Organic chemistry is the object of this exposition. However, its subject is an elementary branch of non-numerical mathematics, the combinatorial theory of graphs. The purpose of the confrontation is to develop more formal representations of the statements and reasoning that underlie a branch of science, with the practical aim of mechanizing some of its intellectual tasks on the computer.

Analytical organic chemistry (hereinafter chemistry) was chosen as a promising branch of science for several reasons. Some of these are its technical importance for terrestrial and cosmic biochemistry; a large literature of successful solutions to intriguing problems; a heavy reliance on experimental data (which distinguishes these efforts from theorem-proving in plane geometry) but data which can be expressed formally without an elaborate translator of natural languages above all, the statements (assertions) of this field are proposed Structures of a kind that can be expressed in a compact and readily computable notation. That is to say, the structural formula is a conventional, widely accepted level of interpretation, that leaves us with some momentary satisfaction without having to reduce it to a complete analysis of inter-atomic forces, distances and angles.

For our purposes, chemistry can be axiomatized with the elementary rules of valence and a small inventory of bonds and atoms. We may also superimpose, in as much detail as we wish, some additional specifications like the alternating symmetry group of tetrahedral carbon, which is a concession to real stereochemistry, i.e., the 4

valences of the C atom are not freely interchangeable, but in general permute under two subgroups, (the conventional D vs. L attributes). With this elaboration, most chemistry is content with a topological description, a statement of connectivity, a graph. In the language of graph theory, the chemical atoms are nodes, the bonds are edges of a graph. The graph is an invariant with respect to its projection on a surface or to topological distortions of the length or form of the edges: all that matters is what is connected to what. Covalent bonds being unpolarized, we concentrate on undirected graphs.

Whether he knows it or not, the chemist is using graph theory from the very start of his studies in the field. Most structure proofs rely, eventually, on the exclusion of all but one of the possible isomers of a given composition. Ethanol must be CH3.CH2.OH because it has a replaceable -OH radical, and this is the only structure of the composition C2H6O that has one. Implied is a proof that there is no other graph with these properties, and that all the other ways of writing the graph, e.g., CH2(CH3).OH are redundant automorphisms that describe the same structure.

Unfortunately for the easy teaching of organic chemistry to college students,or to even balkier computers this step is left to intuition, or at least never supported by a general formal approach. Many students never grasp the concept of the invariant graph for the simple reason that they are never told about it, and then never develop their own techniques for the exhaustive enumeration of isomers that underlies much chemical thinking. The same difficulties plague our

clumsy, tradition-haunted systems of notation, so that in spite of (or because of) 50 pages of definitive rules in fine print, chemists will still argue about the proper name for a graph.

The main result advanced in this review is the development of an algorithm to generate a complete, irredundant set of isomeric graphs. Connected with this is a readable line-notation for structures, and rules for deciding on a unique canonical form for the description of any given structure. This system, named DENDRAL ("dendrite algorithm") has been tested and "proven" by being embodied in a computer program implemented in LISP, a list-processing formal language widely used for work in artificial intelligence. The DENDRAL program also has facilities for solving data-oriented problems in mass spectrometry by a crude emulation of the inductive processes used by human chemists taking cues from mass spectral data. These will be summarized briefly at the conclusion of this article.

The fundamental ideas on which DENDRAL is based are 1) the unique center of any tree-graph, 2) the mapping of rings and atoms onto nodes of a tree 3) the enumeration of rings (cyclic graphs) and 4) the value of a radical. The canonical form of a given structure is obtained by (1) looking in it for its center after (2) reducing its rings, (3) having identified the rings, and (4) having arranged the radicals linked to the center by order of their value. The value of a radical is obtained by looking, recursively, at its apical node and then the values of the radicals attached to it. Conversely, the generator algorithm produces

all possible isomers by making, seriatim, all possible specifications for the atom or bond at the center, then proceeding to produce all the radicals, in canonical order, that might be attached to it, consistent with the composition and other data. These ideas, summarized in table 1, will now be amplified and illustrated.

Trees. A molecule is defined as being a realization of an inseparable graph. That is, any two atoms are so linked, that some path through nodes and edges can be drawn between them. (By this definition, clathrate and catenate ring complexes are not a single molecule.) A pure tree is a 1-connected graph, i.e., cutting any edge will separate the tree into distinct graphs. A pure ring is a complex everywhere at least 2-connected, i.e., at least 2 edges must be cut to separate it. A pure ring will be defined separately, then regarded as a superatom (a complex momentarily treated as a node). This is a familiar idea in conventional notation we call toluene methylbenzene. Toluene is then identified as a tree of two nodes and one bond: ϕ-Me.

Invariant features of trees: center of mass.

In Geneva notation, trees are named as derivatives of the longest alkyl chain, and a complex series of additional rules is then needed to resolve ambiguities. This approach has been quite futile, at least in my hands, for a systematic generator of distinct forms. No existing notations, so far as I am aware, were designed with such an aim in mind. Nor have they been objectively tested, as they might be, for their consistency and uniqueness by implementing them as translator programs. Such programs would compute the correct notation from objective descriptions like connection tables. The approach I have adopted starts with an invariant feature of any tree, its center of mass.

(This choice is convenient, but arbitrary, and other features might be mentioned as alternatives, for example its diametric center.) Jordan ( ) had already shown that any tree has a unique center of mass and a unique center of diameter, not necessarily the same. The same idea is implicit in the first valid calculations of the number of alkanes, C_nH_{2n+2}, by Henze and Blair ( ), which had been incorrectly reasoned by Cayley ( ).

For these calculations, we ignore H's in our count of atoms. They are readily inferred when needed from the unsatisfied valence bonds. The center of mass is that node whose removal most evenly divides the tree into two or more radicals, according to their node-count. The central node is then one, now of whose attached radicals reaches half the total count. A molecule of even count may be entered on an edge, if two equal-count radicals are joined together.

The central partitions of count. For a tree with n atoms, we must first allocate 0 or 1 to the center, there the structure is classified by the integer partition of n among two or more radicals. We avoid redundancy by listing the radicals by their value; the count stands as the most significant cell of the vector that describes the value of a radical (Table 1). Table 2 gives some examples of numerical partitions that follow these rules.

The central node or bond. The compositional formula now contains the candidates for the assignment of the centroid, an atom or bond at the center. H's have been removed, but not before the corresponding number of unsaturations (double-bond-equivalents) are calculated.

The centroid is a unique feature, hence no graph having C allocated to its centroid can be isomorphic with a construction having an N, an O, and S, a ring superatom or a bond(which may be single, double or triple). The systematic use of this principle makes the generator irredundant and exhaustive.

The attached radicals. Once the centroid is fixed, the generator proceeds to form the valid sets of attached radicals from the residue of atoms in the composition pool. The order of decisions made by DENDRAL is often arbitrary, but usually based on some experience of programming convenience and familiarity. It might of course be reversed, for special applications. The existing concrete realization of the program is stated now.

List or vector valuation.

From this point on, DENDRAL scans or emits radicals in the sequence of their list-valuation, a process for which LISP is a particularly apt language. The list is treated as a vector, the cells of which may be either atoms or, again, vectors. Comparisons of a pair of lists are made from left to right, cell by cell, until a non-identity is found. The most significant position (as in natural numbers) is

to the left; hence the leftmost inequality determines the relative ordering of value of the pair. The first item for evaluating a vector is its length. This is the same as to say that vectors of unequal length are packed with implied zeroes to the left so as to equalize them (again like natural numbers, which can be regarded as vectors of digits).

A radical (as the etymology implies) is a rooted tree. To relate it to a list, a radical can be defined as a complex consisting of a atom pendant with zero, one, or more attached atoms or radicals. This definition is, obviously, recursive, since "radical" appears both in the definiendum and the definiens. In LISP notation, the depth of the recursion is indicated by the nesting of parenthesis. Hence, 2-aminoethyl is represented by ((- C (- C (- N)))) and dimethylamino by (- N (- C) (- C)). In the latter case, both C's are attached to the same N as indicated by their being nested in parentheses at the same level. N-methyl, N-ethylamino would be (- N (- C) (- C (- C))). The very monotony of symbols that makes such expressions unreadable to man most readily delivers them to the graces of a computer subroutine. For output, however, the computer readily translates such an expression into the form (.N.. CH CH2.CH3). The reader can work out the principles of interpreting these dot forms, with the help of this example, more quickly than he could read a tedious explanation. At this point he will also understand the entries in table 4 .

The cell that describes a radical then consists of 1) the afferent bond,

2) the apical or pendant atom and 3) the list of radicals attached to this. Table 1 must be consulted for the actual hierarchy of generation and evaluation. To put it briefly, the generator now fetches candidate atoms from the residual pool, partitions the remaining atoms to one or more radicals to be attached to the apical atom, and allocates single, double or triple bonds to the afferent link of the apical atom. Throughout, the program must respect the valence limits of the atoms and the contents of the residual pool, including unsaturations.

Redundancy is avoided by one further constraint: any list of radicals must be in monotonic (non-descending) order of value. This is built into the generator and is effected with a **maximum of anticipation and a minimum of retrospective weeding-out.**

The program also has some rules to keep it from wasting time on futile attempts to build radicals with more double bonds than available valences, and similar foolishness.

At this stage, the generator will build 216 isomers of C3H7NO2, as illustrated in Table. 4 . A perusal of that list will quickly show a fair amount of chemical nonsense, and no matter what their topological validity, we would prefer to eliminate monstrosities like the radical (.N..OH OH) or even (.O.NH.OH).

Chemical commonsense: BADLIST and DICTIONARY.

These features serve two important purposes. The BADLIST is a list of forbidden sub-graphs . It is applied at each round of radical-building to filter out prospective radicals that may contain any member of the list.

Although the matching program takes every advantage of the knowledge that preceding parts of the current structure have all been filtered before, it is still one of the most time-consuming routines in DENDRAL. The DICTIONARY is therefore established as an archives to store lists of the bonafide radicals of a given composition. During radical-building, when a cluster of atoms is allocated to a given radical, the next step would normally be the computation of allowable isomers of that composition. When the DICTIONARY is enabled, the program first searches it for an entry under that composition. If one is already there, the radicals are simply read out of memory rather than be re-computed and re-filtered. If the composition is a new one, the computation proceeds, and the results are then written into the DICTIONARY under the appropriate heading. Thus the results of solved sub-problems are saved from each run and embodied into the current state of the program.

This facility does a great deal to speed up the program. Complete dictionaries have been built of radicals of up to 4 atoms (Table 6). This list has been filtered by the BADLIST indicated in Table 5. Pushing this further tends to exhaust available fast memory, but some start has been made to putting the enlarged dictionaries on external storage (magnetic disc and tape files) and perusing these as indicated for a given problem.

The BADLIST and associated DICTIONARY are, of course, highly context-dependent. Our general-purpose DENDRAL has postulated the environment of natural products, and the arbitrary exclusion of peroxides and some other functional groups listed in table . These matters are under easy control by the programmer-user. However,

when the context is altered, disastrous inconsistencies may arise from attempts to shortcut the production of an entire new DICTIONARY. This is all too familiar an experience in human thought too.

Preferred radicals: GOODLIST.

For many purposes, it would be desirable to bias the order with which various isomers are generated, so that more plausible solutions to specific problems appear earlier in an output which may be, for all practical purposes, inexhaustible. Several facilities have been designed to rearrange the DENDRAL hierarchy without disturbing the eventual content of its output. Analytical data can furnish some of the cues to design the rearrangement.

The most important facility is GOODLIST, a list of preferred radicals. As far as possible, the program uses the atom pool to generate these fragments preferentially, which it does by replacing sub-compositions by corresponding superatoms. Thus the set C + C + O + U is replaced by a monovalent superatom labeled *COOH, and compositions containing this superatom are generated with higher priority. Before a structure is output, the superatoms are translated back to their expanded form so that the whole molecule can be filtered by BADLIST. Some dexterous programming is also needed to inhibit the redundant formation of the same fragments by normal radical-building, and to relate these embellishments to the DICTIONARY. (Human chemists, again, face quite analogous problems.)

Rings.

Owing to their symmetry, rings usually lack an invariant starting point analogous to the center of a tree. The enumerationaand classification of cyclic graphs is in fact a troublesome branch of mathmatics full of unsolved problems ( ). The most famous of these is the map-coloring conjecture. Efficient, general methods of producing all

are notable by their absence. However, it has been possible to develop a workable system by which all ring structures likely to be of practical significance can be computed. Chemical rings are reduced to mappings of linear segments on to the edges of strictly trivalent, cyclic graphs. These graphs have, in turn, been corrupted in a way that assumes their exhaustive listing, and renders them amenable to cannonical representatives. The representatives still have to be scanned for automorphisms. However, at most 2n permutations will have to be tested, usually much less, in contrast to the n! possibilities of the symmetric group.

The specific approach adopted here was selected only after the consideration of a number of alternatives, rejected mainly on account of inordinate computational effort. We require a feature which is invariant under systematic permutation of the orientation of the graph, i.e., the labels of its nodes. A brute force application of matrix algebra to n! permutations of row/column labels of the connection table could give a simple, general approach to identifying a canonical form for a ring of n atoms. However, this would be hopelessly time-consuming for a generator.

Reduction of rings.

The Orthomesh

The first level of simplification is to examine the ring structure without regard to the chemical or bonds identity of the atoms. Any cyclic molecule is mapped onto its saturated carbocyclic analog. This is exactly the approach of the conventional notation, e.g., 1,3-thiazepin postulates "epin" as a generic structure (cycloheptane), whose nodes

are all -C- unless otherwise indicated. This genus of structures is called an orthomesh, being reminescent of a planar mesh where conventionally displayed. Algorithmic numbering rules will soon be established, with which to specify heteroatom replacements in a canonical form.

Especially for deeply caged structures, the orthomeshes have some treacherous isomorphisms (Fig. 1) that we must take care to resolve.

The vertex group, or trivalent, cycliographs.

The next hierarchy of classification is obtained by reducing every linear chain of the orthomesh to a simple edge. The reduction leaves only the branch points or vertices of the ring in a family of structures we call a vertex-group. The most common vertex group is a degenerate one, with no vertices, i.e., the simple ring (Fig. 2).

Except for spiro-compounds, the vertex groups form the set of trivalent, cyclic graphs which have been the subject of some mathematical investigation ( ). The orthomesh can be reconstructed from the vertex-group by mapping a list of chain-lengths onto the list of edges of the vertex-group graph. A chain length is sometimes zero (orthofusion of adjacent cycles), when two vertices are directly linked. Loop pairs, even triples, are however admissible between two nodes, as is a self-loop (e.g. for spiro forms. See Fig. 2).

Now we have but to enumerate the vertex groups. An analytical theory for the cubical graphs does not yet exist ( ) but some practical tables have been produced by a computer program for all reasonable orders of complexity. The program is based on the combinatoric of Hamilton Circuits, which are closed circuits passing once through each node and the same number of edges (2/3 the total) of the trivalent graph.

Not every vertex group has a Hamilton Circuit -- note, for example, Figure 3. However, the occurrence of such graphs can be anticipated from the properties of the graphs of the next lower order. The table of vertex groups can then be made as complete as needed for any practical purpose. The scope of existing tabulations is shown in Table

Spiro-structures, with some quadrivalent vertices, can be dealt with in the same way. These are enumerated from the trivalent graphs by "shrinking" any tetrangular face into a quadrivalent node in all possible ways. Their vertex-groups can, however, be described directly as Hamilton-circuits. Many of the trivalent graphs are planar and can be projected as meshes with no crossing edges. However, starting with Fig. we find a number of non-planar forms. Kuratowski ( ) showed that any trivalent graph, if non-planar, contained this one, which gives a rather easy criterion for detecting

non-planar forms during computation. Remarkably, non-planar forms have yet to be synthesized or identified by organic chemists, perhaps because of the trickiness of the caging already implied by Fig. . However, a hypothetical example of such a molecule is indicated in Fig. , and is perhaps amenable to the present level of art. At any rate, the dictionary of vertex groups is shortened by confining it to planar forms. Nevertheless, many hypothetical vertex groups have not yet found their way into the ring index. One of the smaller examples of a vertex-group yet to be realized, and another challenge of hypothetical chemistry is Fig.

The Hamilton Circuit can be displayed as a polygon. The trivalent graph is then completed by drawing appropriate chords to connect pairs of nodes. This chorded polygon leads to a description of the Hamilton Circuit as a list of span lengths. Each node is attached by its chord distant to another node by a span of 1 or more. From the list of spans, it is easy to reconstruct the chorded polygon.

The span list itself belongs to a set of obvious automorphisms -- obtained by rotations and reflection -- which are easily weeded out. The canonical form is taken as the lowest-valued of the set. In addition, a given graph may show several non-congruent Hamilton circuits. A definitive characterization then depends on a program that efficiently finds all the Hamilton circuits of the graph ( ) working from a connection table.

A complete span list for n vertices takes n characters. By skipping the second terminal, which is already specified by the span

from the first terminal, labels of n/2 characters will suffice. In fact, n/2-1 will do since the final chord is fixed when all the others have been. This process gives some shorthand words, as in table , to simplify lookup. These words are, however, sufficient to define the individual vertex groups.

It is sobering to reflect that much of this fuss is spent on a vanishingly small proportion of rings, Most organic chemistry is concerned with simple rings (the degenerate vertex-group "0"), two-ring fusions like naphthalene, the "hosohedron", vertex-group A, and some three-rings either AA (anthracene) or BB tricyclobutane or ancenapthene.

The canonical form of the vertex group implies the sequence, and therefore the proper numbering of the nodes and edges, which is the order of presentation in the formula. The set of Hamilton circuits of a given graph also delivers its symmetries, which (we will not say again) must be kept in mind during subsequent operations.

The Generator. To recapitulate, the generator is equiped with the compact, precomputed list of vertex groups. For each vertex group, the set of possible orthomeshes is computed by allocating the various allowable partitions of the count of atoms to the list of chain lengths. Then hetero-atoms areeallocated, first to vertices, then to the positions on the chains between vertices. Finally the unsaturations are allocated.

When the ring is completed, it may still have to be hung on a tree, taking account of positional isomerism for the attachments.

Chirality.

The DENDRAL notation implies a relative weight to the attachments on any atom, whether this be in a tree or part of a ring. So far the

four valences of C have been regarded as equivalent, i.e., subject to the symmetric group S_4. For examination of chirality DENDRAL analyzes the absolute assignment of radicals to the four valences. It then asks whether the relative values of the four radicals correspond to those of a standard "even" or "odd" tetrahedron (Fig.). The DENDRAL weights of the pendant radicals do not always correspond to those of conventional notational schemes, but it is easy to incorporate any well-defined scheme into a translator program.

The symmetry of the molecule sometimes obliterates the difference between, say, a DD- and an LL- configuration. DENDRAL prefers to save one of these as canonical, and recognize the symmetry, rather than go through the tortures of cis-/trans and meso- terminologies. Other conventional descriptions of chirality can usually be inferred from absolute assignments. This principle gives us a unified notation for describing any form of isomerism which stems from the division of the tetrahedral symmetric group into two incongruent forms.

Implementation

Noncyclic DENDRAL is fully implemented as a LISP program running on a variety of hardware. At Stanford, we have done most of the programming on the Artificial Intelligence Project's Digital Equipment Corporation PDP-6 with 64K fast core, arranged for limited time-sharing use. The program has routinely been run from Bethesda, Md. with commercial telecommunications. Batch jobs have been run on the

IBM/360-65 of the Stanford Computation Center. Much of the original programming was done by remote connection on the Q-32 experimental time-sharing system of the System Development Corporation, Santa Monica, California. The program takes about 35K storage plus space for the dictionary.

A primitive model of mass spectrometry, mainly the fragmentation patterns for various kinds of bonds and functional groups, has been embodied in the program, and this is now moderately successful in inducing and testing hypotheses of structures to match lists of mass numbers input to it. The main heuristic is the dynamic reformulation of BADLIST and GOODLIST in response to cues from the data, and partial results apparently compatible with them.

This work is actively in progress at present. A partly overlapping group of my colleagues (Dr. C. Djerassi, Dr. E.C. Levinthal) is engaged in the automated handling of samples for mass spectrometry and direct transfer of m.s. data to the computer, for eventual interfacing with the DENDRAL program.

The generation of cyclic structure hypotheses is partially completed. The program still leaves room for a number of shortcuts to apply general information about symmetries in a prospective fashion, to allow about a tenfold economy in the time needed to generate and weed out redundant forms. Table is based in part on a hand simulation of the generator algorithm.

Acknowledgments