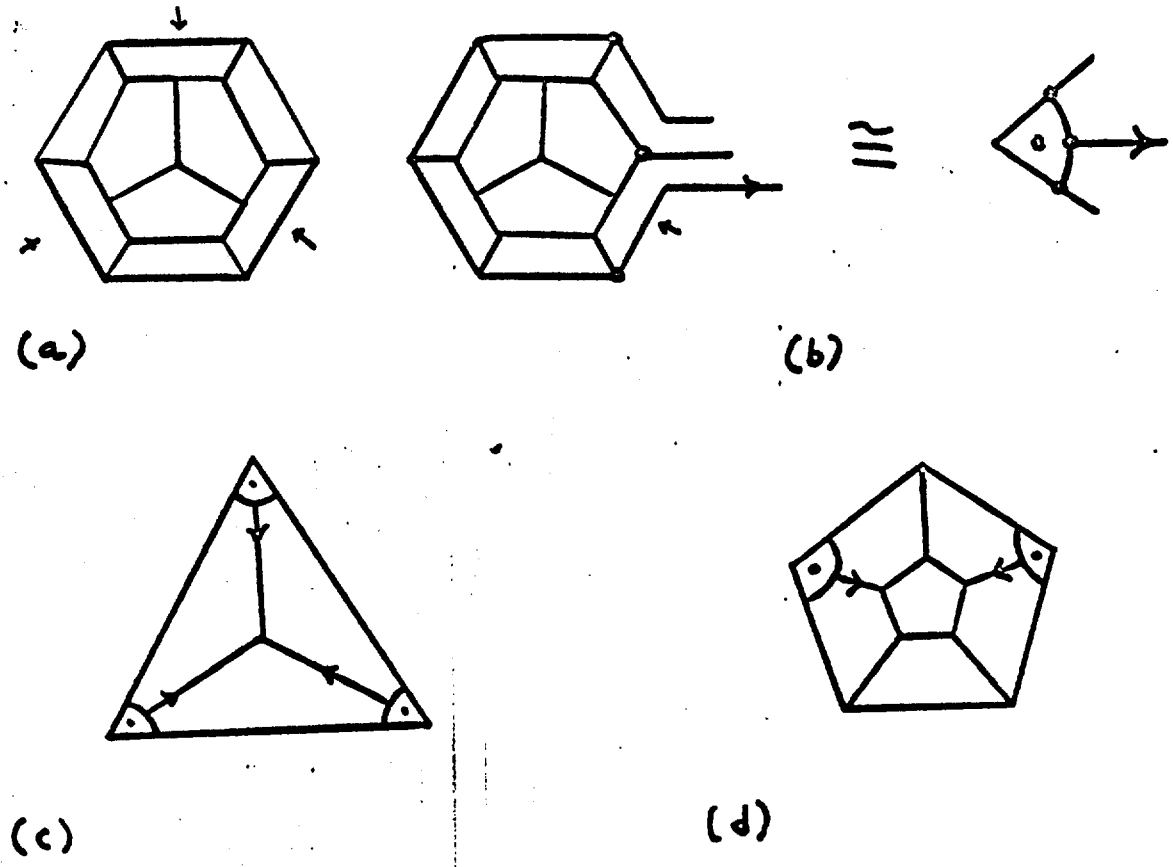
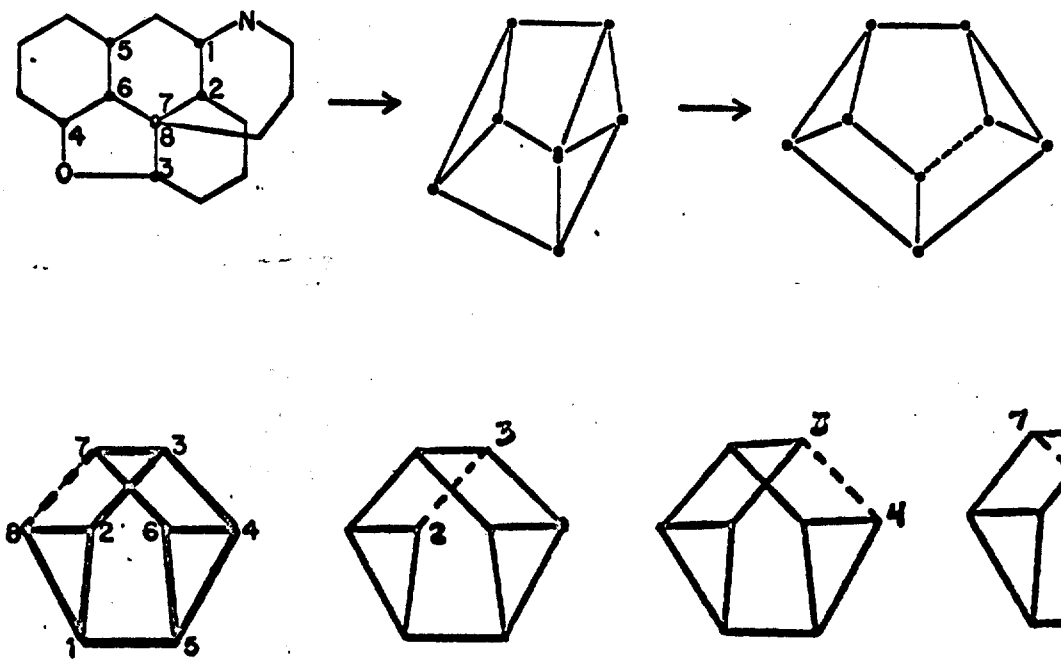
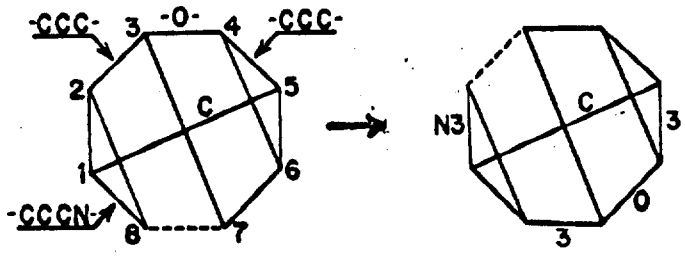



Fig. 7. A graph with special edges and two HC-free polyhedra. (a) has 16 nodes. The marked edges are included in any HC of the graph. Hence the 3-cut (b), with 15 nodes, obligates the marked edge as part of an HC of any graph in which (b) is inserted. This leads to a contradiction, i.e., no Hamilton circuit in (c) Tutte's graph, with 46 nodes and (d) with 38 nodes.




SYMMETRIES




CANONICAL MAP

HAMILTON CIRCUIT REPRESENTATIONS

Fig. 8. Morphine nucleus: symmetry and choice for coding. The dashed edge 7---8 stands for the spiro- (quadrivalent) center in the morphine ring; however, 4 permutations are possible under the symmetry operations. In the canonical form, after account is taken of the mapping of the chemical graph onto the abstract graph, this edge is labelled 2---3. The canonical map would be coded as (8BDD-N.3,$, , ,3,03,,C) each comma marking the next edge of the map. This code is sufficient input for the computer program to reconstruct the molecular structure

ENUMERATION OF THE ALKANES
STEREISOMERISM DISREGARDED

| | | | |
|---|---|---|---|
| 1 | 1 | 26 | 93839412 |
| 2 | 1 | 27 | 240215803 |
| 3 | 1 | 28 | 617105614 |
| 4 | 2 | 29 | 1590507121 * |
| 5 | 3 | 30 | 4111846763 |
| 6 | 5 | 31 | 10660307791 |
| 7 | 9 | 32 | 27711253769 * |
| 8 | 18 | 33 | 72214088660 |
| 9 | 35 | 34 | 188626236139 |
| 10 | 75 | 35 | 493782952902 |
| 11 | 159 | 36 | 1295297588128 |
| 12 | 355 | 37 | 3404490780161 |
| 13 | 802 | 38 | 8964747474595 |
| 14 | 1858 | 39 | 23647478933969 |
| 15 | 4347 | 40 | 62481801147341 * |
| 16 | 10359 | 41 | 165351455535782 |
| 17 | 24894 | 42 | 438242894769226 |
| 18 | 60523 | 43 | 1163169707886427 |
| 19 | 148284 * | 44 | 3091461011836856 |
| 20 | 366319 | 45 | 8227162372221203 |
| 21 | 910726 | 46 | 21921834086683260 |
| 22 | 2278658 | 47 | 58481806621986230 |
| 23 | 5731580 | 48 | 156192366474587200 |
| 24 | 14490245 | 49 | 417612400765371900 |
| 25 | 36797588 | 50 | 1117743651746931000 |

Table 1. Enumeration of isomeric alkanes (disregarding stereoisomerism), from methane to pentacontane. The values marked * disagree in some digits with the values calculated manually by Henze and Blair (1931) and Perry (1932). While this is an amusing exercise for the computer, the discrepancies, needless to say, will have no pragmatic chemical significance. In any case, a proportion of the structures will be unrealizable owing to steric hindrance.

| | | | |
|---|---|---|---|
| .C.C.C O.N=O | .C=N.C C..OO | .O.C.C O.N=C | .C.=CO N.O.C |
| .C..CC O.N=O | .N.C=C C.O.O | .O.C.C C.=NO | .C.=CO O.C.N |
| .C.C=C N.O.O | .N.C=C O.C.O | .O.C.C C=.NO | .C.=CO O.N.C |
| .C.C=C O.N.O | .N.C=C O.O.C | .C..CO C.N=O | .C.=CO C..NO |
| .C.C=C O.O.N | .N.C=C C..OO | .C..CO C=N.O | .C.=CO N..CO |
| .C.C=C N..OO | .N=C.C C.O.O | .C..CO N.C=O | .C=.CO C.N.O |
| .C=C.C N.O.O | .N=C.C O.C.O | .C..CO N=C.O | .C=.CO C.O.N |
| .C=C.C O.N.O | .N=C.C O.O.C | .C..CO O.C=N | .C=.CO N.C.O |
| .C=C.C O.O.N | .N=C.C C..OO | .C..CO O.N=C | .C=.CO N.O.C |
| .C=C.C N..OO | .C.=CN C.O.O | .C..CO C.=NO | .C=.CO O.C.N |
| .C.=CC N.O.O | .C.=CN O.C.O | .C..CO C=.NO | .C=.CO O.N.C |
| .C.=CC O.N.O | .C.=CN O.O.C | .C.C=O C.N.O | .C=.CO C..NO |
| .C.=CC O.O.N | .C.=CN C..OO | .C.C=O C.O.N | .C=.CO N..CO |
| .C.=CC N..OO | .C=.CN C.O.O | .C.C=O N.C.O | =C.C.C N.O.O |
| .C.C.N O.C=O | .C=.CN O.C.O | .C.C=O N.O.C | =C..CC N.O.O |
| .C.C.N C.=OO | .C=.CN O.O.C | .C.C=O O.C.N | =C.C.N C.O.O |
| .C.N.C O.C=O | .C=.CN C..OO | .C.C=O O.N.C | =C.C.N C..OO |
| .C.N.C C.=OO | .C.C.O C.N=O | .C.C=O C..NO | =C.N.C C.O.O |
| .N.C.C O.C=O | .C.C.O C=N.O | .C.C=O N..CO | =C.N.C C..OO |
| .N.C.C C.=OO | .C.C.O N.C=O | .C=C.O C.N.O | =N.C.C C.O.O |
| .C..CN O.C=O | .C.C.O N=C.O | .C=C.O C.O.N | =N.C.C C..OO |
| .C..CN C.=OO | .C.C.O O.C=N | .C=C.O N.C.O | =C..CN C.O.O |
| .N..CC O.C=O | .C.C.O O.N=C | .C=C.O N.O.C | =C..CN C..OO |
| .N..CC C.=OO | .C.C.O C.=NO | .C=C.O O.C.N | =C.C.O C.N.O |
| .C.C=N C.O.O | .C.C.O C=.NO | .C=C.O O.N.C | =C.C.O C.O.N |
| .C.C=N O.C.O | .C.O.C C.N=O | .C=C.O C..NO | =C.C.O N.C.O |
| .C.C=N O.O.C | .C.O.C C=N.O | .C=C.O N..CO | =C.C.O N.O.C |
| .C.C=N C..OO | .C.O.C N.C=O | .O.C=C C.N.O | =C.C.O C..NO |
| .C=C.N C.O.O | .C.O.C N=C.O | .O.C=C C.O.N | =C.O.C C.N.O |
| .C=C.N O.C.O | .C.O.C O.C=N | .O.C=C N.C.O | =C.O.C C.O.N |
| .C=C.N O.O.C | .C.O.C O.N=C | .O.C=C N.O.C | =C.O.C N.C.O |
| .C=C.N C..OO | .C.O.C C.=NO | .O.C=C O.C.N | =C.O.C N.O.C |
| .C.N=C C.O.O | .C.O.C C=.NO | .O.C=C O.N.C | =C.O.C C..NO |
| .C.N=C O.C.O | .O.C.C C.N=O | .O.C=C C..NO | =C..CO C.N.O |
| .C.N=C O.O.C | .O.C.C C=N.O | .O.C=C N..CO | =C..CO C.O.N |
| .C.N=C C..OO | .O.C.C N.C=O | .C.=CO C.N.O | =C..CO N.C.O |
| .C=N.C C.O.O | .O.C.C N=C.O | .C.=CO C.O.N | =C..CO N.O.C |
| .C=N.C O.C.O | .O.C.C O.C=N | .C.=CO N.C.O | =C..CO C..NO |
| .C=N.C O.O.C | | | |

| | | | |
|---|---|---|---|
| C...C C=N O.O | C=..C C.O O.N | C...O C=C O.N | C.=.O C.N C.O |
| C...C N=C O.O | C=..C O.C N.O | C..=O C.C N.O | C.=.O C.N O.C |
| C.=.C C.N O.O | C=..C O.C O.N | C.=.O C.C N.O | C.=.O N.C C.O |
| C.=.C N.C O.O | C...N C=C O.O | C.=.O C.C O.N | C.=.O N.C O.C |
| C=..C C.N O.O | C.=.N C.C O.O | C=..O C.C N.O | C=..O C.N C.O |
| C=..C N.C O.O | C=..N C.C O.O | C=..O C.C O.N | C=..O C.N O.C |
| C...C C.O N=O | C...N C.O C=O | C...O C.N C=O | C=..O N.C C.O |
| C ..C O.C N=O | C...N C=O O.C | C...O N.C C=O | C=..O N.C O.C |
| C...C C=O N.O | C..=N C.O C.O | C...O C=N C.O | N...C C=C O.O |
| C...C C=O O.N | C.=.N C.O O.C | C...O C=N O.C | N...C C.O C=O |
| C..=C C.O N.O | C=..N C.O C.O | C...O N=C C.O | N...C C=O O.C |
| C..=C O.C N.O | C=..N C.O O.C | C...O N=C O.C | N...O C.C C=O |
| C.=.C C.O N.O | C=..N O.C O.C | C..=O C.N C.O | N...O C=C C.O |
| C.=.C C.O O.N | C...O C.C N=O | C..=O N.C C.O | N...O C=C O.C |
| C=..C C.O N.O | C...O C=C N.O | | |

| | |
|---|---|
| C....C C O N=O | C....C O O N=C |
| C....C N O C=O | C....N O O C=C |
| C....C O O C=N | |

Table 2. The isomers of alanine (.C..CN C.=00 ) systematically ordered in DENDRAL-64 notation. Each "." or "=" stands for a single or double bond respectively which must be satisfied by a trailing atom or radical. This will be the first previously unreferenced item in the list to the right of the bond. A leading bond constitutes a central link, which must then be followed by two radicals. A space is used to separate the primary radicals for convenience in reading but has no coding significance. Some 25 of these topological possibilities are recognized chemical forms; an equal number are their tautomers. Most of the remainder are either peroxides or Schiff bases or similar unstable forms. A few, like hydracrylaldoxime, (.C.C.O C=N.O) might be realizable but were not found in a cursory search of the literature.

## COUNT OF CYCLIC TRIVALENT GRAPHS

[and genera of known chemical graphs]

| Vertices | Number of Chemical Rings[+] | With Hamilton Circuits: Polyhedra | With Hamilton Circuits: Unions (Planar) | With Hamilton Circuits: Gauche Forms (Non-Planar) | Without Hamilton Circuits: Planar Unions |
|---|---|---|---|---|---|
| 0 | 1 | 1* | | | |
| 2 | 2 | 1* | | | |
| 4 | 3 | 1* | 1* | | |
| 6 | 4 | 1* | 3* | 1[0] | |
| 8 | 5 | 2* | 10 [9] | 3[0] | 1* |
| 10 | 6 | 5 [4] | 37 [20] | 18[0] | 5 [2] |
| 12 | 7 | 14 [3] | 183 [35] | 133[0] | 30 [7] |
| 14 | 8 | 50 [3] | [45] | | |
| 16 | 9 | 233[3] [2] | [46] | | |
| 18 | 10 | 1249[3] [5] | [25] | | |
| 20 | 11 | [1] | [21] | | |
| 22 | 12 | [1] | [6] | | |
| 24 | 13 | [2] | [9] | | |
| ≥ 26 | ≥ 14 | [0] | | | |

Table 3.

Numbers in brackets are the count of genera of known examples from the ring index. * signifies all. Spiro forms are excluded from this count.

[1] Figures drawn in Lederberg (1965)
[2] Listed in Lederberg (1965)
[3] According to Grace (1965).
[+] This is one less than the number of faces of a polyhedron.