OFFPRINTS FROM
BIOCHEMICAL APPLICATIONS OF MASS SPECTROMETRY
EDITED BY GEORGE R. WALLER
JOHN WILEY & SONS, 1972

CHAPTER 7

USE OF A COMPUTER TO IDENTIFY UNKNOWN COMPOUNDS: THE AUTOMATION OF SCIENTIFIC INFERENCE*

JOSHUA LEDERBERG

Department of Genetics, School of Medicine, Stanford University, Stanford, California

A. INTRODUCTION

The Argentinian writer Jorge Luis Borges, in a short story called "The Library of Babel," showed that all knowledge can be reduced to a problem of selection. He portrayed a library of infinite dimensions filled with books printed in an obscure code in which familiar phrases occasionally appeared. Eventually, a mathematician-inhabitant of this space surmised that each book was one of all possible random concatenations of letters. After a few centuries of discouragement, the inhabitants were inspired by a new revelation—that the library must in fact contain all knowledge. The problem was merely one of selecting the proper texts.

The identification of an unknown compound presents a similar challenge. If the universe of possibilities were infinite, the problem might not be rigorously soluble. Practical solutions depend upon the ingenuity with which the domain of acceptable solutions can be narrowed within a particular experimental context and the efficiency with which tentative solutions can be tested against the data.

The previous chapter deals with the pragmatics of searching the index to a finite library, i.e., the catalog of mass spectra of previously studied molecules, with occasional extensions to related structures. The present chapter deals with chemical structures in more theoretical terms, as part of an effort to embody scientific inference in a computer program. Instead of listing known structures, this program, DENDRAL,* incorporates rules by which all con-

*This report is a summary of the current status of the Heuristic DENDRAL project conducted jointly by the Departments of Chemistry, Computer Science, and Genetics at Stanford University under the direction of Professors Carl Djerassi, Edward A. Feigenbaum, and Joshua Lederberg. This research was financed by the Advanced Research Projects Agency (Contract SD-183), the National Aeronautics and Space Administration (Grant NGR-05-020-004), and the National Institutes of Health (Grant AM-04257). Most of the programming reported here was done by Dr. Bruce Buchanan, Mrs. Georgia Sutherland, Mr. Allan Delfino, and Dr. Armand Buchs.

*The program is called DENDRAL (for DENDRitic ALgorithm). It is written in the list-processing language LISP. It requires 40,000 or more words of memory, depending on the number of atoms in the

ceivable structures can be generated and encoded into a fairly legible but computer-compatible notation (1). In the general case, the generator is constrained only by the elementary rules of valence of the various atoms. In practice it also includes many heuristics that limit its speculations to plausibly stable structures, and further to those of particular interest to the line of chemistry in which it is applied. Besides allowing for the exhaustive enumeration of all possible structures, DENTRAL is also devised to be irredundant—it allows for the presentation of a given structure in a single standardized, or *canonical* notation. The program is also prospectively efficient, so that most redundancies are anticipated and prevented, rather than having to be weeded out after having been formulated.

The primary motivation of the Heuristic DENDRAL project is to study and model processes of inductive inference in science, in particular, the formation of hypotheses that best explain given sets of empirical data. The task chosen for detailed study is the structure determination of organic molecules, and this has been advanced furthest with MS data (1–8). However, the principles are readily generalized to other data for which some chemical theory can be formulated.

The motivation and a general outline of the approach are presented first. Next, a sketch is given of how the program works and how good its performance is at this stage. Last, an example, taken from our group's recent work on aliphatic ethers (2), is shown.

B. MOTIVATION

The DENDRAL project aims at emulating in a computer program the inductive behavior of the scientist in an important but sharply limited area of science, organic chemistry. Most of our work is addressed to the following problem: Given the data of the mass spectrum of an unknown compound, induce a workable number of plausible solutions, i.e. a small list of candidate molecular structures. In order to complete the task, the DENDRAL program then deduces the mass spectrum predicted by the theory of mass spectrometry for each of the candidates and selects the most productive hypothesis, i.e., the structure whose predicted spectrum most closely matches the data.

We have designed, engineered, and demonstrated a computer program that manifests many aspects of human problem-solving techniques. It also works faster than human intelligence in solving problems chosen from an appropriately limited domain of types of compounds, as illustrated in the cited publications (1, 2).

Some of the essential features of the DENDRAL program include the following:

1. Conceptualizing organic chemistry in terms of topological graph theory, i.e., a general theory of ways of combining atoms.
2. Embodying this approach in an exhaustive HYPOTHESIS GENERATOR. This is a program that is capable, in principle, of "imagining" every conceivable molecular structure.
3. Organizing the GENERATOR so that it avoids duplication and irrelevancy and moves from structure to structure in an orderly and predictable way.

The key concept is that induction becomes a process of efficient selection from the domain of all possible structures. Heuristic search and evaluation is used to implement this "efficient selection." Most of the ingenuity in the program is devoted to heuristic modifications of the GENERATOR. Some of these modifications result in early pruning of unproductive or implausible branches of the search tree. Other modifications require that the program consult the data for cues (feature analysis) that can be used by the GENERATOR as a plan for a more effective order of priorities during hypothesis generation. The program incorporates a memory of solved subproblems that can be consulted to look up a result rather than compute it over and over again. The program is aimed at facilitating the entry of new ideas by the chemist when discrepancies are perceived between the actual functioning of the program and his expectation of it.

C. IMPLEMENTATION

1. Generator

As just noted, (11, 13–15), the DENDRAL program contains a structure GENERATOR as its core, abundantly constrained by a set of relevant heuristics. The GENERATOR is built upon a consideration of the conventional structure representation as a topological graph, i.e., the connectivity relations of a set of chemical atoms taken as nodes. 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

composition and the speed with which one wants to see the answers. Many options are available to the chemist at the teletype console; for instance, he can revise the program's theory of chemical instability (BADLIST), he can restrict structure generation to molecules of a specified class (GOODLIST), or he can monitor the structure-generation process through a dialogue with the program. Programming details are available (9).

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. Difficult problems such as the analytical enumeration of polyhedra remain unsolved.

This chapter reviews some elementary features of graphs that may be used for a systematic outline of organic chemistry.

2. 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, 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 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 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 then reflects the 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 bond or edge that joins them.

Each of the centers (Fig. 7-1) 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, 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. Computer programs typically reduce the most complicated descriptions, including matrices of arbitrary dimensions, to linear strings of symbols. The internal description of chemical graphs within the DENDRAL program is a technicality we need not elaborate on here. The choice is arbitrary but includes a compromise between compactness of the code and its compatibility with the conventions of the LISP language.

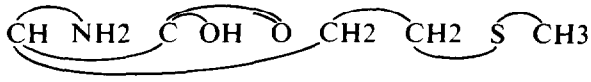
An external linear notation for chemical graphs (i.e., structures) has, however, also been defined. In conjunction with the canons of ordering the radicals, it yields an unambiguous but readily decipherable, fairly compact code for any molecule. It may then be useful for problems of retrieval in library searches, as well as writing dictionaries and catalogs, as well as for the computer input and output for which it was constructed. Notation is, however, a secondary problem in the immediate environment of a computer, for its programs can readily be formed to translate from any format to any other. Programs to translate from DENDRAL notation to connection tables and back to canonical DENDRAL have been operative for some five years.

Here is an elementary example of the external,

DENDRAL "dot" notation, illustrated with methionine:

CH . . . NH2 C . : OH O CH2 . CH2 . S . CH3

a code which is interpreted



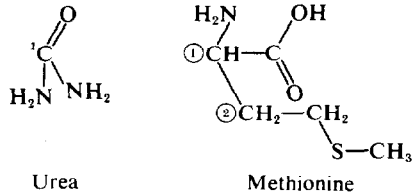
the . and : symbols denoting bonds from the preceding atom. In fact, the H's are fully implied, given the known valence of the other atoms. The formula could also be encoded, and the atoms then could be numbered in canonical sequence as follows:

C . . . NC . : OOC . C . S . C
1 23 456 7 8 9

a form nicely handled in the computer, and appropriate for dictionary codes, but a needless obstacle for the human chemist to interpret.

One can also specify an unlimited number of arbitrary abbreviations for various clusters of atoms, as has been done in the well-known Wiswesser line notation. The designation of —COOH or . C. : OHO as VQ confers a small advantage in brevity, which could also be automatically computed and incorporated in the DENDRAL notation. According to our own experience, the difficulty in reading such abbreviations diminishes their practical value. For example, we do not encode the syllables of English words by compact, arbitrary designators except for special purposes such as telegraphic transmission—and these can be dealt with ad hoc by the computer.

Nevertheless, a few trivial abbreviations have been incorporated into the output conversion routines of DENDRAL and are modified according to the taste of each user. They include such constructions as —CH2OH, —CHO, —COOH, and (for *n*-alkyl) forms like —C3H7 for the corresponding DENDRAL codes: . CH2 . OH, . CH : O, . C . : OH O, and . CH2 . CH2 . CH3; further, as mentioned, the user can insert any others he wishes.



Urea Methionine

Figure 7-1. 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 . . . 134. Carbon atom 2 is the radius center, on a diameter of 7, that is, the center of a largest string

(C—S—C—C—C—C=O).

For both analyses, we ignore hydrogen atoms.

Some 30 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, $C_{11}H_{24}$, one atom is designated as center, leaving 10 to be allocated among 2, 3, or 4 branches. Only the following partitions shown in Fig. 7-1 satisfy the rules (leaving dissymmetry out of account):

Branches	*Partitions*	*Number of Partitions*
2 C(□)(□)	5 | 5	1
3 C(□)(□)(□)	1 2 2 3 | 4 3 4 3 | 5 5 4 4	4
4 C(□)(□)(□)(□)	1 1 1 1 1 2 2 | 1 2 1 2 3 2 2 | 3 2 4 3 3 2 3 | 5 5 4 4 3 4 3	7

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 eicosanes, $C_{20}H_{42}$, and 5.622,109 eicosanols, $C_{20}H_{41}OH$ (see Table 7-I).

The total range of acyclic compounds containing atoms other than that of the hydrocarbons (C, H) is, of course, very much larger than these subsets. To

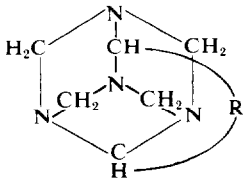
Table 7-I Counting the Different Arrangements of Compounds of Carbon and Hydrogen Containing No Double or Triple Bonds and No Rings (general formula $C_n H_{2n+2}$).

Number of Carbon Atoms	Number of Possible Isomers	Number of Carbon Atoms	Number of Possible Isomers
1	1	11	159
2	1	12	355
3	1	13	802
4	2	14	1.858
5	3	15	4,347
6	5	16	10,359
7	9	17	24,894
8	18	18	60,523
9	35	19	148,284
10	75	20	366,319

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, e.g., 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}.

3. 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. Supplement III (1965) carried this listing to 14,265 rings, indexing the forms that had appeared in the literature up to that time. Faithfully reporting the actual practice of chemical nomenclature, the Ring Index also summarizes a profusion of synonyms and and arbitrary numbering systems. Many thousands of additional rings have been reported since 1965, and these are still a small proportion of the possible topological combinations. Indeed, no ring has yet been reported that would correspond to the whole genus of nonplanar graphs, e.g., the hypothetical



which is related to the "gauche" structure labeled CCC in Fig. 7-2.

Molecules may also contain both acyclic and cyclic parts. However, if a strictly cyclic part has been 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 chemical graphs. 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 7-2 gives some of the more familiar cyclic hydrocarbons to illustrate these correspondences.

The trivalent graphs relevant to chemical problems have been exhaustively generated through a consideration of the Hamiltonian circuits, i.e., circular paths that pass once through every node. It is then possible, in principle, to extend the GENERATOR to the full set of cyclic molecules. In practice, the context of a problem or specific cues from the data usually make this effort unnecessary. This style of exhaustive enumeration has, however, been helpful in solving structural problems without recourse to the computer (16). The *efficient* implementation of a cyclic structure generator is still in the process of completion: inefficient and restricted versions have been exercised.

4. Heuristics

The HEURISTIC DENDRAL process of analyzing a mass spectrum consists of three phases. The first, preliminary inference (or planning), obtains clues from the data as to which classes of chemical compounds are suggested or forbidden by the data. The second phase, structure generation, enumerates chemically plausible structural hypotheses which are compatible with the inferences made in phase one. The third phase, prediction and testing (or hypothesis validation), predicts consequences from each structural hypothesis and compares this prediction with the original spectrum to choose the hypothesis that best explains the data. Corresponding to these three phases are three subprograms. The program(s) have been detailed in previous publications, primarily in the book *Machine Intelligence 4* (9) and in a series of Stanford Artificial Intelligence Project Memos (9–12).

The PRELIMINARY INFERENCE MAKER program contains a list of names of structural fragments, each of which has special characteristics with respect to its activity in a mass spectrometer. These are called "functional groups." Each functional group has associated with it a set of spectral values and relationships among these values that are, to the best of our present knowledge, "diagnostic" for the chemical functional group. Other properties of the functional group indicate which other groups are related to this one – as special or general cases.

The program progresses through the group list, checking the conditions for each group. Two lists are constructed for output: GOODLIST enumerates functional groups that might be present, and BADLIST

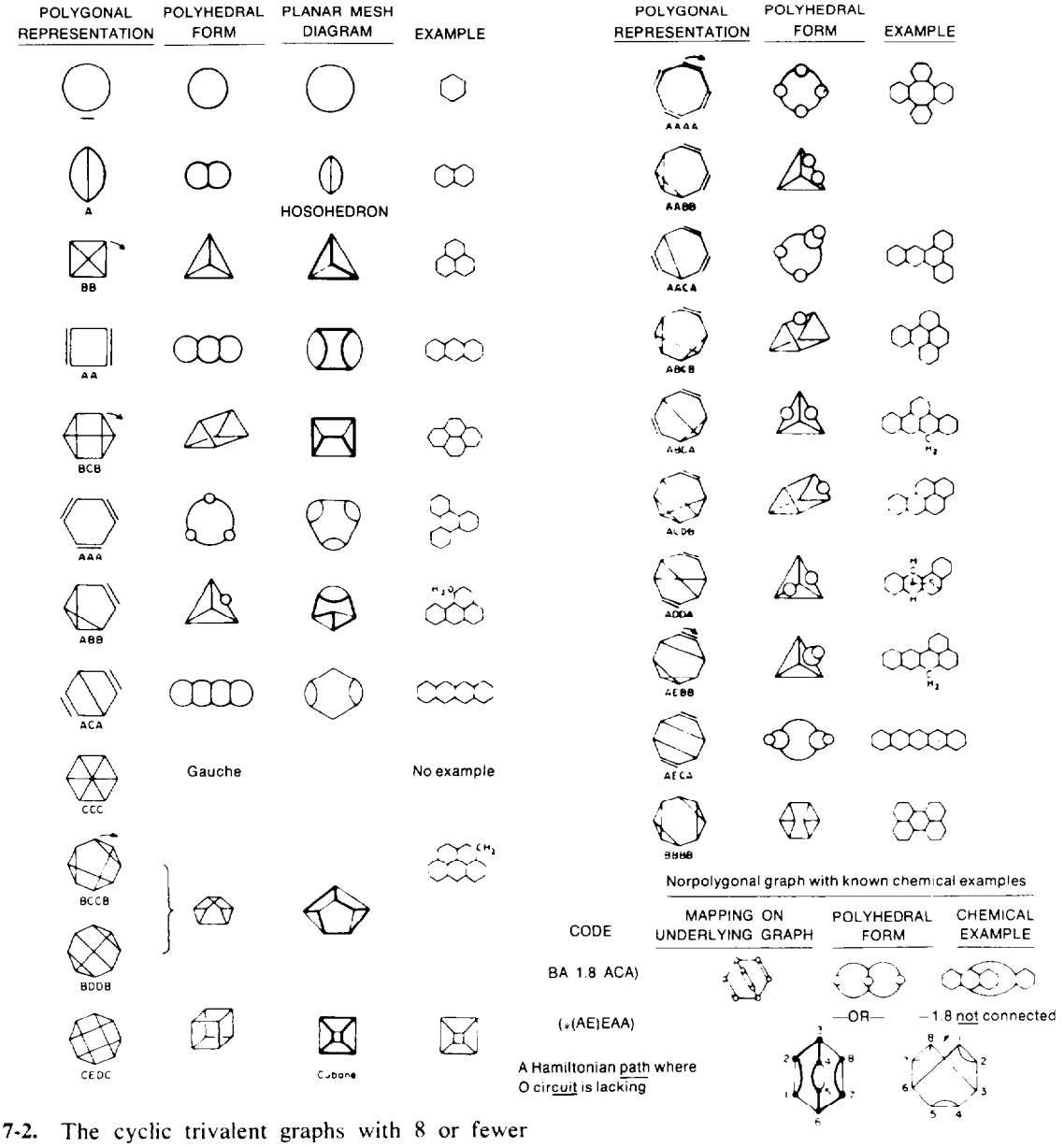


Figure 7-2. The cyclic trivalent graphs with 8 or fewer nodes. Up to 6 nodes, these all have Hamilton circuits but may also be represented in other ways. In a few examples, the circuits are drawn with emphasis on planar map representations. Complete tables of chord lists like those shown under the circuit (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. 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.

lists functional groups that cannot be in the substance that was introduced to the mass spectrometer.

GOODLIST and BADLIST are the inputs to the STRUCTURE GENERATOR, which is a generator of isomers (topologically possible graphs) of a given empirical formula (collection of atoms). GOODLIST and BADLIST control and constrain the generation of paths in this space. Each GOODLIST item is treated as a "superatom," so that any functional group inferred from the data by the PRELIMINARY INFERENCE MAKER will be guaranteed to appear in the list of candidate hypotheses output by the STRUCTURE GENERATOR.

The third subprogram is the Mass Spectrum PREDICTOR, which contains what has been referred to as the "complex theory of mass spectrometry." This is a deductive model of the processes that affect a structure when it is placed in a mass spectrometer. Some of these rules determine the likelihood that individual bonds will break, given the total environment of the bond. Other rules are concerned with larger fragments of a structure such as the functional groups which are the basis of the PRELIMINARY INFERENCE MAKER. All these rules are applied (recursively) to each structural hypothesis coming from the STRUCTURE GENERATOR. The result is a list of mass-intensity number pairs, which is the predicted mass spectrum for each candidate molecule.

Any structure is discarded which appears to be inconsistent with the original data (i.e., its predicted spectrum is incompatible with the given spectrum). The remaining structures are ranked from most to least plausible on the basis of how well their spectra compare with the data. The top ranked structure is considered to be the "best explanation."

Thanks to the collaboration of Dr. Gustav Schroll, an NMR (Nuclear Magnetic Resonance) PREDICTOR and INFERENCE MAKER have been added to the program. Thus the program can confirm and rank candidate structures through predictions independent of mass spectroscopy, bringing the whole process more in line with standard accounts of "the scientific method." Thus the HEURISTIC DENDRAL program is expanding from the "automatic mass spectroscopist" to the "automatic analytical chemist." Other analytical tools, such as infrared spectroscopy, will be incorporated eventually. Only the clumsiness of the language hinders further extensions to conventional "wet chemistry" reactions.

Interaction and interdependence of the three subprograms of HEURISTIC DENDRAL must be mentioned when discussing these computer programs. Because of the size of the combined programs, it is more practical to run them separately than to run them together. One supervisor takes care of the interaction by having each subprogram write an output file which is then the input file for the next phase of program operation. The PRELIMINARY INFERENCE MAKER writes the file containing the empirical formula and the GOODLIST and BADLIST to be used by the STRUCTURE GENERATOR. That program, in turn, reads this file and writes another file containing the single output list of structures which it generates according to the GOODLIST and BADLIST specifications. The PREDICTOR then reads this file to obtain its input and calculates a mass spectrum for each structure in the file. If other tests such as NMR prediction are to be made on the candidate structures, the supervisor interfaces the appropriate program to these others in the same way.

D. COMMENTARY

One reason for the high level of performance of the program is the large amount of MS knowledge chemists have imparted to the program. Obtaining this has been one of the biggest bottlenecks in developing the program. At present there is no axiomatic or even well organized theory of mass spectrometry which we could transfer to the program from a textbook or from an expert. Most of the chemical theory has been put into the program by a programmer who is not a chemist but who spent many hours in eliciting the theory from the chemist-expert. In many cases the chemist's theory was only tentative or incompletely formulated, so that many iterations of rule formulating, programming, and testing were necessary to bring the DENDRAL program to its present level of competence.

A few general points of strategy have emerged from the DENDRAL effort. With regard to the theoretical knowledge of the task domain in the program, we believe that the following considerations are important:

1. It is important that the program's "theory of the real world," i.e., of pertinent branches of chemistry, be centralized and unified. Otherwise, during the evolution of a complex program, any stage of which is an arbitrary simplification, inconsistencies will accumulate. For example, one module of the theory may expect organic compounds to contain sulfur, although sulfur is denied in another portion of the theory.

2. It would be advantageous for the program to derive planning (Preliminary Inference) cues from its own theory, by introspection, rather than from external data which may not yet have been assimilated into its theory. The success of the program depends in every case on the validity of the theory, so there is no use going beyond it. It is more efficient for the computer to generate hypothetical spectra and search for the relevant "diagnostic" patterns in them than to wait for experimental data. The theory should be responsive to the data; then the list of inference cues should be generated from the theory.

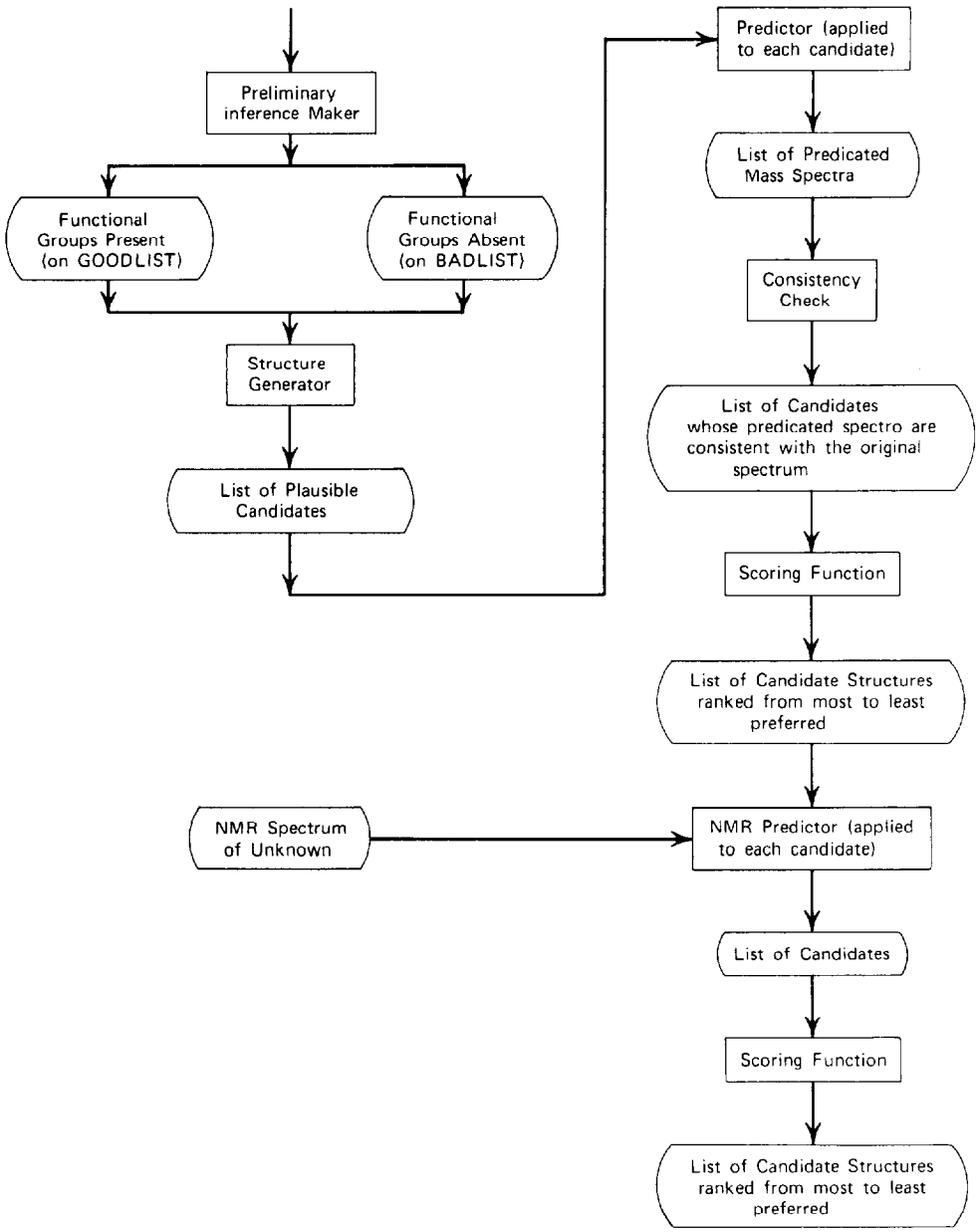
3. Separating the theory from the routine which uses it facilitates changing the theory to improve it, on the one hand, or to experiment with variations of it, on the other. Although scattering the theory in the program's LISP code increases running efficiency, it seems more desirable, at this point, to increase the program's flexibility. This has led us to design the programs in a form we refer to as "table-driven."

Reference 12 contains a more complete discussion of this effort.

E. EXAMPLE [DIAGNOSING THE STRUCTURE OF AN ALIPHATIC ETHER FROM LOW-RESOLUTION MASS SPECTRA AND NUCLEAR MAGNETIC RESONANCE DATA (2)]

A diagrammatic representation of Heuristic DENDRAL is depicted in Scheme 7-A. Given an unknown mass spectrum (Fig. 7-3) and the empirical formula of the molecular ion, the program must infer the presence of the correct functional group, which is the ether group here. This information is obtained by the PRELIMINARY INFERENCE MAKER* and is then used by the STRUCTURE GENERATOR to compile exhaustive and irredundant lists of candidate structures containing this functional group. Truncation of the list of candidate structures is achieved by the PREDICTOR section of Heuristic DENDRAL, in which a predicted mass spectrum for each possible structure is compared to

*Program MODULES are labeled in small capital letters.



Scheme 7-A. Conceptualization of Heuristic Dendral.

the original unknown (Fig. 7-3). Any irreconcilable difference between the unknown and predicted mass spectra results in the rejection of that candidate structure from further consideration. All the viable structures are then processed by the SCORING FUNCTION, which ranks them in order of preference. At this level of the program an NMR spectrum is predicted for each surviving candidate and the results are compared to the NMR spectrum of the unknown compound. In our experience this yields only one acceptable structure. The decision rules and the structure of Heuristic DENDRAL are perhaps best appreciated in a step-by-step discussion of its solution to a given problem.

The criteria for Heuristic DENDRAL to infer the presence of an ether function from an examination of an unknown low-resolution mass spectrum and the composition of the molecular ion are summarized in Scheme 7-B*. The program acknowledges the presence of the ether subgraph by checking for affirmative answers to the following specific points. Peaks corresponding to the loss of 17 and 18 amu, respectively, are below 2% relative abundance:† the empirical composition of the molecular ion must be consistent with the presence of an ether linkage within a saturated molecule and two alkyl ions

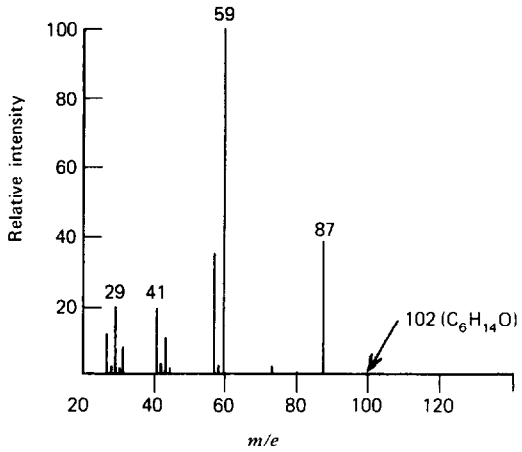
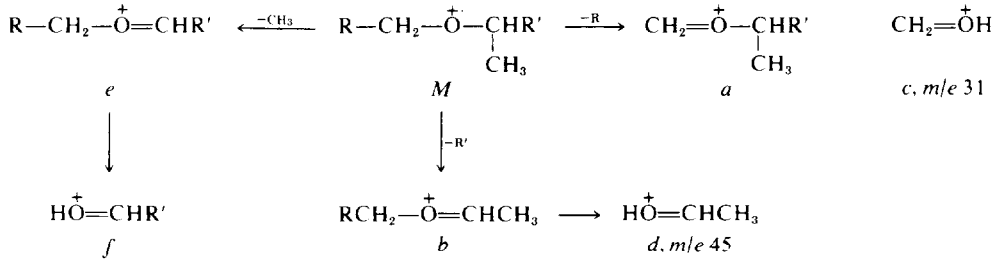


Figure 7-3. Mass spectrum of unknown aliphatic ether of composition $C_6H_{14}O$.

matical relationship of the α-cleavage processes, e.g., of an ether 3: $M+58=a+b$. Thus for the program to respond that an ether 3 subgraph is present, it must recognize two peaks whose sum is equal to the molecular weight plus 58 amu. For an ether 2, ether 4, ether 4A, ether 5, and ether 6, the masses of the radicals duplicated in α-cleavage are 44, 72, 72, 86, and 100 amu, respectively. The values depicted



corresponding to the alkyl chains flanking the ether-oxygen atom must be present. Should these conditions be satisfied, then Heuristic DENDRAL attempts to expand the ether subgraph into any of the six subgraphs depicted in Scheme 7-B. If any condition fails, none of these other ethers will be considered. The degree of substitution on either α-carbon atom will affect the masses of the products of α-cleavage of aliphatic ethers. The α-cleavage peaks referred to in Scheme 7-B have their origin in the following mathe-

in Scheme 7-B as 31...high, 45...high, etc., correspond to the mass of the rearrangement ions *c* and *d* for the case of an ether 3.

The following responses were generated by Heuristic DENDRAL as it processed a typical problem. The operator initiates the program by typing the following command†:

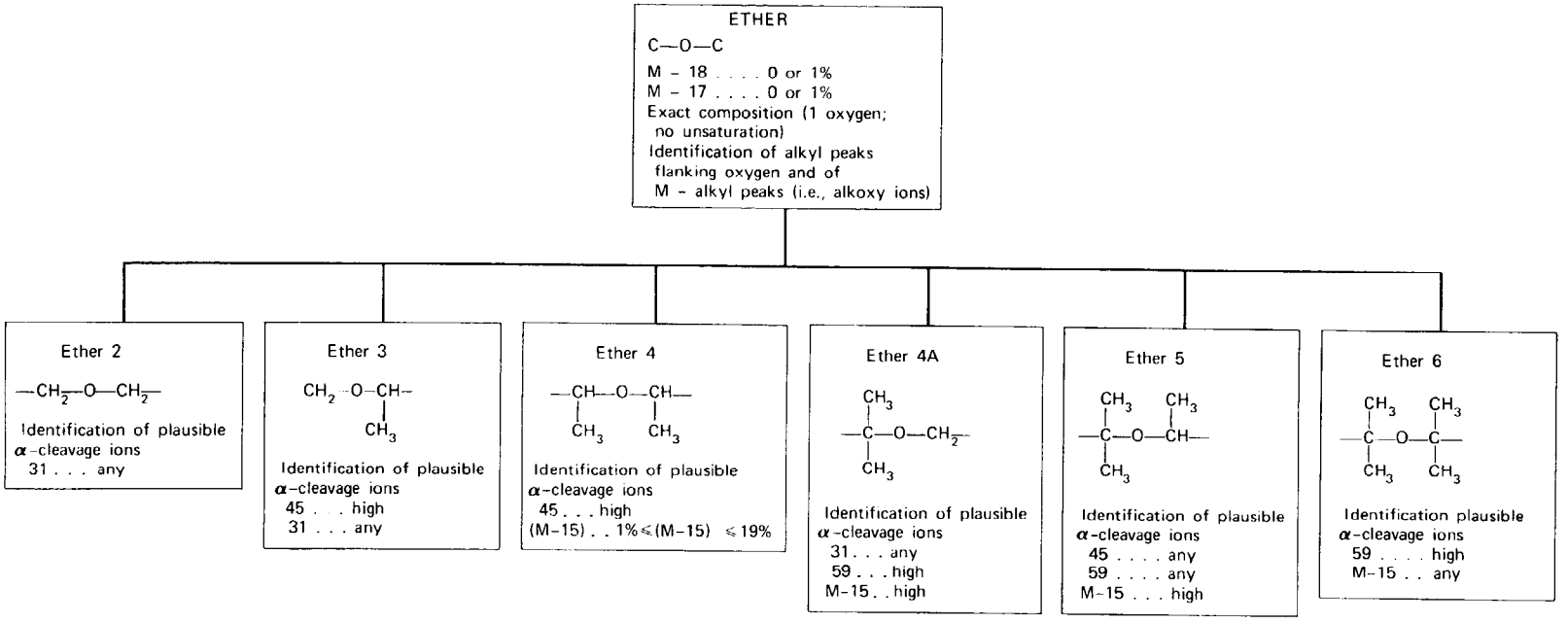
```
*(INFER (QUOTE C6H140) S:ETH-TERT-BUT
        (QUOTE TEST!!))
```

The program fetches the low-resolution mass spectrum in question, and following an initial examination

*High, > 10% relative abundance; any, ≥ 1% relative abundance.

†The empirical composition of an ether is also compatible with the presence of a hydroxyl group. However, alcohols show appreciable peaks (> 2% relative abundance) in their mass spectra corresponding to the loss of water from their molecular ions. Furthermore, the mass spectra of some aliphatic ethers (18) display weak peaks (< 2% relative abundance) due to the expulsion of 17 amu.

†S-ETH-TERT-BUT is the code under which the "unknown" low-resolution mass spectrum (Fig. 7-3) is filed. It corresponds to the data recorded (18) for ethyl *t*-butyl ether and TEST 11 is the name of the storage location in which results will be kept for later use.

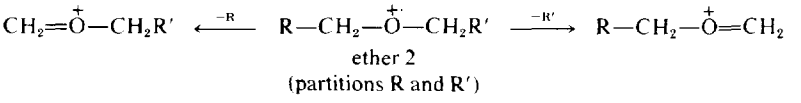


Scheme 7-B. Rules for ether identification

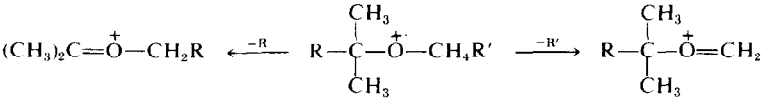
of Fig. 3 by the PRELIMINARY INFERENCE MAKER the computer responds with

```
*GOODLIST = (*ETHER2!* *ETHER4A!*)
*PARTITIONS = ((*ETHER 2!* 15. 43.)
               (*ETHER4A!* 15. 15.))
```

The program deduces that both the ether 2 and ether 4A subgraphs (Scheme 7-B) are consistent with the information contained in Fig. 7-3. (GOODLIST, as the name implies, is a list of subgraphs thought to be particularly good for solving the problem at hand.) Furthermore, it defines partitions which correspond to the alkyl chains expelled in the α-cleavage fragmentation of an ether 2 and an ether 4A



ether 2
(partitions R and R')



ether 4A
(partitions R and R')

(where R and R' are partitions for hypothetical ether 2 and ether 4A subgraphs). Finally, subgraphs that appear to be poor solutions for this problem—subgraphs whose conditions are violated by Fig. 7-3—are placed on BADLIST. For example, alcohol subgraphs are placed on BADLIST since Fig. 7-3 contains no prominent peak due to the loss of water from the molecular ion.

```
*BADLIST = (*C-2-ALCOHOL* *PRIMARY-
ALCOHOL* * ALCOHOL* *ETHER* *ETHER
4* *ETHER3*
```

The command†

```
*(EXPLAIN (QUOTE TEST11) (QUOTE
TEST11A)(QUOTE MAR2ø))
```

instructs that part of the program known as the STRUCTURE GENERATOR to locate the output of the PRELIMINARY INFERENCE MAKER (in file TEST 11) and the STRUCTURE GENERATOR then builds all the candidate structures consistent with the GOODLIST and BADLIST constraints, leaving the result in the external file under the label TEST 11A. The teletype response is in the following form:‡

```
(FILE READ)
(NOVEMBER-15-1968 VERSION)
C4*ETHER2!*H1ø
MOLECULES    NO DOUBLE BOND EQUIVS
  1.          CH2..C3H7        O.C2H5,
  2.          CH2..CH..CH3 CH3 O.C2H5,

(NOVEMBER-15-1968-VERSION)
C2*ETHER4A!*H6
MOLECULES    NO DOUBLE BOND EQUIVS
  1.          C....CH3 CH3 CH3 O.C2H5,

DONE
*
```

The PREDICTOR section of Heuristic DENDRAL (see Scheme 7-A) is made operational by typing the sentence

```
*(SCORE (QUOTE TEST11A) S:ETH-TERT-
BUT)
```

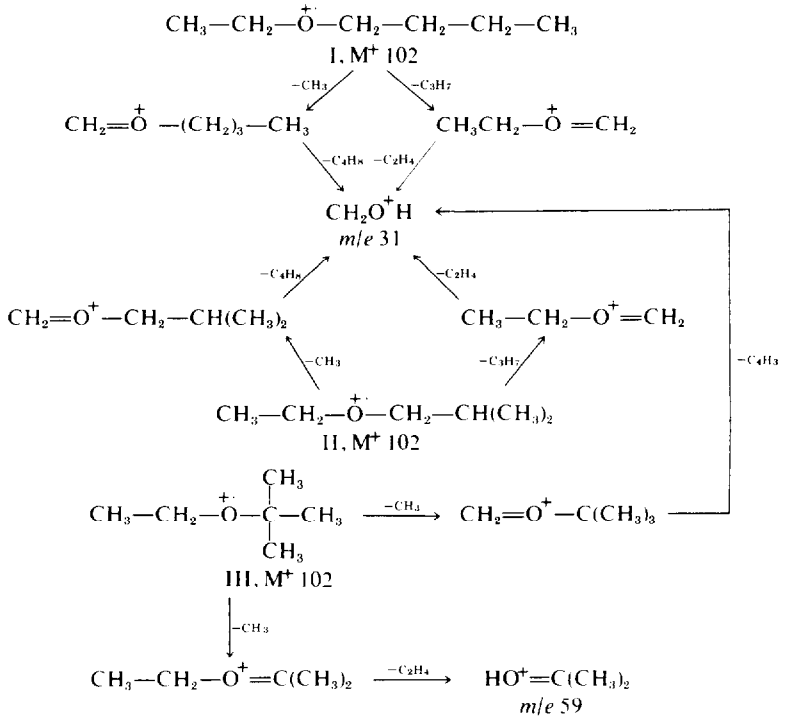
The predicted abbreviated mass spectrum for each of the three candidate structures (read from TEST 11A) is then compared to Fig. 7-3 to determine whether any fundamental inconsistencies exist. Those structures remaining (none were eliminated in the example under scrutiny) are then processed by the SCORING FUNCTION, which ranks them in order of preference. The order depends on the number of peaks considered to be significant in the predicted mass spectrum† and on their estimated relative degrees of significance. For example, ions *a*, *b*, and *e* are assigned degree 3 and

†"QUOTE" is an idiosyncracy of LISP to distinguish a label from the contents of the corresponding list.

‡The three candidate structures represented in DENDRAL dot notation are ethyl *n*-butyl ether, ethyl isobutyl ether (both belonging to the ether 2 subgroup), and ethyl *t*-butyl ether (ether 3 subgroup), respectively. C4*ETHER2!*H10 and C2*ETHER 4A!*H6 correspond to the empirical formula $C_6H_{14}O$ when the compositions (see Scheme 7-B) CH_4O and C_4H_8O of an ether 2 and ether 4A, respectively, are included.

†In the predicted mass spectra the *m/e* value and relative intensity are listed as a dotted pair [e.g., "(57.61)" refers to *m/e* 57 of 61% relative intensity]. No significance should be attached to the relative intensity values as they are calculated from parameters which are at best only crude approximations.

rearrangement ions *c*, *d*, and *f* also have degree 3. It will be observed that the SCORING FUNCTION ranks candidate 3 (ethyl *t*-butyl ether) as its first preference (score of 23). This example received an inflated score relative to the other two structures because of the branching of the *t*-butyl entity. Thus every methyl of this group is available for elimination by α-cleavage (Scheme 7-C) and each of these resulting ions can yield the rearrangement ion of *m/e* 59. Hence the



Scheme 7-C **α-Cleavage** patterns of several ethyl-butyl ethers.

greater the numbers of possible α-cleavages, the more significant peaks and the higher the score of that candidate in the present program. We have deferred the further refinement of the SCORING FUNCTION in favor of an NMR section of Heuristic DENDRAL since this promised to yield a more unambiguous result (see Fig. 7-4).

We frequently found that mass spectrometry alone was insufficient to separate the correct structure from those of three or four other dialkyl ethers but that unequivocal answers could be obtained by incorporating into Heuristic DENDRAL some knowledge of NMR spectroscopy. Thus a new subroutine of the program was applied to all tenable structures passed by the SCORING FUNCTION. It should be noted that the program can profitably use NMR data if it is available but does not require it.

The NMR program accepts two arguments: (1) a list of candidates (from the SCORING FUNCTION) and (2) the NMR spectrum of the unknown compound. For each viable structure an NMR spectrum is predicted.* This is then compared with the unknown's NMR spectrum and the chemical shift information must agree to within ±0.3 ppm. The predicted resonance must display the same multiplicity and integral value as the unknown. If the recorded signal is a multiplet then the predicted NMR spectrum must contain one or more signals within ±0.3 ppm of this chemical shift and the values of the integrals must be compatible. If the signal requirements are not satisfied between the predicted and unknown's NMR spectrum, then the disparity is noted and utilized by the NMR SCORING FUNCTION. For any candidate the score is zero if all the signals in the unknown spectrum were assigned. Otherwise the score is the product of all the integrals of the un-

*The NMR data necessary for the prediction of chemical shifts are stored as correlation tables taken from K. Nakanishi, *Infrared Absorption Spectroscopy*, Holden-Day, San Francisco, Calif., 1962, p. 223. The integral values for a given structure are predicted as the actual number of hydrogens giving rise to each predicted signal. The multiplicity of the predicted signal is determined by the following rules (the term "α-carbon" refers to the carbon atom adjacent to the C-H under discussion): if more than one α-carbon possesses hydrogens M (multiplet); if no α-hydrogens present S (singlet); if one α-hydrogen present D (doublet); if two α-hydrogens present T (triplet); if three α-hydrogens present Q (quartet). No use is currently made of coupling constants or other data (spin decoupling measurements) but it is anticipated that these could be incorporated into the program as required.

assigned signals multiplied by 0.75 for each multiplet. The lower the score, the higher the priority for any structure.

The recorded NMR spectrum (3 hydrogens, triplet at δ 1.09; 9 hydrogens, singlet at δ 1.13; and 2 hydrogens, quartet at δ 3.33) of the unknown compound (ethyl *t*-butyl ether) is already available in the literature (17). It was presented to the program as

((1.09 3 T) (1.13 9 S) (3.33 2 Q))

and output from the program given in Fig. 7-5 appeared at the teletype.†

```
1)
C..C.C.CO.C.C

((Ø . Ø) (29 . 3Ø) (31 . 1ØØ) (57 . 61) (59 . 33) (87 . 66) (1Ø2 . 5))

2)
C..C..CCO.C.C

((Ø . Ø) (29 . 37) (31 . 1ØØ) (57 . 75) (59 . 18) (87 . 81) (1Ø2 . 6))

3)
C....CCCO.C.C

((Ø . Ø)(29 . 8)(31 . 25)(57 . 5)(59 . 75)(87 . 1ØØ)(1Ø2 . 1))

*LIST OF RANKED MOLECULES:

1 #3
  S = 23
  P = ((31 . 3) (87 . 3) (59 . 3) (87 . 3) (59 . 3) (87 . 3) (59 . 3)
      (37 . 2))
  U = NIL

2 #1
  S = 11
  P = ((31 . 3) (59 . 3) (31 . 3) (87 . 2))
  U = NIL

3 #2
  S = 11
  P = ((31 . 3) (59 . 3) (31 . 3) (87 . 2))
  U = NIL

*   1. N MEANS THE FIRST RANKED
      MOLECULE IS THE NTH IN THE
      ORIGINAL NUMBERED LIST ABOVE.
S = THE SCORE (HIGHEST = BEST) BASED ON
      THE NUMBER OF SIGNIFICANT
      PREDICTED PEAKS IN THE ORIGINAL
      GRAPH.
P = THE LIST OF SIGNIFICANT PREDICTED
      PEAKS.
U = THE POSSIBLY SIGNIFICANT PEAKS
      USED TO RESOLVE SCORING TIES (THE
      FEWER IN DOUBT THE BETTER).
DONE
#
```

Figure 7-4.

†The STRING NOTATION used for candidates 1, 2, and 3 is represented in an alternative DENDRAL format in which 1 designates a single bond. These three candidates translate to I, II, and III, respectively.

```
PREDICTED NMR-SPECTRA:
CANDIDATE NUMBER: 1
STRING-NOTATION: O11C1CC1C1C1C
```

DELTA-VALUE	NUMBER OF HYDROGENS	MULTIPLICITY
Ø.9Ø	3	T
1.3Ø	3	T
1.4Ø	2	M
1.9Ø	2	M
3.4Ø	2	T
3.4Ø	2	Q

```
CANDIDATE NUMBER: 2
STRING-NOTATION: O11C1CC1C11CC
```

DELTA-VALUE	NUMBER OF HYDROGENS	MULTIPLICITY
Ø.9Ø	6	D
1.3Ø	3	T
2.ØØ	1	M
3.4Ø	2	D
3.4Ø	2	Q

```
CANDIDATE NUMBER: 3
STRING-NOTATION: O11C1CC111CCC
```

DELTA-VALUE	NUMBER OF HYDROGENS	MULTIPLICITY
1.3Ø	9	S
1.3Ø	3	T
3.4Ø	2	Q

```
LIST OF RANKED MOLECULES:
```

CANDIDATE:	RANK:	NON-ASSIGNED SIGNALS:
3	1	NIL
2	2	((1.1299999 9 S))
1	3	((1.1299999 9 S))

```
DONE
#
```

Figure 7-5

The program predicted chemical shifts for the protons of candidates 1, 2, and 3 according to the values in parentheses in structures I, II, and III. Heuristic DENDRAL correctly identified the unknown from its mass and NMR spectra as ethyl *t*-butyl ether. Table 7-II records other examples in which DENDRAL examined known spectra as "unknown" utilizing solely the MS information or combining it with an NMR spectrum.

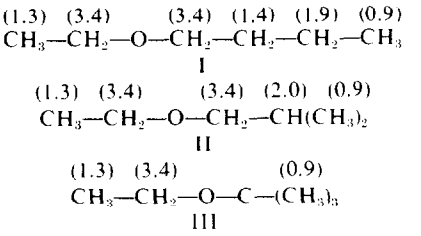
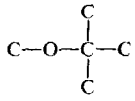
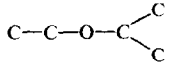
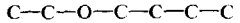
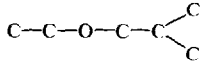
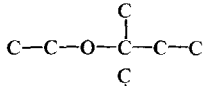
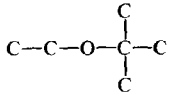
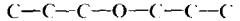
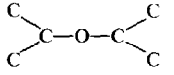
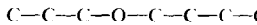
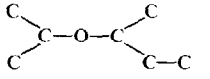
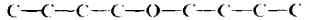
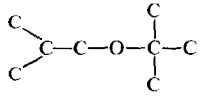
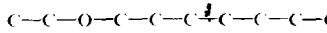
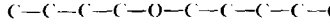
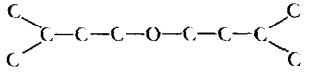


Table 7-11 Heuristic DENDRAL Interpretation of the Mass Spectra[a] of Some Aliphatic Ethers

	Compound	Number of Aliphatic: Isomers	Ethers	Number of Candidates from: Structure Generator	Consistency Check	Ranking of Candidates
1.		14	6	2	2	Correct structure ranked below ethyl *n*-propyl
2.		14	6	4	4	Correct structure ranked first
3.		32	15	2	2	Correct structure tied with ethyl isobutyl
4.		32	15	2	2	Correct structure tied with ethyl *n*-butyl
5.		32	15	6	6	Correct structure tied with *n*-propyl isopropyl
6.		32	15	3	3	Correct structure ranked first[b]
7.		32	15	1	1	Correct structure ranked first[b]
8.		32	15	10	10	Correct structure ranked first[b]
9.		72	33	2	2	Correct structure tied with *n*-propyl isobutyl
10.		72	33	1	1	Correct structure ranked first
11		171	82	3	3	Correct structure tied with *n*-butyl isobutyl and diisobutyl
12.		171	82	15	15	Di-*t*-butyl ranked first Correct structure tied for second with isopropyl isoamyl
13.		405	194	17	13	Correct structure tied with 12 other ethyl ethers
14.		405	194	8	8	Correct structured tied with 7 other (C_4)—O—(C_5) ethers
15.		989	482	10	10	Correct structure tied with 9 others (C_5)—O—(C_5) ethers
16.		989	482	10	10	Correct structure ranked first[b]

[a] The mass spectra used as "unknown" were taken from the literature (18).

[b] NMR spectra correctly differentiated the correct structure from the other candidates. Without the NMR input data the correct structure tied for first, together with the number of candidates listed under consistency check.

Although we recognize that the assignment of the correct structure to an unknown aliphatic ether is a fairly simple problem, it nonetheless represents a starting point for demonstrating the potential power inherent in computer interpretation of experimental data. Even when no unambiguous answers can be obtained, it is impressive to note that the number of possible candidates is reduced drastically (e.g., 10 candidates out of 989 theoretical possibilities in examples 15 and 16 in Table 7-II). In the case of mass spectra taken directly from GC effluents, the program would not be able to utilize NMR input data. Thus multiple solutions would be possible for a particular problem. However, as stated previously, a significant degree of truncation considering all possible aliphatic ethers would be achieved. Clearly one can program other physical data (for instance, infrared and ultraviolet spectral parameters) to supplement the MS and NMR data currently used. With added experimental data and sophisticated programming the computer should be able to solve more complex problems and it is to this end that future research in our laboratories is being directed.

REFERENCES

1. Duffield, A. M., Robertson, A. V., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A., and Lederberg, J., *J. Amer. Chem. Soc.* **91**, 2977 (1969).
2. Schroll, G., Duffield, A. M., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A., and Lederberg, J., *J. Amer. Chem. Soc.* **91**, 7440 (1969).
3. Pettersson, B., and Ryhage, R., *Ark. Kemi.* **26**, 293 (1967).
4. Crawford, L. R., and Morrison, J. D., *Anal. Chem.* **40**, 1469 (1968), **41**; 994 (1969).
5. Venkataraghavan, R., McLafferty, F. W., and VanLear, G. E., *Org. Mass Spectrom.* **2**, (1) (1969).
6. Sasoki, S., Abe, H., and Ouki, T., *Anal. Chem.* **40**, 2220 (1968).
7. Biemann, K., and Fennessey, P. V., *Abstr. papers, 14th Ann. Conf. Mass Spectrom. Dallas, Tex.*, 322 (1966).
8. Mandelbaum, A., Fennessey, P. V., and Biemann, K., *Abstr. papers, 15th Ann. Conf. Mass Spectrom. Denver, Colo.*, 111 (1967).
9. Buchanan, B. G., Sutherland, G. L., and Feigenbaum, E. A., "HEURISTIC DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry," *in* Meltzer, B., and Michie, D. (Eds.), *Machine Intelligence 4*, Edinburgh University Press, 1969 (also Stanford Artificial Intelligence Project Memo No. 62).
10. Sutherland, G., "HEURISTIC DENDRAL: A Family of LISP Programs," to appear in Bobrow, D. (Ed.), *LISP Applications* (also Stanford Artificial Intelligence Project Memo No. 80).
11. Lederberg, J., and Feigenbaum, E. A., "Mechanization of Inductive Inference in Organic Chemistry," *in* Kleinmuntz, B. (Ed.), *Formal Representations for Human Judgment*, Wiley, 1968 (also Stanford Artificial Intelligence Project Memo No. 54).
12. Buchanan, B. G., Sutherland, G. L., and Feigenbaum, E. A., "Rediscovering Some Problems of Artificial intelligence in the Context of Organic Chemistry," *in* Meltzer, B., and Michie, D. (Eds.), *Machine Intelligence 5*, Edinburgh University Press (1970) (also Stanford Artificial Intelligence Project Memo No. 99).
13. Lederberg, J., "DENDRAL-64—A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs," technical report to NASA, CR 57029 (1964); also available from the author and summarized in (11), (14), and (15).
14. Lederberg, J., Sutherland, G. L., Buchanan, B. G., Feigenbaum, E. A., Robertson, A. V., Duffield, A. M., and Djerassi, C., *J. Amer. Chem. Soc.* **91**, 11 (1969).
15. Lederberg, J., "Topology of Molecules" *The Mathematical Sciences, A Collection of Essays*, M.I.T. Press, Cambridge, Mass., 1969, p. 37.
16. Paquette, L. A., Kirschner, S., and Malpass, J. R., *J. Amer. Chem. Soc.* **92**, 4330 (1970).
17. Brune, H. A., and Schulte, D., *Chem. Ber.* **100**, 3438 (1967).
18. McLafferty, F. W., *Anal. Chem.* **29**, 1782 (1957).