Technical Report No. IRL 1158

CYTOCHEMICAL STUDIES OF PLANETARY MICROORGANISMS EXPLORATIONS IN EXOBIOLOGY

Status Report Covering Period January 1, 1972 to December 31, 1972
For
National Aeronautics and Space Administration
Grant NGR-05-020-004

Instrumentation Research Laboratory, Department of Genetics
Stanford University School of Medicine
Stanford, California 94305

Report to the National Aeronautics and Space Administration

"Cytochemical Studies of Planetary Microorganisms - Explorations in Exobiology

NGR 05-020-004

Summary Report Covering Period January 1, 1972 to December 31, 1972

Instrumentation Research Laboratory, Department of Genetics

Stanford University School of Medicine

Stanford, California

Joshua Lederberg
Principal Investigator

Elliott C. Levinthal, Director
Instrumentation Research Laboratory

TABLE OF CONTENTS

A. INTRODUCTION

This report covers the activities of the Instrumentation Research Laboratory during the calendar year January 1, 1972 to December 31, 1972.

The main support of the IRL activities during this period continued to be the NASA grant NGR 05-020-004. Some funds have come from other grants, other agencies, and in some cases private institutions. This report includes all the activities of the laboratory which relate to or have benefited from NASA support regardless of whether or not they were primarily supported by this NASA grant.

The Dendral project receives direct support from NIH grant GM 00612 and indirectly from the ACME grant NIH RR 00311 and Professor Feigenbaum's ARPA grant SD-183.

Our work in the area of cell separation was supported in this period by NIH grant GM 17367.

The efforts on image processing for the 1971 Mars Mariner are a joint effort with the Artificial Intelligence Laboratory of the Computer Science Department under Professor John McCarthy. It receives support from NASA grant NGR-05-020-508 and JPL Contract 952489. Our work on Viking lander camera imagery receives support under Langley contract NAS 1-9682.

During the last six months of this report period we have had extensive discussions with the Manned Spacecraft Center at Houston on ways to exploit our experience in analytical methodology using gas chromatography and mass spectroscopy for the purpose of improved physiological monitoring of astronauts. In particular the work we have been doing on the analysis of the metabolic constituents of urine has led to a specific proposal. This proposal has been funded (NGR-05-020-632) for one year starting May 15, 1973.

We have continued to collaborate with Professor Marvin Chodorow of the Applied Physics Department on a preliminary evaluation of the application of microwave acoustic scattering to problems of cell discrimination and detection.

The general areas of the Program Resume, Part B of the report, are:

I. Chlorination of DNA Bases

II. Mass Fragmentography

III. Mass Spectrometry

IV. Urine Analysis

V. Analysis of Natural Products by Mass Spectrometry

VI. Computer Aided Research (DENDRAL)

VII. Cell Separation

VIII. Mariner Mars 1971 Orbiter Photography

IX. Viking Lander Imagery

B. PROGRAM RESUME

I. Chlorination of DNA Bases

The study of the action of aqueous hypochlorous acid under physiological conditions upon the bases present in DNA has been continued. Under these conditions thymine yields 5-chloro-6-hydroxy-5, 6-dihydrothymine while with one and two equivalents of reagent 6-methyluracil is converted into 5-chloro-6-methyluracil and 5,5-dichloro-6-hydroxy-5, 6-dihydro-6-methyluracil respectively. 1,3-Dimethyluracil reacts with both one and two equivalents of hypochlorous acid to form 5,5-dichloro-6-hydroxy-5,6-dihydro-1,3-dimethyluracil.

DNA bases containing the purine ring system, for instance guanine and adenine, react with hypochlorous acid to yield parabanic acid. Xanthine also yielded this product while the N-methylated purines caffeine and theophiline afforded N-methyl parabanic acid. Mechanistically these observations suggest that parabanic acid is derived from the six-membered ring of the purine system.

Our chlorination studies were extended to the nucleosides, cytidine and deoxycytidine and the nucleotide cytidine-5'-monophosphate. The products from all three compounds were the corresponding 4-N-chloro compound. This was determined from physical measurements including mass spectrometry and especially from their NMR spectra which

exhibited a characteristic 0.42-0.52 ppm diamagnetic shift for the chemical shift of the C-6 proton.

In addition the following papers on chlorination by hypochlorous acid have been either published or accepted for publication:

> Chlorination Studies. I. The Reaction of Aqueous Hypochlorous Acid with Cytosine. By W. Patton, V. Bacon, A. M. Duffield, B. Halpern, Y. Hoyano, W. Pereira and J. Lederberg. Biochem. Biophys. Res. Commun., 48, 880 (1972).
>
> Chlorination Studies. II. The Reaction of Aqueous Hypochlorous Acid with α-Amino Acids and Dipeptides. By. W. E. Pereira, Y. Hoyano, R. Summons, V. A. Bacon and A. M. Duffield. Biophys. Biochem. Acta (in press).

II. Mass Fragmentography

Methods for the detection and quantitation of amino acids in varied environments by mass spectrometry is being pursued. As the approach used involves mass fragmentography we require the preparation of a suitable derivative which must enhance the mass spectral identification and in addition must possess suitable gas chromatographic properties. The derivative of choice for these purposes would appear to be the amino acid O-butyl ester, N-trifluoroacetate. Quantitation of the amino acid levels present in crude soil extracts has been achieved by the addition of an internal standard consisting of a known quantity of the deuterated amino acids to be analyzed. Using the quadrupole-mass spectrometer-computer system in the technique of mass fragmentography we have been able to simultaneously quantitate up to ten of the amino acids present in soil. The experimental details of this novel method await publication, see "The Simultaneous Quantitation of Ten Amino Acids in Soil Extracts by Mass Fragmentograph" by W. E. Pereira, Y. Hoyano, W. E. Reynolds, R. E. Summons and A. M. Duffield, Anal. Biochem., in press.

We have also extended this quantitative technique to the measurement of phenylalanine in plasma (i.e. phenylketonuria). See "The Determination of Phenylalanine in Serum by Mass Fragmentography" by W. E. Pereira, V. A. Bacon, Y. Hoyano, R. Summons and A. M. Duffield, Clinical Biochem., in press.

III. Mass Spectrometry

Phenothiazines represent a class of drugs commonly prescribed in medicine as tranquilizers. Although the mass spectra of this group of compounds have been investigated in several laboratories no definitive study of their mass spectral decomposition processes using deuterium labeling has appeared in the literature. Suitable deuterated derivatives of promazine and promazine sulfoxide were prepared and from their mass spectra the types of rearrangements occurring in the mass spectral fragmentation processes of promazine and its sulphoxide were identified. See "A Study of the Electron Impact Fragmentation of Promazine Sulphoxide and Promazine using Specifically Dueterated Analogs," by M. D. Solomon, R. Summons, W. Pereira and A. M. Duffield, Austral. J. Chem., 26, 325 (1973).

IV. Urine Analysis

Work is progressing on the organic chemical constituents of the urine of premature babies hospitalized in the Pediatrics Ward of the Stanford University Medical Center. Premature infants were selected for this study because they are on strict diets and their urinary metabolites should reflect body metabolites rather than food artifacts. Each urine specimen is processed for their free acid and amino acid constituents and these assays are repeated following hydrolysis of the urine. The complex series of compounds contained in each of the four fractions is separated by gas chromatography and the constituents of each chromatographic peak identified (where possible) by their mass spectra signatures. Currently our system can identify organic compounds present in derivatized extracts below the microgram level.

As an example of the application of GC-MS to biomedical problems we can cite preliminary studies on approximately 80 urine samples from a total of 11 premature or "small for gestational age" infants. This project was undertaken to investigate the phenomenon of late metabolic acidosis. This condition is characterized by low blood pH levels and poor weight gain and, as distinct from respiratory acidosis, occurs after the 2nd day of life. Its incidence is higher in infants whose birthweight is less than 1750 g (one study shows 92% incidence for these children) than in infants with birthweight greater than 1750 g (28%).

Of the 11 patients studied we were able to observe 6 closely and continuously for periods ranging from 6 to 8 weeks from day 3 of life. Three of these infants had birthweights below 1000 g and the other three were born weighing less than 1500 g. Of the 6, five showed symptoms corresponding to late metabolic acidosis and the other showed normal and even development. The five infants showing the acidosis all excreted very large amounts of p-hydroxyphenyllactic acid together with smaller amounts of p-hydroxyphenylpyruvic acid and p-hydroxyphenylacetic acid. After reaching a peak, the occurrence of these compounds in the urine gradually diminished and were almost completely absent at the time blood pH and weight gain had returned to normal. The infant who did not show symptoms of acidosis only excreted minute amounts of these compounds during the period of observation.

The occurrence of large amounts of these compounds in the urine indicates a temporary defect in phenylalanine - tyrosine metabolism and dietary factors such as protein and vitamin intake can be shown to affect the incidence and the severity of the condition. It is hoped that further studies will result in a clearer picture of relationships between the condition and diet and hence lead to a reduction in its occurrence.

V. Analysis of Natural Products by Mass Spectrometry

During the past year research has continued on the structural analysis by mass spectrometry of natural products isolated from plant, animal and marine sources. A list of publications resulting from this experimentation follows:

E. Ali, P. P. Gosh Dastidar, S. C. Pakrashi, L. J. Durham and A. M. Duffield. "Studies on Indian Medicine Plants - XXVII Sequiterpene Lactones of Enhydra Fluctuans Lour. Structures of Enhydrin, Fluctuanin and Fluctuadin." Tetrahedron 28, 2285 (1972).

A. M. Duffield and O. Buchardt. "Thermal Fragmentation of Quinoline and Isoquinoline N-Oxides in the Ion Source of a Mass Spectrometer." Acta Chem. Scand., 26, 2423 (1972).

A. N. H. Yeo and C. Djerassi. "Mass Spectrometry in Structural and Stereochemical Problems. CCX. Evidence for Transition States of Different Ring Sizes in the Loss of C_4H_8 from Phenyl n-Butyl Ether in the Mass Spectrometer." J. Am. Chem. Soc. 94, 482 (1972).

B. A. Brady, W. I. O'Sullivan and A. M. Duffield. "The Electron-Impact Promoted Fragmentation of Aurone Epoxides." Organic Mass Spectrometry 6, 199 (1972).

P. Sedmera, A. Klasek, A. M. Duffield and F. Santavy. "Pyrrolizidine Alkaloids. XIX. Structure of the Alkaloid Eruoifoline." Coll. Czech. Chem. Commun., 37, 4112 (1972).

P. Perros, J. P. Morizur, J. Kossanyi and A. M. Duffield. "Spectrometrie de Masse VIII. Elimination d'can Induite par Impact Electronique dans le Tetrahydro-1,2,3,4-Naphtal-ene-diol-1,2. Org. Mass Spectrom. 7, 357 (1973).

Y. M. Sheikh, R. J. Liedtke, A. M. Duffield and C. Djerassi. "Mass Spectrometry in Structural and Stereochemical Problems CCXVII. Electron Impact Promoted Fragmentation of O-Methyl Oximes of some α,β-Unsaturated Ketones and Methyl Substituted Cyclohexanones." Canad. Jour. Chem. 50, 2776 (1972).

VI. Dendral

A. Computer Science

Recent Progress in the Dendral Artificial Intelligence Project is summarized in the following report.

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 2 3 4 5 6 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.

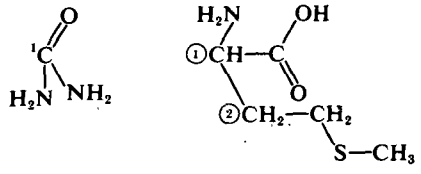
Nevertheless, a few trivial abbreviations have been incorporated into the output conversion routines of DENDRAL and are modified according to the taste of each user. They include such constructions as —CH2OH, —CHO, —COOH, and (for *n*-alkyl) forms like —C3H7 for the corresponding DENDRAL codes: .CH2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

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	5 5	1
3	1 2 2 3 / 4 3 4 3 / 5 5 4 4	4
4	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



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.

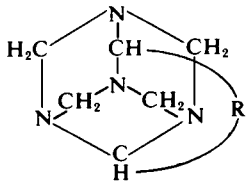
Table 7-I Counting the Different Arrangements of Compounds of Carbon and Hydrogen Containing No Double or Triple Bonds and No Rings (general formula $C_n H_{2n+2}$).

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

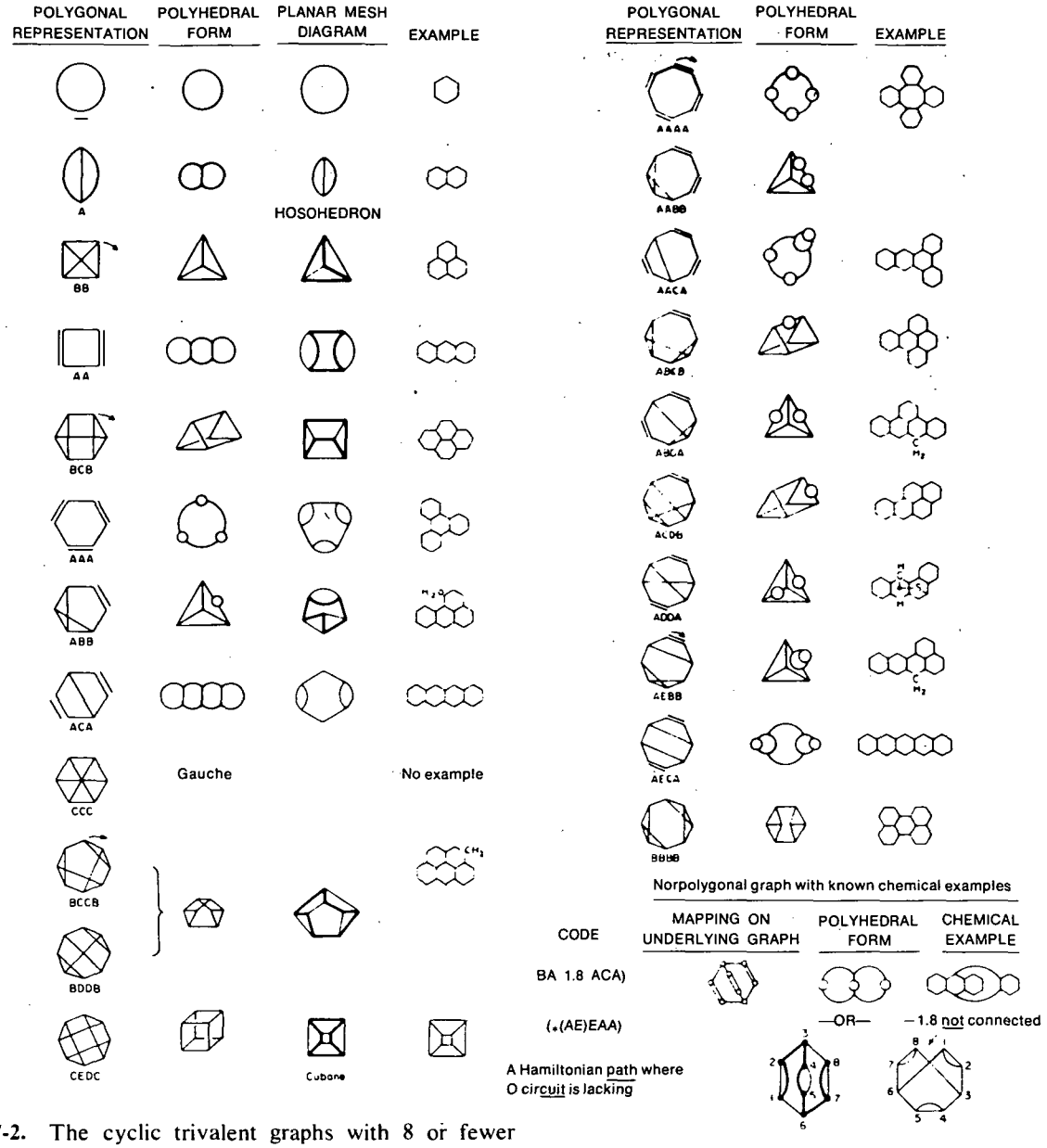


Figure 7-2. The cyclic trivalent graphs with 8 or fewer nodes. Up to 6 nodes, these all have Hamilton circuits but may also be represented in other ways. In a few examples, the circuits are drawn with emphasis on planar map representations. Complete tables of chord lists like those shown under the circuit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

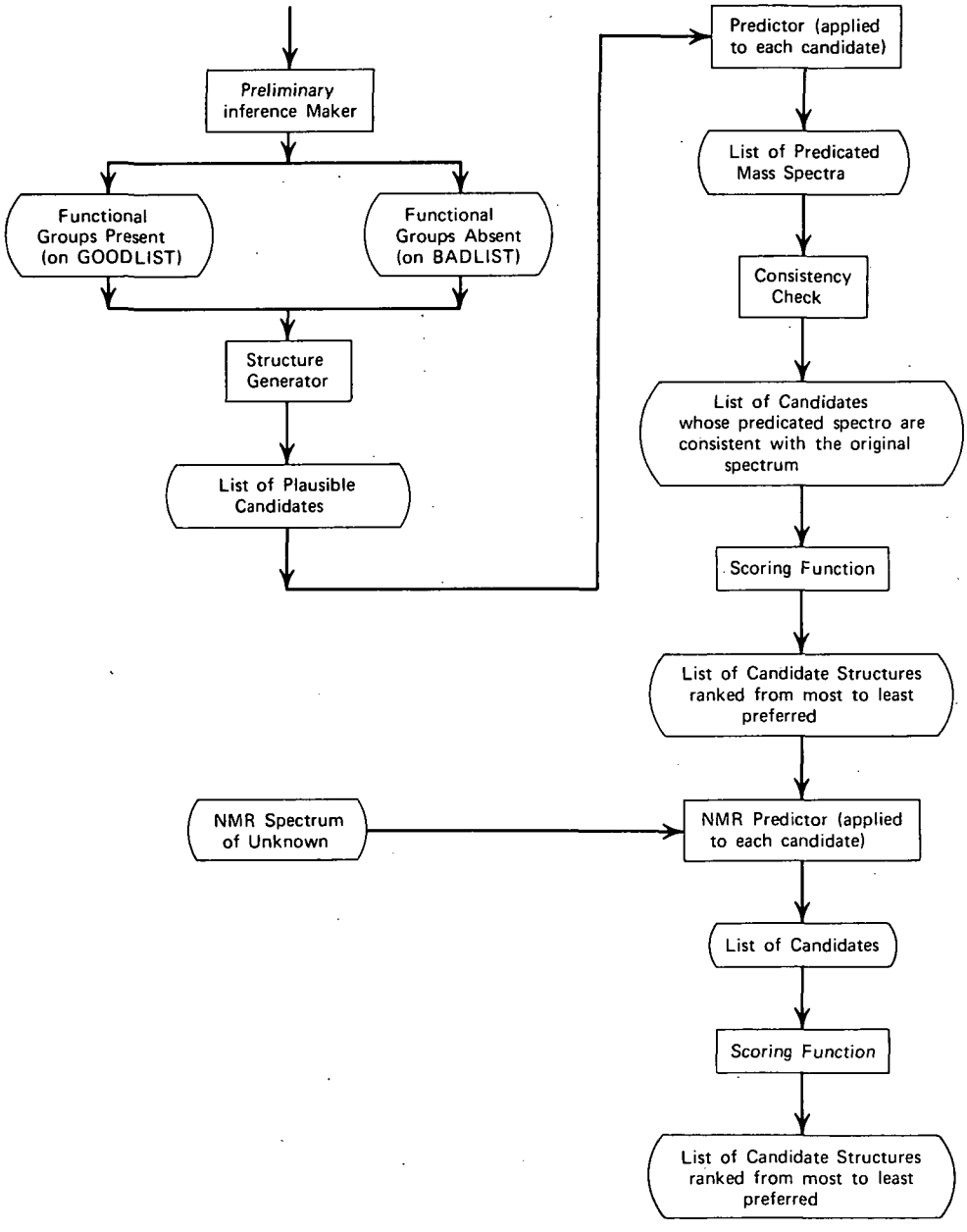
3. Separating the theory from the routine which uses it facilitates changing the theory to improve it, on the one hand, or to experiment with variations of it, on the other. Although scattering the theory in the program's LISP code increases running efficiency, it seems more desirable, at this point, 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

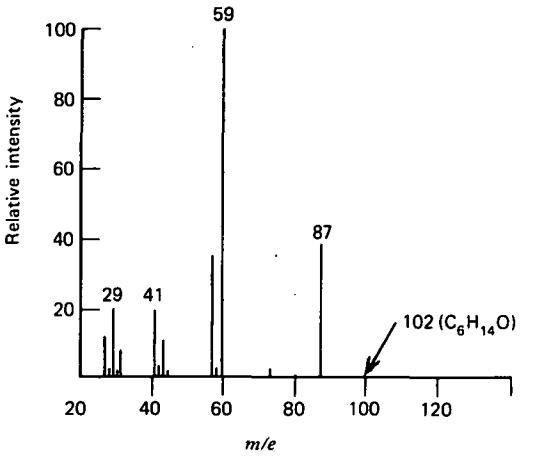
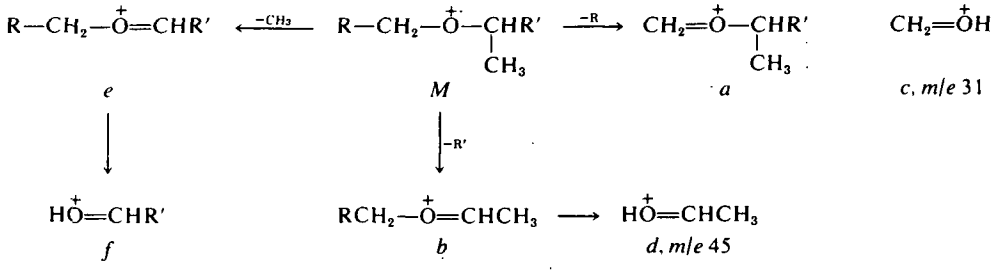


Figure 7-3. Mass spectrum of unknown aliphatic ether of composition $C_6H_{14}O$.

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†:

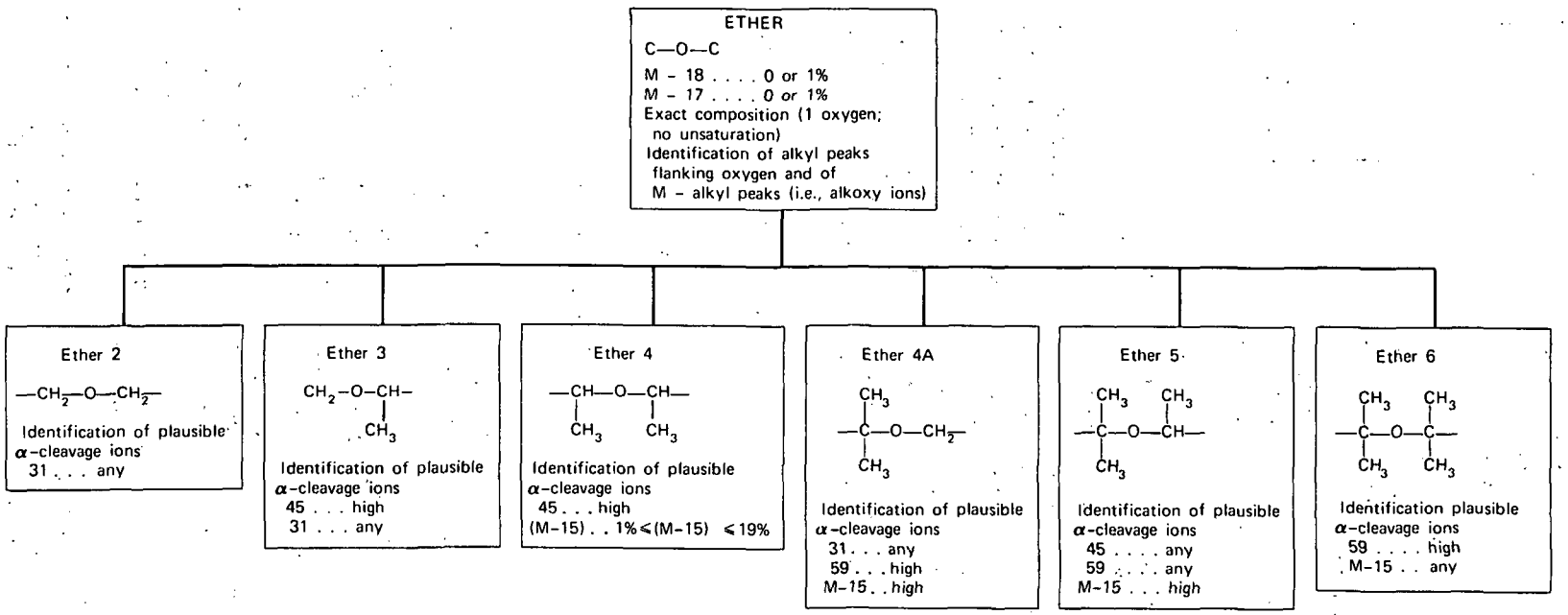
```
*(INFER (QUOTE C6H140) S:ETH-TERT-BUT
        (QUOTE TEST!!))
```

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.

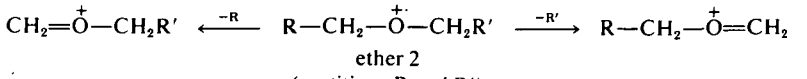


Scheme 7-B. Rules for ether identification

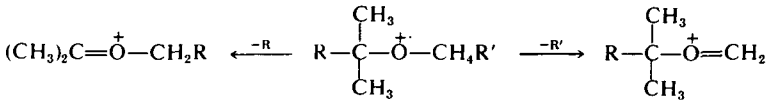
of Fig. 3 by the PRELIMINARY INFERENCE MAKER the computer responds with

```
*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)
```

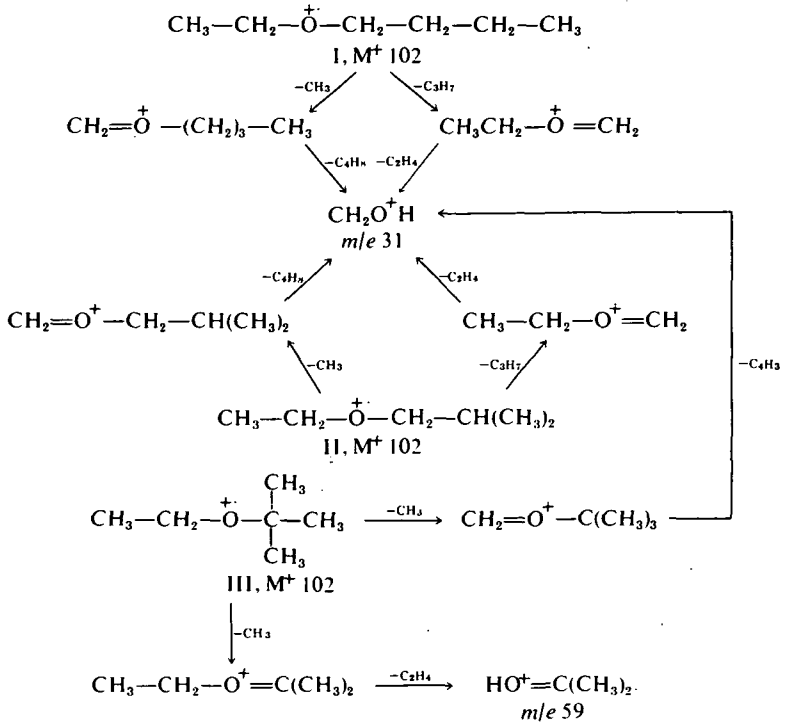
The predicted abbreviated mass spectrum for each of the three candidate structures (read from TEST 11A) is then compared to Fig. 7-3 to determine whether any fundamental inconsistencies exist. Those structures remaining (none were eliminated in the example under scrutiny) are then processed by the SCORING 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: O1ICICCICICIC

                    NUMBER OF
DELTA-VALUE         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: O1ICICCICIICC

                    NUMBER OF
DELTA-VALUE         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: O1ICICCIIICCC

                    NUMBER OF
DELTA-VALUE         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.

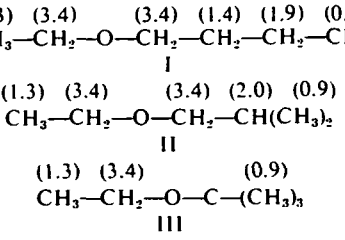
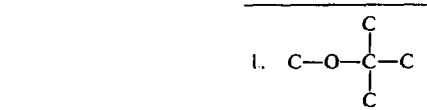
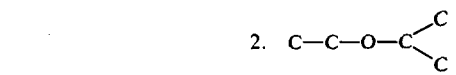
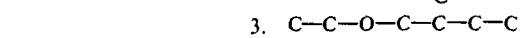
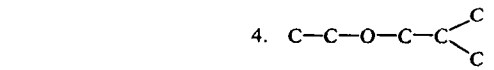
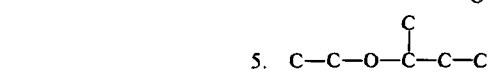
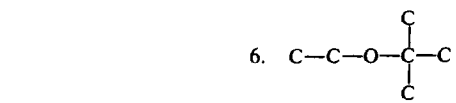
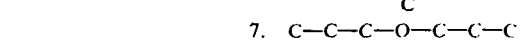
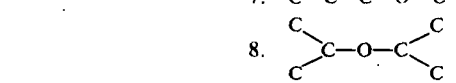
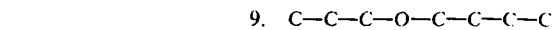
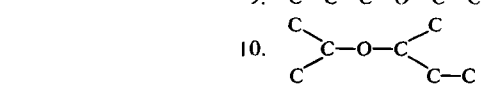
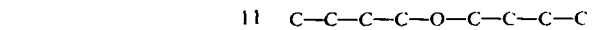
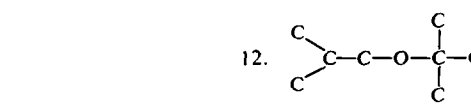
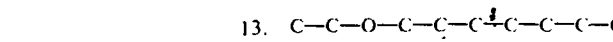
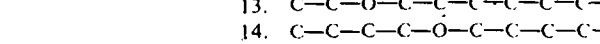
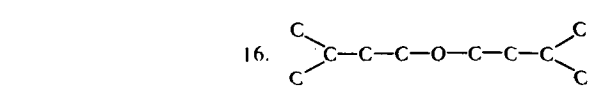


Table 7-II Heuristic DENDRAL Interpretation of the Mass Spectra[a] of Some Aliphatic Ethers

	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).

B. Computer Aided Research Instrumentation

1. Objectives and Background

The objectives of this work are to develop and demonstrate computer techniques for automating complex laboratory instrumentation and procedures. Within these objectives, three main areas of investigation are of interest including: 1) methodologies for computer modeling and control of instrument performance, 2) the integration of intelligent data analysis and interpretation programs into closed loop systems to optimize and economize problem solutions, and 3) the organization of computing and hardware resources needed to implement automated systems. We have chosen gas chromatography/mass spectrometry (GC/MS) as a specific environment in which to develop and test ideas. This choice is based on the importance of GC/MS to on-going research in medicine, biochemistry, and computer science described elsewhere in this report, as well as a genuine need for system automation because of severe data volume and analysis complexity problems.

2. Progress

Since the last report, we have made progress in developing aspects of computer-aided GC/MS instrumentation as summarized below.

a) MASS SPECTROMETER DATA SYSTEM AUTOMATION

Concentrating initially on the MAT-711 high resolution mass spectrometer, we have made progress toward a reliable, automated data acquisition and reduction system for scanned low and high resolution spectra. This system is largely failsafe and requires no operator support or intervention in the calculation procedures. Output and warnings to the operator are provided on a CRT adjacent to the mass spectrometer. The system contains many interactive features which permit the operator to examine selected features of the data at his leisure. The feedback currently provided to the operator to assist in instrument set-up and operation can just as well be routed to hardware control elements for these functions thereby allowing computer maintenance of optimum instrument performance.

Progress in this area is an integration of our efforts in hardware and software improvements:

HARDWARE - The basic system consists of the mass spectrometer interfaced to a PDP-11/20 computer for data acquisition, pre-filtering, and time buffering into the ACME time-shared 360/50. The more complex aspects of data reduction are done in the 360/50 since the PDP-11 has limited memory and arithmetic capabilities. New interfaces for mass spectrometer operation and control have been developed. The interfaces can handle (through an analog

multiplexer) several analog inputs and outputs which require that the PDP-11 computer be relatively near the mass spectrometer. We now have the capability for the following kinds of operation through the new interfaces.

i) Computer selection of digitization rate

ii) Computer selection of data path (interrupt mode or direct memory access (DMA)

iii) Direct memory access for faster operation in the data acquisition mode.

iv) Computer selection of analog input and output channels.

v) Sensing of several analog channels through a multiplexer (e.g., ion signal, total ion current).

vi) Magnet scan control. This control can be exercised manually or set by the computer. It controls both time of scan and flyback time. Coupled with selection of scan rate, any desired mass range can be scanned at any desired scan rate.

vii) The computer can monitor the mass spectrometer's mass marker output as additional information which will be used to effect calibration.

Another development has been a signal conditioner for

the ion signal which incorporates a box-type integrator to sum the ion signal between A/D converter readings. This modification makes successive intensity readings independent of each other because the integrator is reset after each reading. It also provides for low pass filtering the ion current signal with a bandwidth automatically adjusted correctly for different sampling rates and hence lessens intensity measurement uncertainties caused by external noises.

SOFTWARE - Automatic instrument calibration and data reduction programs have been developed to a high degree of sophistication. We can now accurately model the behavior of the MAT-711 mass spectrometer over a variety of scan rates and resolving powers. Our instrument diagnostic routines are depended upon by the spectrometer operator to indicate successful operation or to help point to instrument malfunctions or set-up errors. Some features of these programs are described below.

i) Data Acquisition. Programs have been written which permit acquisition of peak profile data at high data rates using the PDP-11 as an intermediate data filter and buffer store between the mass spectrometer and ACME. This allows data acquisition to proceed even under the time constraints of the time-sharing system. Storage of peak profiles rather than all data collected has greatly reduced the storage

requirements of the program and saves time as the background data (below threshold) are removed in real-time. An automatic thresholding program is in operation which statistically evaluates background noise and thresholds subsequent data accordingly. Amplifier drift can thus be compensated. We have developed some theoretical models of the data acquisition process which suggest that high data acquisition rates are not necessary to maintain the integrity of the data. Demonstration of this fact with actual data has helped relieve the burden of high data rates on the computer system, particularly as imposed by GC/MS operation, and permits more data reduction to be accomplished in real-time or alternatively reduces the required data acquisition computer capacity.

ii) Instrument Evaluation. A high resolution mass spectrometer operating in a dynamic scanning mode is a complex instrument and many things can go wrong which are difficult for the operator to detect in real-time. In order for the computer to assist in maintaining data quality, it must have a model of spectrometer operation on the basis of which data quality can be assessed and processing suitably adapted as well as instrument performance optimized. We have developed a program which monitors the state of the mass spectrometer. This preliminary program checks the following items:

1) Data acquisition parameters such as scan range and time constants, background threshold, a dynamic peak model to determine resolution and threshold acceptance levels for peak width and intensity, the number of peaks collected, and data storage utilization statistics.

2) Calibration of the mass/time relation to be used as a model for subsequent spectra, output of the mass range over which the scale is calibrated, calibration peaks missed, if any, and a graph of extrapolation error versus mass. Any irregularities in this output point to scan problems.

3) The dynamic resolution versus mass is determined and output as a graph. This allows the operator to adjust to more constant resolution over the mass range.

iii) Data Reduction. A program has been written which allows automatic reduction of high resolution data based on the results of the prior instrument evaluation data. Conversion of peak positions in time to the corresponding mass values is effected by mapping each spectrum into the calibration model developed previously. The interpolation algorithm between reference calibration points incorporates a quadratically varying exponential time constant to account for the second order character of a magnet discharging through a resistance and a capacitance. It also takes into account a mass offset at infinite time which affects

accuracy in determining low masses and which results from residual magnetization at the end of a scan.

Perfluorokerosene (PFK) peaks, introduced into high resolution mass spectra for internal mass calibration, are distinguished from unknown peaks by a pattern recognition algorithm which compares the relationships between sequences of reference peaks in the calibration run with the set of possible corresponding sequences in the sample run. The candidate sequence is selected which best approximates calibrated performance within constraints of internally consistent scan model variations. This approach minimizes the need for selection criteria such as greatest negative mass defect for reference peaks, the validity of which cannot be guaranteed. Excellent performance results from using sequences containing 10 reference peaks.

Unresolved peaks are separated by a new analytical algorithm, the operation of which is based on a calculated model peak derived from known singlet peaks rather than the assumption of a particular parametric shape (e.g., triangular, Gaussian, etc.) This algorithm provides an effective increase in system resolution by a factor of three thereby effectively increasing system sensitivity. By measuring and comparing successive moments of the sample and model peaks, a series of hypotheses are tested to establish the multiplicity of the peak, minimizing computing

requirements for the usually encountered simple peaks. Analytic expressions for the amplitudes and positions of component peaks have been derived in the doublet case in terms of the first four moments of the peak complex. This eliminates time consuming iteration procedures for this important multiplet case. Iteration is still required for more complex multiplets.

Elemental compositions are calculated from high resolution mass values with a new, efficient table look-up algorithm developed by Lederberg (ref. 1) and appended herewith.

Future work will extend these ideas to a system for the acquisition of selected metastable information as well as to include the quadrupole system used in the routine low resolution clinical work.

b) GAS CHROMATOGRAPHY/HIGH RESOLUTION MASS SPECTROMETRY

We have recently verified the feasibility of combined gas chromatography/ high resolution mass spectrometry (GC/HRMS). Using the programs described above we can acquire selected scans and reduce them automatically, although the procedures are slow compared to "real-time" due to the limitations of the time-shared ACME facility. We have recorded sufficient spectra of standard compounds to show that the system is performing well. A typical experiment

which illustrates some of the parameters involved was the following. A mixture (approximately 1 microgram/ component) of methyl palmitate and methyl stearate was analyzed by GC under conditions such that the GC peaks were well separated and of approximately 25 sec. duration. The mass spectrometer was scanned at a rate of 10.5 sec/decade, and a resolving power of 5000. The resulting mass spectra displayed peaks over a dynamic range of 100 to 1 and were automatically reduced to masses and elemental compositions without difficulty. Mass measurement accuracy appears to be 10 ppm over this dynamic range. A more definitive study of mass measurement accuracy will be carried out shortly to accurately determine the performance of the system.

We have begun to exercise the GC/HRMS system on urine fractions containing significant components whose structures have not been elucidated on the basis of low resolution spectra alone. Whereas more work is required to establish system performance capabilities, two things have become clear: 1) GC/HRMS will be a useful analytical adjunct to our low resolution GC/MS clinical studies to assist in the identification of significant components whose structures are not elucidated on the basis of low resolution spectra alone, and 2) the sensitivity of the present system limits analysis to relatively intense GC peaks. This sensitivity limitation is inherent in scanning instruments where one gives up a factor of 20-50 in sensitivity over photographic

image plane systems in return for on-line data read-out. This limitation may be relieved by using television read-out systems in conjunction with extended channeltron detector arrays as has been proposed by researchers at the Jet Propulsion Laboratory. The development of such a sensor system is beyond the current scope of our effort. We can nevertheless make progress in applying GC/HRMS techniques to accessible effluent peaks and can adapt the improved sensor capability when available.

Recent experiments in operation of the mass spectrometer in conjunction with the gas chromatograph have also shown that the present ACME computer facility cannot provide the rapid service required to acquire repetitive scans at either high or low resolving powers. We can, however, acquire scans on a periodic basis, meaning most GC peaks in a run can be scanned once at high resolving power. We are presently implementing a disk on the PDP-11 to act as a temporary data buffer between the mass spectrometer and ACME. This disk will allow acquisition of repetitive scans, while data reduction must be deferred to completion of the GC run.

c) AUTOMATED GC/MS DATA REDUCTION

The application of GC/MS techniques to clinical problems has made clear the need for automating the analysis of the results of a GC/MS experiment. Previous paragraphs

dealt with the problems of reducing raw data in preparation for analysis. At this point the data must be analyzed with a minimum of human interaction in terms of locating and identifying specific constituents of the GC effluent. The subsequent problem of identification is addressed by the library search and DENDRAL mass spectrum interpretation programs. The problem of locating effluent components in the GC/MS output involves extracting from the approximately 700 spectra collected during a GC run, the 50 or so representing components of the body fluid sample. The raw spectra are in part contaminated with background "column bleed" and in part composited with adjacent constituent spectra unresolved by the GC.

We have begun to develop a solution to this problem with very promising results. By using a unique disk oriented matrix transposition algorithm developed for image processing applications, we can rotate the entire array of 700 spectra by 500 mass samples per spectrum to gain convenient access to the "mass fragmentogram" form of the data. This form of the data, displayed at a few selected mass values, has been used at Stanford, MIT, and elsewhere for some time to evaluate the GC effluent profile as seen from these masses. Mass fragmentograms have the important property of displaying much higher resolution in localizing GC effluent constituents. Thus by transposing the raw data to the mass chromatogram domain we can systematically

analyze these data for baselines, peak positions, and amplitudes, and thus derive idealized mass spectra for the constituent materials free from background contamination and influences of adjacent GC peaks unresolved in the overall gas chromatogram. These spectra can then be analyzed by library search techniques or first principles as necessary.

The results of this work can also lead to reliable prescreening analysis of GC traces alone by having available a detailed list of GC effluent positions and expected amplitudes for say a urine fraction. By dynamically determining peak shape parameters for detected GC singlet peaks, interpretation of more complex peaks can be made to determine if unexpected constituents or abnormal amounts of expected constituents are present.

d) CLOSED-LOOP INSTRUMENT CONTROL

In the long term, it would be possible for the data interpretation software to direct the acquisition of data in order to remove ambiguities from interpretation procedures and to optimize system efficiency. The achievement of this goal is a long way off but we feel the above developments along with progress in the computer interpretation of mass spectra (DENDRAL) represent important preliminary steps toward closed-loop control.

The task of collecting different types of mass spectral

information (e.g., high resolution spectra, low ionizing voltage spectra and selected metastable information) under closed loop control during a GC/MS experiment is difficult and may not be realizable with current technology. We are studying this problem in a manner which will allow the system to be used for important research problems (e.g., routine analysis of urine fractions without fully closed loop control) while aspects of instrument control strategy are developed in an incremental fashion.

The essence of this approach is to develop a multi (two or three)-pass system which permits collection of one type of data (e.g., high resolution mass spectra) during the first GC/MS analysis. Processing of these data by DENDRAL will reveal what additional data are necessary on specific GC peaks during a subsequent GC/MS run to uniquely solve the structure or at least to reduce the number of candidate structures. This simulated closed-loop procedure will demonstrate the ability of DENDRAL type programs to examine data, determine solutions and propose additional strategies, but will not have the requirement of operating in real-time, although some parameters in the acquisition of metastable data will require change between consecutive GC peaks.

Studies such as these will identify in some detail the feasibility and necessity of closed-loop automation as well as the portions of the procedure which must be improved to

meet the time constraints imposed by limited sample quantities and GC/MS operation. We have already identified the problem of the rate at which resolution can be changed and have determined a potential solution. Additional problems under study are those of instrument sensitivity and strategies for metastable ion measurement.

3. Future Plans

Our future plans represent extensions of the on-going work described above under "PROGRESS" as well as the continued routine maintenance of the GC/MS systems. Specifics are briefly summarized below. A significant impact will occur with the termination of NIH support of the ACME computing facility in July 1973. We now perform most of our data reduction processing on the ACME 360/50. The follow-on facility to ACME will be an unsubsidized, fee-for-service facility mounted on a 370/158 computer along with other Stanford Hospital administrative computing functions.

We are in the process of implementing interim, stand-alone support for our GC/MS work on available PDP-11/20 machines. We have a number of proposals pending with NIH which would support longer term solutions, either through augmented stand-alone PDP-11 capabilities or through funds to operate on the fee-for-service 370/158.

a) MASS SPECTROMETER DATA SYSTEM AUTOMATION

Future efforts will include the transition of the existing ACME-based system to a follow-on system, the character of which will depend on NIH funding approval for one of the alternatives. We will adapt previously developed concepts for use in the Finnigan low resolution GC/MS system being used for routine urine analysis. We will develop data system extensions for the MAT-711 system which allow semi-automated acquisition and reduction of metastable information to support fragmentation pathway studies, Heuristic DENDRAL program development, and closed-loop simulation. This metastable system will incorporate calibration procedures and automated peak detection and resolution procedures based on the high resolution system. The existing hardware interface will be used to control source or electrostatic analyzer voltages in conjunction with the magnet scan to measure specific parent-daughter ion relationships.

b) GAS CHROMATOGRAPHY/HIGH RESOLUTION MASS SPECTROMETRY

We will complete the intermediate disk buffer in conjunction with the follow-on computing system transition to allow routine collection and filing of sequential spectra. We will exercize the system on body fluid samples in support of our clinical applications and the development of interpretation programs. As developments occur which improve sensitivity, we will incorporate these to extend the

power of the system.

c) AUTOMATED GC/MS DATA REDUCTION

The approach described above is still in the formative stage. We will complete the development and implementation of these ideas, test them in the clinical application domain and produce an automated system suitable for routine use by the biochemist.

d) CLOSED-LOOP INSTRUMENT CONTROL

With the development of a more automated method for acquiring information on metastable peaks (under subtask (a) plans), we will develop and exercise the strategy planning aspects of the Heuristic DENDRAL programs in connection with managing a urine analysis GC/MS run. This will be a simulation of closed-loop operation intended to demonstrate the feasibility and need for an actual implementation of these ideas. In support of these closed-loop simulations we will investigate the feasibility of instrument mode switching and simple control function such as ion source and electrostatic analyzer potentials and magnet scan.

REFERENCE

1) Lederberg, Joshua, "Rapid Calculation of Molecular Formulas from Mass Values," Journal of Chemical Education, Vol. 49, Page 613, September, 1972.

VII. Cell Separation

While this work was initiated by the subject grant, most of the support is now coming from NIH Grant GM 17367. The work has obvious applications in the medical field as well as possible applications for exobiology.

A. High Speed Fluorescent Cell Sorter

This unit is designed to measure the fluorescence of cells in a jet of liquid, break-up the jet into uniform drops and collect the drops in a series of containers, with all drops containing cells with similar fluorescent characteristics collected in the same container.

Our last annual report for the period ending December 31, 1971, described a multichannel cell separator which had just been completed and upon which testing was just commencing. The new instrument simultaneously measures fluorescence and scattering cross section of each cell. Both signals are used as sorting parameters. Much of our recent efforts have been devoted to evaluating, operating and improving this new instrument.

Installation of a second, more powerful laser (4 watts) has improved the sensitivity and made it possible to operate the new system completely independently of the old instrument which continues to be frequently used. Many modifications to the fluid and sample handling components

of the system have resulted in more reliable operation, accurate monitoring of sample flow rate, rapid flushing and sample changing, and fast recovery from nozzle blockages. The electronic signal processing logic has been replaced with improved circuitry that permits independent specification of scattering and fluorescence signal limits for each of two separated fractions. Detected cells that do not satisfy the criteria for sorting are positively isolated to the undeflected fraction, thus preventing contamination of the wanted cells. Purity of the separated fractions has also been improved by identifying pairs of cells too closely spaced to properly sort, and also isolating these cells to the undeflected fraction. These improvements also keep most empty droplets and much unwanted debris from either separated fraction. Cells are now processed at rates of several thousand per second with separated fraction purities routinely between 90% and 99%. Further refinement of the decision making circuitry is expected to increase both processing rate and fraction purity.

Programs for processing and plotting data from the two parameter 1024 channel pulse height analyzer have been written. A storage indicator and circuitry for displaying two-parameter "scatter plots" of cell parameters has been added to the system.

The minimum detectable number of molecules of rhodamine or fluorescein has been determined using radioassay techniques and conconavalin A (Con A), a protein from jack bean meal which binds to carbohydrate

residues on the cell surfaces. The Con A was labelled with both the fluor and with I^{125}. The I^{125} count on cell suspensions was used to determine the number of Con A molecules, and thus of fluor molecules bound. The number of fluor molecules bound was reduced until it was just possible to detect a signal above background noise in the cell separator. This minimum number seemed to be less than 4000 for either fluor.

In most of our work fluors are conjugated with immunoglobulins. Considering that each immunoglobulin molecule will usually contain more than one fluor molecule, and that further amplification can be obtained using the immunofluorescence "sandwich" technique, it appeared that cells with on the order of a few hundred active sites on their surface should be separable under ideal conditions using this unit. This was confirmed by tests on thymocytes carrying a few hundred molecules of synthetic polypeptide, T, G, A-----L, labelled with I^{125}, as a marker for autoradiography, reacted with a fluorescent antibody to T, G, A-----L and separated. The practical separation limit is usually set by nonspecific adhesion of fluor-containing molecules to the cells.

A major effort has been devoted to achieving sufficiently asceptic operation that separated fractions can be cultured. This has involved basic redesign of the flow system and much more care in sample and machine treatment. The work has been successful and several cultures have been grown, indicating anew that cells are viable after passing

through the unit.

The scatter channel has proved more useful than anticipated. The signal it provided with various red blood cell samples was compared to that from a Coulter counter. The variation with volume appeared to be somewhat less than linear. However larger cells gave larger signals, and the resolution of neighboring signals was significantly better in the scatter channel than in the Coulter.

This equipment makes many desirable but previously impossible biological experiments not only possible but relatively easy. The following are among the many applications which have been made:

1. Rabbit lymphocytes bearing different membrane antibody light chain markers have been separated and shown to give rise preferentially to plasma cells bearing the same markers.

2. Spleen cells from mice immunized with two different antigens were incubated with one of the antigens after the latter had been made fluorescent. These fluorescent cells were then separated. The non-fluorescent portion retained all of its antibody titer to the second antigen but lost most of its titer to the first, indicating that the activity was on different cells.

3. Populations of different apparent sizes have been separated from both mouse spleen and mouse thymus lymphocyte suspensions. Preliminary indications are that the spleen lymphocytes giving

the smaller scatter signal are dead, but the smaller thymus fraction appears to be functionally different than the larger, or at least to react differently to hydroxycortisone treatment of the animal. Experiments in this area, and in those discussed below are continuing.

4. Electron micrographs of antigen reactive cells have been made showing the location of molecules of antigen on the cell surface.

5. Preliminary results on treatment of separated cells with mitogens like PHA (phytohemagglutinin) and PWM (Pokeweed mitogen) indicate these reagents can successfully stimulate division in T but not B cells. Interferon is also produced in T but not B cell cultures, in contradiction to earlier speculations that B-cells are also interferon producers.

6. A series of experiments was conducted to determine the minimum leakage of Rh+ fetal cells into an Rh- maternal circulation which could be detected. In these tests Rh+ antibody was added to suspensions of Rh- cells containing various proportions of Rh+ cells, and the suspension then treated with fluorescent goat antihuman gamma globulin. Results showed that Rh+ cells could be detected easily by the separator at dilutions of 10^{-5}, and with care as low as 10^{-6}.

7. Work has started on using the separator to fractionate fetal lymphocytes from maternal peripheral blood using HLA antigens as markers. Preliminary experiments indicate that a sandwich technique, in which the cells are treated with antiserum to the father's HLA antigens, then with fluoresceinated rabbit anti-human gamma globulin, and run through the separator should be able to provide considerable enrichment of the fetal cells. It may be possible eventually to use this technique to provide fetal cell cultures for use in prenatal diagnosis of chromosome abnormalities.

VIII. Mariner Mars 1971 Orbiter Photography

The image processing work being carried out in cooperation with the Stanford Artificial Intelligence Project has resulted in about 500 image difference operations on MM'72 photos. Thirty-one images of the martian moons, Phobos and Deimos, were also processed to enhance their contrast and high frequency information. Late in the reporting period special emphasis was put on differencing images of the proposed landing sites for Viking 1975.

The image differencing operation which is described in detail in Technical Report IRL 1123 has represented our major interest. Candidate images for differencing are selected by ourselves and members of the MM-71 Image Team (primarily C. Sagan and J. Veverka). The necessary image data and navigation data is read from magnetic tapes supplied by Jet Propulsion Laboratory. The scientist requesting a particular set of image differences then has the opportunity to observe the work in progress and make decisions while the processing is being done. This interactive approach is possible through the availability of the Stanford AI time-sharing system operating on a PDP-10 computer and the associated image processing equipment.

Work has also been done in the area of image information management. An information retrieval capability has been implemented at the Stanford AI project which enables us to quickly review the planet coverage of the MM-71 TV Mission. It is primarily oriented toward revealing the extent

of repeated TV coverage of any area specified by latitude and longitude. It enables the user to quickly determine if an area has been photographed, and if so, how many times, on which orbits, by which camera, and by which pictures within an orbit. On a display screen is shown the disk of the planet, the footprint of the images, and vectors indicating view and sun angles (See Figure 1). Since more than seven thousand images exist, the need for such a system is obvious. This capability could prove quite useful in picture targeting and landing site refinement for the Viking mission.

It is important to note that this is an interactive system oriented towards the needs of the scientists. Its success depends on its ability to present data in a manner consistent in format and organization with the way the experimenters view the object under investigation.

The above mentioned capability actually represents the initial phase of the process for the projection and differencing operation. With the identifiers and footprints of all the images before the user the list can be pruned until just the footprints of interest are present. The user can then proceed directly to the projection and differencing steps.

The above capacity, when combined with a disc based storage system gives the scientist a significant degree of flexibility to review the image data and the processing carried out on it.

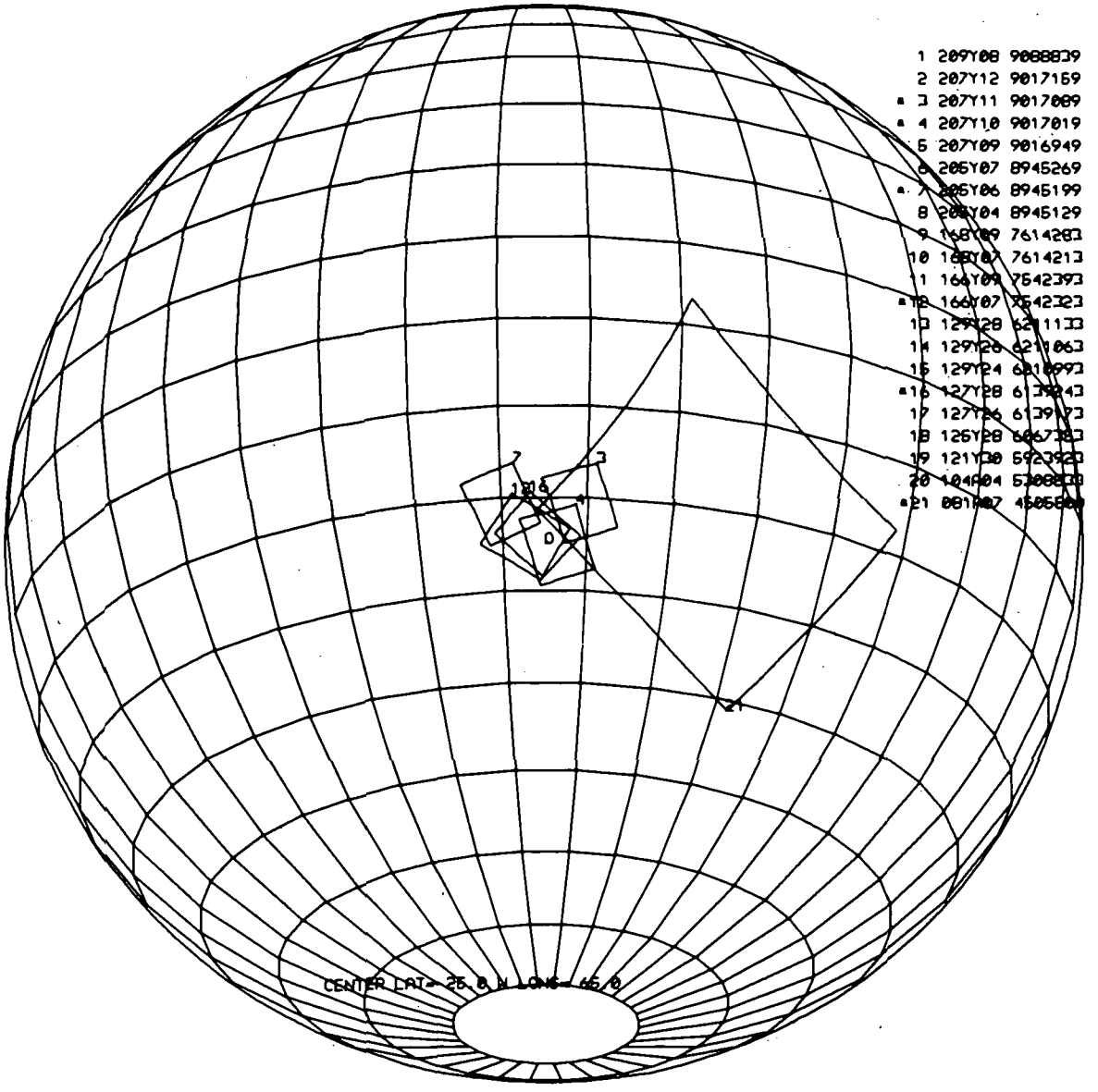

1 209Y08 9088839
2 207Y12 9017159
3 207Y11 9017089
4 207Y10 9017019
5 207Y09 9016949
6 206Y07 8945269
7 206Y06 8945199
8 206Y04 8945129
9 168Y09 7614283
10 168Y07 7614213
11 166Y09 7542393
12 166Y07 7542323
13 129Y28 6211133
14 129Y26 6211063
15 129Y24 6210993
16 127Y28 6139143
17 127Y26 6139073
18 126Y28 6067283
19 121Y30 5923923
20 104Y04 5200853
21 081Y07 4505893
CENTER LAT= 25.0 N LONG= 65.0

Figure 1.

A total of forty-nine picture differences were made which covered the areas at or near nine of the ten candidate landing sites for the Viking 1975 mission. One such difference can be seen in Figure 2. This example shows the area around Viking Landing Site three. The upper left image was taken on orbit 168 and the upper right on orbit 209 forty days later. During this time significant clearing took place, the differences left minus right and right minus left are shown in the lower left and right respectively.

A portion of the Viking related work was presented at a Viking landing site selection meeting in November 1972.

Projects for the next reporting period include a comprehensive study of the albedo changes which occurred between the MM-6 and MM-7 period and the MM-9 mission, completing work on a catalog of enhanced Phobos and Deimos images and a limited number of additional MM-9 differencing.

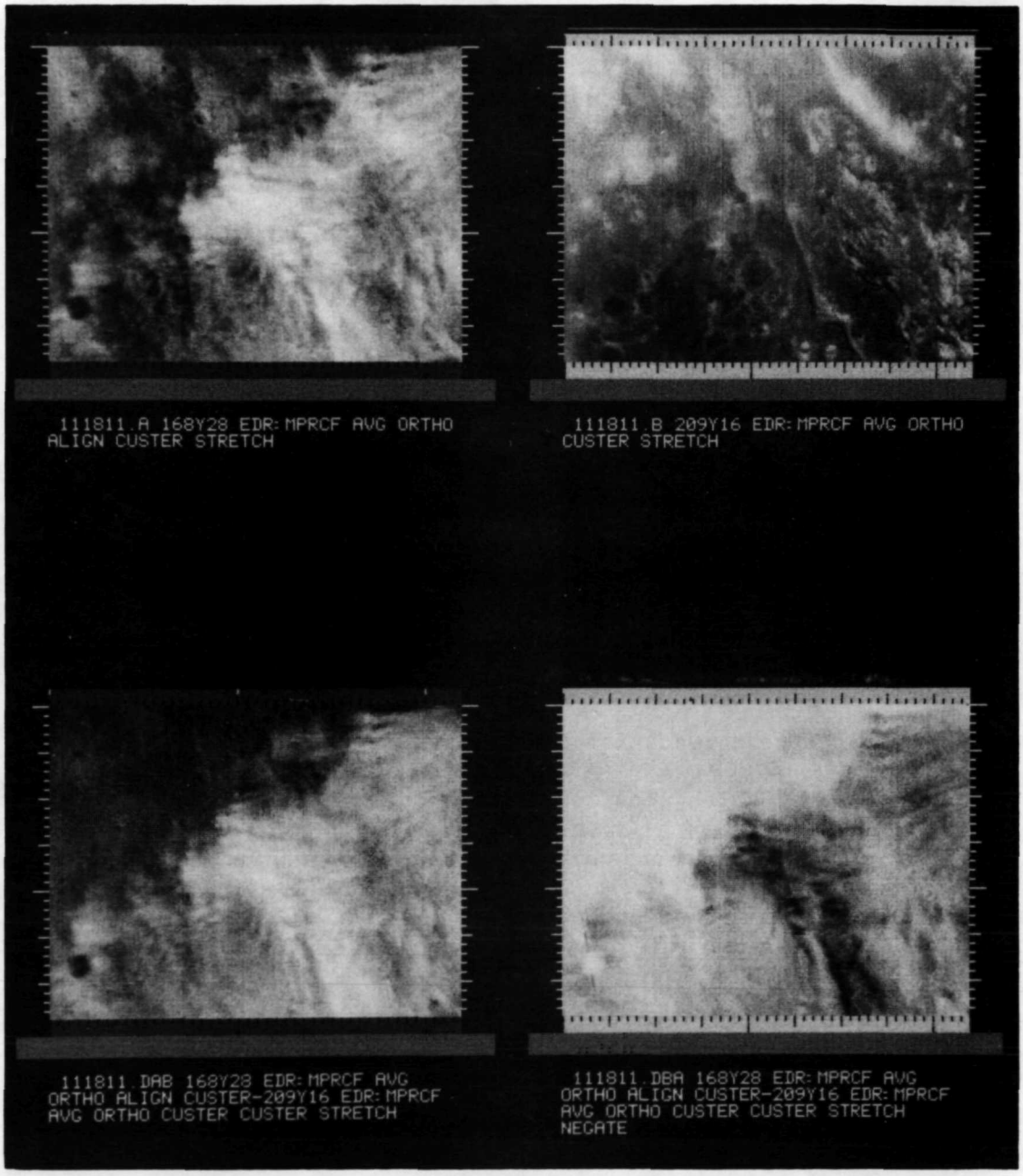

111811.A 168Y28 EDR: MPRCF AVG ORTHO ALIGN CUSTER STRETCH
111811.B 209Y16 EDR: MPRCF AVG ORTHO CUSTER STRETCH
111811.DAB 168Y28 EDR: MPRCF AVG ORTHO ALIGN CUSTER-209Y16 EDR: MPRCF AVG ORTHO CUSTER CUSTER STRETCH
111811.DBA 168Y28 EDR: MPRCF AVG ORTHO ALIGN CUSTER-209Y16 EDR: MPRCF AVG ORTHO CUSTER CUSTER STRETCH NEGATE

Figure 2

IX. Viking Lander Imagery

During the last year, Drs. Levinthal and Liebes, as members of the Lander Camera Science Team have assumed team responsibility for the Science Operations Requirements Document (SORD), the Software Functional Descriptions (SFD's) derived from SORD and the more detailed Software Requirements Documents (SRD's) that, in turn, are derived from the SFD's. In addition they are part of the ad hoc Viking Image Processing System Steering Committee which is planning the hardware configuration to meet the requirements for the imaging science component of Mission Operations. All of the above activities, while separately supported by Langley, relate to, and benefit from previous efforts in connection with the Mariner Mission and some of the basic work supported by this grant.

C. REPORTS AND PAPERS

This section lists reports and papers not referred to in preceding sections of the Program Resumé.

REPORTS

1. Annual Report for Period July 1, 1971 to June 30, 1972. JPL Contract 95289. Instrumentation Research Laboratory, Genetics Department, Stanford University. "Mariner Mars 1972". IRL Report 1143 (1972).

PUBLICATIONS

1. A. M. Duffield, W. E. Reynolds, D. A. Anderson, R. A. Stillman, Jr., and C. E. Carroll. "Computer Recognition of Metastable Ions." Proc. Nineteenth Annual Conference on Mass Spectrometry and Allied Topics, D4, 63-67 (1971).

2. W. E. Pereira, M. Solomon, B. Halpern. "The Use of (+)-2,2,2-Trifluoro-1-Phenylethylhydrazine in the Optical Analysis of Asymmetric Ketones by Gas Chromatography." Aust. Jour. Chem. 24, 1103 (1971).

3. W. A. Bonner, H. R. Hulett, R. G. Sweet and L. A. Herzenberg. "Fluorescence Activated Cell Sorting." Rev. Sci. Instr. 43, 404 (1972).

4. M. D. Solomon, W. E. Pereira, and A. M. Duffield. "The Determination of Cyclohexylamine in Aqueous Solutions of Sodium Cyclamate by Electron-Capture Gas Chromatography." Analytical Letters, 4(5), 301 (1971).

5. B. G. Buchanan, A. M. Duffield, A. V. Robertson. "An Application of Artificial Intelligence to the Interpretation of Mass Spectra." In "Mass Spectrometry: Techniques and Appliances" edited by George W. A. Milne, Published by John Wiley & Sons, Inc. 1971, pp. 121-178.

6. J. Cymerman Craig, W. E. Pereira, B. Halpern and J. W. Westley. "Optical Rotatory Dispersion and Absolute Configuration-XVII α-Alkylphenylacetic Acids." Tetrahedron 27, 1173 (1971).

7. L. H. Quam, S. Liebes, Jr., R. B. Tucker, M. J. Hannah, B. G. Eross. "Computer Interactive Picture Processing." Stanford Artificial Intelligence Report Memo AIM-166, STAN-CS-72-281. (1972).

8. L. G. Brookes, M. A. Holmes, I. S. Forrest, V. A. Bacon, A. M. Duffield and M. D. Solomon. "Chlorpromazine Metabolism in Sheep II. In Vitro Metabolism and Preparation of 3H-7-Hydroxychlorpromazine." Agressologie 12, 333 (1971).

9. W. E. Pereira and B. Halpern. "The Steric Analysis of Aliphatic Amines with Two Asymmetric Centres by Gas Liquid Chromatography of Diastereoisomeric Amides." Aust. J. Chem. 25, 667 (1972).

10. A. Buchs, A. B. Delfino, C. Djerassi, A. M. Duffield, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, G. Schroll and G. L. Sutherland. "The Application of Artificial Intelligence in the Interpretation of Low Resolution Mass Spectra." Advances in Mass Spectrometry, Vol. 5 314 (1972).

11. E. Steed, W. E. Pereira, B. Halpern, M. D. Solomon and A. M. Duffield. "An Automated Gas Chromatographic Analysis of Phenylalanine in Serum." Clinical Biochemistry, in press.

12. V. Tortorella, G. Bettoni, B. Halpern, P. Crabbe. "Optical Properties of Dimedonyl Derivative of Aromatic Amines and Amino Acids." Tetrahedron 28, 2991 (1972).

13. E. C. Levinthal, W. B. Green, J. A. Cutts, E. D. Jahelka, R. A. Johansen, M. J. Sander, J. B. Seidman, A. T. Young and L. A. Soderblom. "Mariner 9 - Image Processing and Products." Icarus, in press.

14. T. A. Mutch, A. B. Binder, F. O. Huck, E. C. Levinthal, E. C. Morris, C. Sagan, and A. T. Young. "Imaging Experiment: The Viking Lander" Icarus, in press.

15. C. Sagan, J. Veverka, P. Fox, R. Dubisch, J. Lederberg, E. Levinthal, L. Quam, R. Tucker, J. B. Pollack, and B. Smith. "Variable Features on Mars: Preliminary Mariner 9 Television Results." Icarus, in press.