Form Approved
Budget Bureau No. 68

SECTION I

DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE

APPLICATION
FOR CONTINUATION GRANT

REVIEW GROUP	TYPE	PROGRAM	GRANT NUMBER (Give Prefix)
			5 R24 RR00612-05

TOTAL PROJECT PERIOD
From: May 1, 1971 Through: April 30, 1974

REQUESTED BUDGET PERIOD
From: Jan. 1, 1974 Through: April 30, 1974

1. TITLE
Resource-Related Research -- Computers and Chemistry

2A. PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR (Name and Address, Street, City, State, Zip Code)
Dr. Edward A. Feigenbaum
Professor of Computer Science
Stanford University
Stanford, Ca. 94305

4. APPLICANT ORGANIZATION (Name and Address-Street, City, State, Zip Code)
Stanford University
Stanford, California 94305

2B. DEGREE
Ph.D.

2C. SOCIAL SECURITY NO.
[redacted]

5. PHS ACCOUNT NUMBER
458210

2D. DEPARTMENT, SERVICE LABORATORY OR EQUIVALENT
Computer Science Department

2E. MAJOR SUBDIVISION
School of Humanities & Sciences

3. ORGANIZATIONAL COMPONENT TO RECEIVE CREDIT FOR INSTITUTIONAL GRANT PURPOSES

6. TITLE AND ADDRESS OF OFFICIAL IN BUSINESS OFFICE OF APPLICANT ORGANIZATION
K. D. Creighton
Deputy Vice President for Business & Finan
Stanford University
Stanford, California 94305

7. RESEARCH INVOLVING HUMAN SUBJECTS (See Instructions)
☒ No ☐ Yes APPROVED: ____ Date

8. INVENTIONS (See Instructions)
☒ No ☐ Yes-not previously reported
☐ Yes-previously reported

9. PERFORMANCE SITE(S)
Computer Science Department
Department of Genetics
Department of Chemistry
Stanford University
Stanford, California 94305

TELEPHONE INFORMATION

	Area code	Tele. No.
11A. PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR (Item 2a)	415	321-2300 Ext. 4878
11B. Name of business official (Item 6) K. D. Creighton	415	321-2300 Ext. 2251
11C. Name and title of administrative official (Item 15b) Kathleen Butler Sponsored Projects Officer	415	321-2300 Ext. 2883

10. DIRECT COSTS REQUESTED FOR BUDGET PERIOD
$61,412

12A. CONGRESSIONAL DISTRICT OF APPLICANT ORGANIZATION SHOWN IN ITEM 4
Congressional District #17

12B. COUNTY OF APPLICANT ORGANIZATION SHOWN IN ITEM 4
Santa Clara

13. DO NOT USE THIS SPACE

14. CERTIFICATION AND ACCEPTANCE. We, the undersigned, certify that the statements herein are true and complete to the best of our knowledge and accept, as to any grant awarded, the obligation to comply with Public Health Service terms and conditions in effect at the time of the award.

15. SIGNATURES
(Signatures required on original only. Use ink. "Per" signatures not acceptable.)

15A. PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR — DATE

15B. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION — DATE

PHS 398-1 (Previous editions obsolete)
REV 1-70

...ing the Science Information ... changes.

Not for publication or publication reference.

U. S. Department of
HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE

NOTICE OF RESEARCH PROJECT

PROJECT NO. (DO NOT USE THIS SPACE)

TITLE OF PROJECT

Resource-Related Research -- Computers and Chemistry

GIVE NAMES, DEPARTMENTS, AND OFFICIAL TITLES OF PRINCIPAL INVESTIGATORS OR PROJECT DIRECTORS AND ALL OTHER PROFESSIONAL PERSONNEL ENGAGED ON THE PROJECT.

Edward Feigenbaum, Principal Investigator; Bruce Buchanan, Computer Scientist; N. Sridharan, Computer Scientist; Alan Duffield, Research Associate; Ray Carhart, Research Associate; Harold Brown, Research Associate; Geoff Dromey, Research Associate; Tom Rindfleisch, Research Associate; Dennis Smith, Research Associate; Ernest Steed, Research Engineer; Nicholas Veizades, Research Engineer; Robert Tucker, Computer Programmer; William White, Computer Programmer.

NAME AND ADDRESS OF APPLICANT INSTITUTION

Stanford University
Stanford, California 94305

SUMMARY OF PROPOSED WORK – *(200 words or less – Omit Confidential data.)*

In the Science Information Exchange summaries of work in progress are exchanged with government and private agencies supporting research in the biosciences and are forwarded to investigators who request such information. Your summary is to be used for these purposes.

A computer program, named Heuristic DENDRAL, has been developed in the Heuristic Programming Project at Stanford to work specifically with the problem of interpreting mass spectra. It has already demonstrated its ability to interpret the low resolution mass spectra of aliphatic ketones, ethers and amines and the high resolution mass spectra of estrogenic steroids. The objective of the proposed work is to expand the capabilities of the Heuristic DENDRAL computer program in a variety of ways in order to make it a tool of wider utility.

The original proposal was broken into three separable, but related, proposals. All of them enhance the power of the mass spectrometer as a tool for organic and biochemists and enhance the effectiveness of the Stanford interdepartmental mass spectrometry facility (Medical School and Chemistry Department) as a research resource. New work is proposed within the original three parts.

Part A. The first part proposes to entend the analytic capabilities of the Heuristic DENDRAL program to the mass spectra of complex organic compounds. In particular, the efficient implementation of the cyclic structure generating algorithm will be the focus of this work.

Part B. This section of the proposal for computer interpretation of the mass spectra of metabolites has been divided into two categories, reflecting instrumentation and laboratory support for this goal.

Part B (1). The first subpart is devoted to the improvement of GC/MS data system capabilities and the coupling of extracted data to the Heuristic DENDRAL programs for analysis.

PROFESSIONAL SCHOOL *(medical, dental, etc.)* WITH WHICH THIS PROJECT SHOULD BE IDENTIFIED	SIGNATURE OF PRINCIPAL INVESTIGATOR	DATE
Humanities and Sciences		

DO NOT WRITE BELOW THIS LINE – FOR OFFICE USE ONLY

SUPPORTING AGENCY

TYPE OF SUPPORT *(Check one)*

☐ Agency Staff (Intramural) ☐ Negotiated Contract ☐ Special Project Grant ☐ Research Grant ☐ Other (Specify)

OBLIGATED CURRENT F.Y.	NUMBER OF FUTURE YEARS TENTATIVELY ASSURED BEYOND CURRENT FISCAL YEAR	BEGINNING DATE	ESTIMATED COMPLETION DATE

NIH - 1946 (FORMERLY PHS - 166)
REV. 6-72

Part B(ii). The second subpart is aimed at analysis of body fluids, such as urine, by gas chromatography/mass spectrometry in the service of clinical problems.

Part C. The last part is work aimed at extending the knowledge about mass spectrometry, and thus extending the power of the mass spectrometer, by using a computer to codify and reason about large collections of mass spectra.

Progress Report

Part A. APPLICATIONS OF ARTIFICIAL INTELLIGENCE TO MASS SPECTROMETRY

OBJECTIVES:

Research activities carried out under Part A of this project have been directed toward extending the reasoning power of Heuristic DENDRAL. Heuristic DENDRAL reresents a paradigm for attacking problems in one of the major areas of importance to any scientific discipline dealing with molecules, the area of structure elucidation. We have focused our attention on the use of heuristic programming techniques for analysis of mass spectra and ancillary analytical data which can be obtained utilizing a mass spectrometer. It is convenient to discuss objectives, progress and plans by examining three broad areas of activity in research connected with Part A. We wish to note that these areas conform to our overall strategy of PLAN-GENERATE-TEST. We have shown, earlier, how powerful this strategy is when applied to the task of structure elucidation utilizing mass spectral data. The areas and their objectives are the following:

(I) PLANNER:
- (a) Extend the programs used for structure elucidation to structural analysis of complex molecules.
- (b) Assess the capabilities and limitations of the PLANNER.
- (c) Generalize the programming techniques to reduce compound class dependence.
- (d) Explore the utility of ancillary data available from the mass spectrometer.

(II) STRUCTURE GENERATOR:
- (a) Complete the exhaustive, irredundant generator of molecular structures.
- (b) Develop efficient constraints on the generator to exploit its potential utility.
- (c) Exploit the concepts developed for the structure generator in solving various structure-problems (related to m.s. and others) and isomer-problems.

(III) PREDICTOR
- (a) Extend the Predictor to still more complex molecular structures.
- (b) Explore the design of experimental strategies, utilizing Predictor functions, to differentiate among candidate solutions.

We point out that the PLAN-GENERATE-TEST strategy, although applied to structure elucidation, has potential utility as a strategy for solving other chemical problems. Similarly, although we utilize mass spectral data almost exclusively, the same heuristic programming techniques allow facile extension to analysis of data from other types of analytical instrumentation. These were not objectives of the original research proposal but seem logical extensions for future work. We have illustrated the potential of these techniques for analysis of 13C NMR data (Carhart and Djerassi, 1973). This is discussed briefly under the PLANS section, below.

PROGRESS:

(I) PLANNER

The function of the Planner is to analyze mass spectral data acquired on a compound. The Planner attempts to derive structural information from these data using the rules of behavior of compounds in the mass spectrometer.

Objective (a): Extend Programs.

The Planner is presently embodied in a program which also contains a set of functions to assemble this structural information into complete molecules (a primitive Structure Generator) and to test these molecular structures with other, not necessarily mass spectral, rules (a primitive Predictor). This performance program was written in this way to provide a useful tool for chemical studies while more general versions of the Structure Generator and Predictor were being developed. This program and its performance have been described in some detail in a publication and in previous progress reports. A manuscript (Smith, et.al., 1973) has now appeared describing the application of this program to the analysis of mixtures of compounds without prior separation.

Objective (b): Assess Capabilities.

We have extended the capabilities of the Planner so that we can analyze both low and high resolution mass spectral data. A low resolution mass spectrum is regarded by the program as a pseudo high resolution spectrum wherein possible elemental compositions of each peak are limited only by the inferred molecular formula of the compound. This results in more ambiguity with a commensurate increase in number of candidate solutions as would be expected considering the lower specificity of low resolution data as compared to high resolution data.

We have extended our capabilities for molecular ion determination utilizing a heuristic search technique through the space of plausible molecular ions. This technique has had significant success even when dealing with the low resolution mass spectra of compounds which display no molecular ion, for example the class of derivatized amino acids (trifluoracetyl, n-butyl esters) important to studies carried out under Part B, below.

We have segmented the performance program to decrease the amount of memory required for its operation. This should increase the chances for other groups to make use of the program.

The limitations of the present performance program are primarily the requirement that some information about the class of compounds be known, and that, for each class, relatively detailed rules about the mass spectral fragmentation of this class be available. The former limitation results primarily from the nature of the program in that a complete structure generator is not incorporated. The primitive structure generator available to the program can only place substituents about an assumed skeleton. This limitation will be alleviated when a structure generator with GOODLIST and BADLIST constraints is available (see Structure Generator, below). The latter limitation is more fundamental, but is characteristic of every spectroscopic technique to one degree or another. It must be assumed that analysis of a mass spectrum, alone, may not lead to sufficiently unambiguous information about the structure of the compound yielding the spectrum. It is for this reason that extensions of the programming techniques to encompass data from other spectroscopic techniques are attractive.

Objective (c): Generalize Techniques.

We have carried out several successful experiments to ensure that the performance program, used originally for analysis of estrogenic steroids, retains only procedures which are compound-class independent. By supplying fragmentation rules for other classes of compounds, we have successfully carried out structure elucidation of molecules in several diverse classes including other steroidal hormones and related compounds (progesterones, testosterones, androsterones), steroidal sapogenins and derivatized amino acids.

Objective (d): Explore Utility

Previous progress reports have summarized in some detail the ways in which data from ancillary techniques in mass spectrometry (metastable ion and low ionizing voltage data, labile hydrogen exchange) can be used by the program. The utility of metastable ions for aid in stucture elucidation continues as an active area of interest. Experience with the program has inspired studies on metastable ions, first, to help delineate the course of fragmentation of molecules with the purpose of extending and refining fragmentation rules used by the program (Smith, Duffield and Djerassi, 1973). Experience with the increased specificity of structural information with concomitant reduction in analysis time when metastable ion information is available (Smith, et.al., 1973) has led to a study of a new technique for detection and analysis of metastable ions (Direct Analysis of Daughter Ions, or DADI) and has illustrated the utility of this technique in mixture analysis (Smith, Djerassi, Maurer and Rapp, 1973).

Experience with the PLANNER has led to several research activities related to, but not supported by, this grant. Our studies of estrogen mixtures isolated from pregnancy urine have suggested new compounds likely to be important in the human metabolism of estrogens. Some of these compounds are hitherto unreported structures and a synthesis program is underway in Professor Djerassi's laboratory to produce some of these compounds. The Planner will be used as one method of comparison of the synthesized, authentic standards with those isolated from pregnancy urine.

Work is also being carried out to explore the fragmentation of model systems possessing two heteroatoms in close proximity. It is clear from the first of these studies (Block, Smith, and Djerassi, 1973) that the fragmentation of these difunctional systems does not reflect that of monofunctional analogs. More groundwork is required in this area to obtain better fragmentation rules for these systems.

II. STRUCTURE GENERATOR

Objective (a): Complete the Generator

The last progress report discussed the completion of both the basic structure generator algorithm and program, which provide the capability for exhaustive generation of graph isomers of a given empirical formula, with prospective avoidance of duplicate structures. Since the time of the submission of that report, manuscripts describing the structure generator, directed specifically to an audience of chemists, have been submitted (Masinter, Sridharan, Lederberg, and Smith, 1973; Masinter and Sridharan, 1973). Some effort over the past year has been devoted to

verification of the completeness and irredundancy of the method. We have extended existing combinatorial counting algorithms to check that the numbers of isomers generated are correct. We have used an interactive version of the generator to verify that variations (allowed by the algorithm) of the mechanism of generation yield the same set of isomers. In this way we are now increasingly confident that the program's performance accurately reflects the mathematically proven algorithm on which it is based.

The Structure Generator has been briefly described, and placed in its context within Heuristic DENDRAL, in an invited paper presented at a NATO/CRS sponsored conference on Computer Representation and Manipulation of Chemical Information, held in Amsterdam in June, 1973 (Smith, Masinter and Sridharan, 1973).

We have also begun to develop techniques to expand the scope of the generator. One example, which has been completed, is adding extensions to the CATALOG. The CATALOG contains the set of vertex-graphs from which structures are assembled. The original CATALOG was not sufficient to generate all isomers of some potentially interesting compositions, e.g., those involving graphs possessing nodes of degree >3. We now have a program which constructs complete sets of vertex-graphs containing nodes of degree >3 from the set of trivalent graphs in the original CATALOG. We have thus extended the capabilities of the generator. Other such extensions are discussed in the PLANS section, below.

Objective (b): Develop Constraints

It is absolutely essential that we provide the mechanism for constraining the Structure Generator: without constraints it is merely a legal move generator, as in a chess-playing program. For structure elucidation problems, the Planner can determine many features of the molecular structure from various types of experimental data such as presence of functional groups, and the numbers of double bonds and rings. Partial information of this sort can be used to constrain the Structure Generator to the space of plausible candidate structures. From a graph-theoretic point of view, however, constraining the graph generating algorithm is a difficult unsolved problem.

We are presently formulating several types of constraints to apply to the Structure Generator. Some types of constraint await the development of new mathematical tools (see PLANS), while others can be immediately implemented with relatively minor alterations to the algorithm. The class of constraints presently receiving attention deals with types of unsaturation (rings or double bonds) desired in the final structures. Related to this constraint is the constraint of number of quaternary carbons present. The former information (number and nature of multiple bonds) is readily available from several spectroscopic techniques, while the latter may be obtained from 13C NMR. The implementation of this class of constraints will be used as the model for future implementation of a GOODLIST (structural features known to be present) and a BADLIST (structural features known to be absent).

It is possible that some types of constraints may not be easily implemented within the algorithm. Thus, retrospective tests of isomers may be required to search for desired or unwanted features. We have developed some new approaches to graph matching which seem to be significantly more efficient than previous methods. Should prospective implementation of a constraint prove difficult, we will

have at our disposal some powerful graph matching tools to exercise the constraint.

Objective (c): Exploit the Generator for Structure Elucidation

We have demonstrated the utility of some subsystems of the structure generator, e.g., the LABELLER, by exploring some problems of isomerism noted in the chemical literature. We have corrected the member and presented the identities of isomers formed by different substitutions of alkyl chains about a porphyrin nucleus. We are presently exploring some problems of isomerism of carbocyclic ring systems, specifically C10H10 and (CH)10 and C10H2n-4 tricyclic ring systems, n = 8 - 12, related to the mechanistics of isomeric interconversion.

We have the complete list of all topologically possible 1176 6-membered Diels-Alder ring systems, using any combination of C,N,O and S. This list was generated using the PARTITIONER and an extended version of the LABELLER. These are all the 6-membered ring systems that can be embedded in structures resulting from the well-known Diels-Alder reaction. Of the 1176 possible ring systems, approximately 80% are unreported in the Ring Index. Many of these are chemically unstable - underscoring the need for a BADLIST implementation for the Cyclic Structure Generator. However, many of these unreported ring systems are certainly chemically plausible. Awareness of such gaps in relatively simple synthetic categories might lead to discovery of new categories of compounds with important biological effects.

(III) PREDICTOR

The function of the Predictor in the PLAN-GENERATE-TEST strategy is to perform a detailed evaluation of candidate solutions (structures) to a structure elucidation problem. It may use a more detailed model of spectroscopic behavior than that embodied in a Planner to attempt to differentiate among possible solutions.

Objective (a): Extend the Program

We have extended and generalized the Predictor used previously for saturated, aliphatic, monofunctional compounds. Given a list of structures and rules of fragmentation processes, it will predict a mass spectrum for each structure. Prediction of relative ion abundances is crude, but previous work has shown that even crude measures of ion abundance are usually satisfactory. The predicted spectrum can be matched then with the observed and candidates ranked according to the quality of the match. The program works with structures and rules of any complexity. An interesting philosophical question is how much knowledge should be brought to bear on interpretation of the data at the Planning vs. Predicting stages of analysis. It is our feeling that if more can be accomplished during Planning to constrain the Structure Generator, the analysis will be more efficient. On the other hand, some knowledge can be utilized only if a complete structure is specified, so that its use is restricted to a predictive role. Moreover, Predictor Functions have a greater utility, as indicated in the subsequent section.

Objective (b): Differentiate Structures

The Predictor has a more obvious application in the design of

experimental strategies to differentiate among candidate structures. Rules of spectroscopic behavior utilized during Planning demand the presence of some data to evaluate. The Predictor can then be used to request additional data from any source to aid in differentiation. We have explored this approach by analyzing the spectrum of a compound with the performance program. The Predictor was used to evaluate the the set of candidate structures to define the minimum number of metastable defocussing experiments necessary to achieve a unique solution. Thus, no time need be spent acquiring unnecessary or redundant data. Clearly, this has important implications for future work in that many different types of data (e.g., NMR, IR) might be requested by the Predictor to facilitate identification.

PLANS

For the remaining period of this grant we propose to carry out the following extensions of the research outlined above.

(I) PLANNER

The major area of activity related to the present version of the Planner will be to focus our attention on using the program in support of chemical studies outlined under Part B (see below). The chemical extraction and derivitization procedures used in the analysis of body fluids restricts the types of compounds present in each separated fraction. Such simplifications make this a problem more amenable to attack. Only certain classes of compounds are present in each fraction, and we have some knowledge of the mass spectral fragmentation of these classes. We wish to couple the program to the results of library matching procedures so that we direct our efforts to structure elucidation of those components which have not been previously identified. This is particularly important in the context of analysis problems such as those discussed under Part B.

We propose increasing the utility of the program by removing two present constraints: (a) allow unspecified "dummy" atoms in the skeleton instead of requiring a rigidly fixed structural skeleton, and (b) allow fragmentation processes to be specified more flexibly - in particular, allow fragmentations in substituents on the skeleton instead of requiring all fragmentations to cut through the skeleton.

Although we are presently uncomfortable with immediate coupling of the Structure Generator to the Planner, we propose continued exploration of the problems of controlling the generator automatically. Actual implementation awaits a more comprehensive treatment of the problem of constraints.

II. STRUCTURE GENERATOR

The inclusion of a reasonable set of constraints is obviously required and will be the subject of most of our future development work. We plan to develop an interface to the present interactive version of the Structure Generator that speaks a more chemical language. This interface will be designed to avoid the present requirement that the user know something about the program before he can use it. As the optimum method for implementation of a constraint is determined, the interface will be extended to translate the usual specification of the constraint in chemical terms into rules acting at the level of the program. As we stressed in development of the PLANNER, there are considerable advantages to building a powerful program in an

incremental fashion. These steps are logically directed to our longer term goal of developing a useful structure elucidation tool for the chemist, based on the structure generator.

There are several other areas of interest which are peripherally related to the problem of constraints and which will occupy our attention. The Structure Generator knows no chemistry other than atom names and their associated valence. There are several important areas where this is an immediate problem. For example, the program has no explicit awareness of the aromatic resonances, leading to a remediable redundancy in the list of isomers. An aromaticity-predictor is also indispensable for anticipating chemical behavior of a structure.

We wish to deal with types of isomers besides simple connectivity isomers. We need to have the facility for assembly of molecular sub-structures (the usual type of information inferred from spectroscopic data) when such an assembly yields new rings or multiple bonds. All the above questions need a reexamination of the fundamental mathematical considerations. The present algorithm has been proven to yield complete and irredundant solutions. In devising new algorithms or variants of the present one, the burden of proof can be reduced to (the usually easier) equivalence to the previous algorithm. Professor Harold Brown, who was the mathematician instrumental in initial development of the labelling algorithms for structure generation, will be with us again for several months to help attack the problems outlined above.

III. PREDICTOR

Although the Predictor has been essentially finished for our own internal use, we propose to spend a modest amount of time in the coming months making it more usable by others. In particular, we wish to extend the initial work on predicting the new experiments necessary for distinguishing among candidate structures (e.g., predicting that a metastable peak at mass 70.1 would confirm one structure and disconfirm another). In addition, we plan to work on cataloging some existing sets of mass spectrometry rules in such a way that the program can be easily used for different classes of problems.

part A references (Published or submitted during year)

R. Carhart and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference XI....

D.H. Smith, B.G. Buchanan, R.S. Engelmore, H. Adlercreutz, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference, IX. Analysis of Mixtures Without Prior Separation as Illustrated for Estrogens," J. Amer. Chem. Soc., September 5, 1973.

D.H. Smith, A.M. Duffield, and C. Djerassi, "Mass Spectrometry in Structural and Stereochemical Problems, CCXXII. Delineation of Competing Fragmentation Pathways of Complex Molecules from a Study of Metastable Ion Transitions of Deuterated Derivatives," Org. Mass Spectrom., 7, 367 (1973).

D.H. Smith, C. Djerassi, K.H. Maurer, and U. Rapp, "Mass Spectrometry in Structural and Stereochemical Problems, CCXXXIV. Applications of DADI, A Technique for Study of Metastable Ions, to Mixture Analysis," J. Amer. Chem. Soc., submitted (1973).

J.M. Block, D.H. Smith, and C. Djerassi, "Mass Spectrometry in Structural and Stereochemical Problems, CCXXXVIII. The Effect of Heteroatoms upon the Mass Spectrometric Fragmentation of Cyclohexanones," J. Org. Chem., submitted (1973).

L.M. Masinter, N.S. Sridharan, J. Lederberg, and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference XII. Exhaustive Generation of Cyclic and Acyclic Isomers," J. Amer. Chem. Soc., submitted (1973).

L.M. Masinter and N.S. Sridharan, "Applications of Artificial Intelligence for Chemical Inference, XIII. Labelling Objects Having Symmetry, J. Amer. Chem. Soc., submitted (1973).

D.H. Smith, L.M. Masinter, and N.S. Sridharan, "Heuristic DENDRAL: Analysis of Molecular Structures," to be published in the proceedings of the NATO/CRS Advanced Study Institute on Computer Representation and Manipulation of Chemical Information.

N.S. Sridharan, "Computer Generation of Vertex Graphs", Stanford Computer Science Memo CS-73-381, Stanford University, July, 1973.

Part B-i Gas Chromatograph - Mass Spectrometer Data System Development

OBJECTIVES AND RATIONALE

The objectives of this part of the research project are the improvement of GC/MS data system capabilities and the coupling of extracted data to the Heuristic DENDRAL programs for analysis. We ultimately seek a substantial degree of interaction between the instrumentation and the analysis programs including computer specification and control of the data to be collected. In addition to the development goals, this portion of the project provides for the day-to-day operation of the GC/MS systems in support of mass spectrum interpretation computer program development (Parts A and C) and applications of GC and MS to biomedical and natural product sample analysis with collaborators.

Our rationale for this approach is that the overall system should be designed for problem solving rather than just for data acquisition. This implies that analytical computer programs, after review of available experimental data, could be able to specify additional information needed to confirm a solution or distinguish between alternative solutions. Such requests could be passed back to an instrument management program to set up proper instrument parameters and collect the additional information. Our initial objectives to implement an on-line, closed-loop system using the ACME computer facility have met with a number of difficulties. These grow principally out of ACME's limited computing capacity and commitments as a general time-sharing service. In addition, the scanning high resolution mass spectrometer has inherent sensitivity limitations, which do not preclude a demonstration but rather limit the practical sample volume which could be analyzed. Until such limitations can be overcome, particularly in terms of computing support, we have focussed our efforts on an open-loop demonstration of such an approach.

PROGRESS

Progress has been made in demonstrating a GC/High Resolution Mass Spectrometry capability, in further developing automated data analysis algorithms, and in planning for the implementation of a data system for the collection of metastable ion information. Progress in these and other areas directed toward the main research goals has been impacted by a transition in computing support which is still underway. This transition, discussed in more detail below, was occasioned by the phase-over of the ACME computing facility, which we had been using, from NIH grant subsidy to a fully fee-for-service operation under Stanford University auspices.

Summaries of the results and problems encountered in each of the areas follow.

Gas Chromatography/High Resolution Mass Spectrometry (GC/HRMS)

We have verified the feasibility of combined gas chromatography/high resolution mass spectrometry (GC/HRMS). Using programs described in previous reports, we can acquire selected scans and reduce them automatically. The procedures are slow compared to "real-time" because of 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.

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 can be a useful analytical adjunct to our low resolution GC/MS clinical studies (Part B-ii), 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. We can nevertheless make progress in applying GC/HRMS techniques to accessible effluent peaks and can adapt the improved sensor capability when available.

These experiments have also shown that the ACME computer facility cannot reliably provide the rapid service required to acquire and file repetitive spectometer scans. This problem is to be expected in a heavily used time-shared facility without special configuration for high rate, real time support. Excepting possible requirements for real time data analysis (such as in a closed-loop system), this problem could be solved by implementing a large local buffer (e.g., disk) at the front-end data acquisition mini-computer. We are exploring this possibility in conjunction with the overall planning for computer support discussed below.

Data Analysis Algorithms

A. Peak Resolution

One of the significant trade-offs to be made in GC/HRMS is that of sensitivity versus resolution. In maintaining high instrument resolution (in the range of 5,000-10,000) while scanning fast enough to analyze a GC effluent peak (approximately 10 sec/decade), system sensitivity is constrained as discussed above. We have worked on a method for reducing instrument resolution requirements through more sophisticated computer analysis of a lower resolution output. In effect this transfers the burden of overlapping peak detection and mass determination to the computer instead of requiring inherently well resolved data out of the instrument. The advantage comes in better system sensitivity.

Unresolved peaks are separated by an analytical algorithm, the operation of which is based on a model peak derived from known singlet peaks in the data. Actual tabulated peak models are used rather than the assumption of a particular parametric shape (e.g., triangular, Gaussian, etc.). This algorithm provides an effective

increase in system resolution by approximately 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.

B. GC Analysis

The application of GC/MS techniques to clinical problems as described in Part B(ii) of this proposal has indicated the desirability of automating the analysis of the results of a GC/MS experiment. 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 promising results. By using a disk-oriented matrix transposition algorithm, the array of 700 spectra by 500 mass samples per spectrum can be rotated to gain convenient access to the "mass fragmentogram" form of the data. The transposition algorithm avoids many successive passes over the input data file as would be required in a straightforward approach. By generating a reorganized intermediate file, time savings by factors of 5-10 are achieved.

The fragmentogram 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 higher resolution in localizing GC effluent constituents. Thus by transposing the raw data to the mass fragmentogram domain, we can systematically analyze these data for baselines, peak positions, and amplitudes, and thus derive better mass spectra for the relatively few constituent materials. These are 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.

We have applied a preliminary version of this algorithm to several urine samples. These contain several apparently simple peaks which in fact consist of multiple components. The algorithm performs well in separating out these constituents although further testing is required.

Closed-loop Instrument Control

In the long term, it could be possible for the data interpretation software to direct the acquisition of data in order to minimize ambiguities in problem solutions and to optimize system efficiency. The task of deciding among and collecting various types of mass spectral information (e.g., high resolution spectra, low ionizing voltage spectra, or selected metastable ion information) under closed-loop control during a GC experiment is difficult. Problems arise because of the large requirements placed on computer resources

and present limitations in instrument sensitivity or data read-out imposed by the time constraints of GC effluent peak widths. Solutions to these problems may not be economically feasible within currently existing technology but seem achievable in the future. We are studying this problem in a manner which would entail a multi (two or three) - pass system. This permits the 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. Such additional data could help to uniquely solve a structure or at least to reduce the number of candidate structures. This simulated closed-loop procedure could demonstrate the utility of DENDRAL type programs to examine data, determine solutions and propose additional strategies, but will not have the requirement of operating in real-time. Some parameters in the acquisition of particular types of information, such as metastable data, will require computer control, even in the open-loop mode.

We have considered plans to implement two aspects of instrument control, in addition to the magnetic scan control implemented for GC operation and reported previously. These include system resolution control, such as would be required to change from normal spectrum scanning mode to metastable scanning mode, and high voltage control necessary to selectively measure metastable ion fragmentation data. In addition to these we have considered the discrete switching of various electronic mode controls which are straightforward and not discussed in detail.

Implementation plans for computer control of these instrument functions have been delayed because of the ACME computing facility transition which diverted the necessary hardware and software manpower.

Resolution control involves changing the widths of the slits at the exit of the ion source and the entrance to the ion multiplier detector. Additional source and electrostatic analyzer voltages must be controlled to optimize performance, as discussed later. Mechanical slit adjustment is accomplished on the MAT-711 instrument by heating wires which support the slit jaws. The resulting expansion or contraction of those wires move the spring-loaded jaws. As implemented by the manufacturer, the time constants involved in heating the control wires are 5-10 seconds. It is possible to speed this up to approximately 0.5 seconds by application of a controlled over voltage decreasing to the appropriate equilibrium value for the desired slit width. This was demonstrated by a series of experiments on an extra slit assembly mounted in a vacuum jar in our laboratory. Cooling of the wires is relatively fast in the way they are mounted so no problem exists in that direction.

It is desirable to have feedback to indicate the actual slit width achieved rather than relying on a slit assembly calibration. Stretching of the support wires or changes in the spring tension under temperature cycling would change this calibration. An optical scheme to measure slit width in situ is possible. We do not contemplate implementing this feedback immediately because it requires major changes to the instrument flight tube.

Two types of metastable ion relationships are obtainable by suitable control of the double-focussing instrument. First, for a given daughter ion, one can trace the parent ions which give rise to

it. Second, for a given parent ion one can trace the various daughters to which it gives rise. The first measurement ("metastable defocussing") is the more straightforward for this instrument since parent ions can be enumerated by a simple scan of the accelerating voltage, holding the electrostatic analyzer (ESA) voltage and magnetic field constant. The second type of scan requires the coordinated scan to two of the three fields. We feel that joint computer control of the accelerating voltage and ESA voltage is the simpler approach since the magnetic field is more difficult to set and monitor because of hysteresis effects. For a resolution of 1000 in the metastable ion mass measurement, the voltages must be set to approximately .01-.02% accuracy. This requires a 14-16 bit digital-to-analog (D/A) converter to control the input (10 volts) to the operational amplifier which generates the high voltage. Similar D/A controls of ion source voltages for ion current and focus optimization can be implemented using optical isolators to allow vernier control of the various high voltages around the nominal 8KV values.

Computing Transition

As mentioned earlier, the transition of the ACME computing facility from NIH subsidy to Stanford-sponsored fee-for-service operation has impacted our development efforts this past year. Both the low resolution instrument used for routine body fluid analysis research and the high resolution instrument are affected. All computing support was previously obtained from the ACME facility, much of it as core research without explicit transfer of funds. The transition has required consideration of both technical and economic factors. The new facility represents a combination of the previous ACME interactive and real time computing load with various administrative and batch computing loads on a new IBM 370/158 computer. This new environment will have even more difficulty in supporting real time computing needs than ACME did. No real time support has been available since the 360/50 service was discontinued on July 31, although terminal service was reestablished in mid-August. Data acquisition service via the IBM 1800 is expected to be operational by early November.

For the high resolution instrument, this transition, as a minimum, necessitates an interface modification (we previously sent data through the IBM 2701 interface no longer to be supported). It also amplifies the problems we encountered in sending and filing high rate mass spectrometer data (particularly during GC/MS runs). These problems would be present to some extent in any general time-sharing service machine without specific hardware and software configuration provision for these needs (such provisions for real time support had been proposed in our SUMEX computer application).

After examining a variety of alternatives, we conclude that a dedicated mini-computer solution (built around a machine with the arithmetic capability of a PDP-11/45) would be highly attractive technically and relatively inexpensive. A stand-alone mini-computer system would cost in the range of $50,000-$60,000, augmenting existing equipment, plus approximately $9,000 per year maintenance and $2,000 per year for supplies. Estimates for 370/158 support, based on current charging algorithms and previous utilization experience, run from $35,000-$50,000 per year. This spread is caused by uncertainties in the effects of planned measures to increase operating efficiency and possible changes to the rate structure. In

any case, the mini-computer approach pays for itself in 1 to 2 years of operation and provides the responsiveness of a dedicated machine for real time support. Unfortunately our existing budget does not provide for this solution. The budget is very marginal for purchase of computing support from the 370/158 as well. This later approach is the only currently available one, however, since it can be implemented with relatively low start-up cost. The effect of budget limitations appears in terms of a reduced number of samples which can be run. We have attempted to minimize the other budget costs (manpower principally) to increase the computing funds available. This will necessarily impact our development goals. We hope, in the renewal application for DENDRAL support, to be able to implement the more effective mini-computer approach for the high resolution spectrometer as a longer term solution.

We have undertaken an interim mini-computer solution for the low resolution spectrometer (Finnigan 1015 quadrupole) which is primarily used for our body fluid analysis studies. For the same reasons outlined above, a mini-computer solution is attractive. In the case of the low resolution quadrupole instrument, a lesser capacity machine will suffice for immediate data acquisition and display functions. We have implemented such an interim system on a PDP-11/20 machine available from other funding sources. This system, which is now operational, allows the acquisition of GC/MS data, limited by the capacity of the DEC tape storage medium to approximately 600 spectra, per experiment. For certain types of GC analyses, up to 1000 spectra per experiment are required so this limits, to some extent, the utility of this interim system. A calcomp plotter is supported for display purposes. A fixed head disk provides for library search procedures which are still being converted from the ACME system. We have applied to the NIH-GMS for funds to augment this system in order to relieve current limitations as part of a Genetics Center research proposal.

FUTURE PLANS

Our future plans are basically to continue development along the lines outlined above. We will complete the computing support transition steps described. These include primarily establishing a connection to the new 370/158 facility to provide interim support for the high resolution system. We will pursue additional software and hardware development goals as far as possible within the limited budget available. These efforts will concentrate for the most part on bringing up a metastable ion analysis data system. It should be reemphasized that the manpower levels proposed in the follow-on budget have been minimized to allow for purchasing computing time on the 370/158. The allocated manpower is required primarily for instrument operation and maintenance with minimal provision for development efforts.

Part B(ii). Analysis of Body Fluids by Gas Chromatography/Mass Spectrometry.

The chemical separation of urine into the following fractions prior to GC/MS analysis has been described in previous DENDRAL Reports:

free acids (analyzed by gc/ms as their methyl esters)
amino acids (analyzed by gc/ms as their N-trifluoroacetate n-butyl ester)
carbohydrates (analyzed by gc/ms as their trimethyl silyl ether derivatives)
hydrolyzed acids (analyzed by gc/ms as their methyl esters)
hydrolyzed amino acids (analyzed by gc/ms as their N-trifluoroacetate n-butyl esters)

During the past year we have extended these methods of fractionation to the following body fluids: blood (after an initial precipitation of proteins by the addition of ethanol) and amniotic and cerebrospinal fluids. The following summarizes the results obtained from an analysis of these fluids during the past year by gas chromatography-mass spectrometry.

URINE ANALYSIS:

A. The Development of a "Metabolic" Profile Characteristic of Neonatal Tyrosenemia Using Combined Gas Chromatography-Mass Spectrometry.

This work was carried out in collaboration with clinical colleagues from the Department of Pediatrics at Stanford University and a joint publication describing this research is in preparation.

The study was based on a total of one hundred and four 24-hour urine samples from sixteen premature or small birthweight infants receiving treatment in the Stanford nursery. After exclusion of infants who became ill, died, or left the nursery, we were able to follow nine infants closely for periods of between 4 and 6 weeks from day 3 of life. All nine infants had birthweights of below 1500g and three of these were below 1000g.

Of the nine infants studied, five showed transient tyrosinemia as shown by a marked elevation in the urinary excretion of the tyrosine metabolites, p-hydroxyphenyllactic acid, p-hydroxyphenylpyruvic acid and p-hydroxyphenylacetic acid. There was also a less marked but distinct elevation in the urinary tyrosine output. Figures 1 and 2 show the metabolic profiles of the same infant (J.L.) in the normal(a) and tyrosinemic(b) states. Figure 1 shows the free acid outputs, chromatographed as the methyl ester-methyl ether derivatives and Figure 2 is an expression of the free amino acids of the same urines, chromatographed as the N-trifluoroacetyl n-butyl ester derivatives. In each case the concentration of each metabolite is a function of the peak height as compared to the height of the internal standard. Table 1 is a summary of the ranges of urinary output of tyrosine and metabolites observed for all the infants in the study.

TABLE 1 Daily Excretion in mg/kg

Tyrosine	p-HPLactic	p-HPPyruvic	p-HPAcetic

Normal	0.2 - 3	0 - 5	0 - 0.5	0.2 - 2
Tyrosinemic	3 - 15	5 - 50	0.5 - 5	0.5 - 5

As shown by Table 1 and Figure 1 neonatal tyrosinemia is characterized by a very large increase in the output of p-hydroxyphenyllactic acid and by a 10-50 fold excess of the latter over p-hydroxyphenylpyruvic acid. Studies of the hereditary defects in tyrosine metabolism initially indicated that p-hydroxyphenylpyruvic acid was the major metabolite although more recently cases have been reported where p-hydroxyphenyllactic is in a 2-5 excess over p-hydroxyphenylpyruvic. These latter determinations were made using GC and GC/MS methods and therefore probably reflect the improved specificity of the analytical procedure (previously colormetric methods were used) rather than a difference of actual metabolic profile. Apart from the very large excess of p-hydroxyphenyllactic acid over its keto analog we could detect no significant differences between the profiles shown in neonatal tyrosinemia and those published for hereditary disease. Other metabolites such as p-hydroxymandelic acid, DOPA N-acetyltyrosine, which have previously been reported in tyrosinemic urine were not seen to be elevated.

B. GC/MS Analysis of Urine from Children Suffering from Leukemia.

This research was carried out with twenty 24-hour urine samples supplied by Drs. Jordan Wilbur and Tom Long of the Stanford Children's Hospital.

The acidic fraction of all urines studied in this project showed no abnormal metabolites nor were gross amounts of known acids detected. The amino acid fraction, however, of six of the urine samples showed the presence of an non-protein amino acid, beta-aminoisobutyric acid (BAIB). In several of these instances the patients were excreting in excess of 1 gram of BAIB per day. The literature contains many references to increased BAIB excretion (genetic excretors, lead poisoning, pulmonary tuberculosis, march hemoglobinuria, thalassaemia and Down's Syndrome). The reported excretion of BAIB by leukemic patients was not substantiated by another investigator. There are several criticisms in the literature of the methods used for the quantitation of BAIB in biological fluids and in order to fill this void a sensitive, specific and rapid method for the quantitation of BAIB has been developed. (SEE: The Quantitation of BAIB in Urine by Mass Fragmentography; W.E. Pereira, R. Summons, W.E. Reynolds, T.C. Rindfleisch and A.M. Duffield, in press).

C. GC/MS Analysis of Urine from Patients Suffering from Hodgkin's Disease.

During this study 20 urine samples from patients with diagnosed Hodgkin's Disease (Department of Oncology, Stanford University Medical Center) were analyzed and in general, no abnormal metabolic profile could be found in any urine. There was one exception in which an individual was noted to excrete massive quantities of adipic acid (of the order of 1 gram per day).

D. Detection of Metabolic Errors by GC/MS Analysis of Body Fluids.

This project results from a collaborative effort between the Departments of Genetics and Pediatrics of the Stanford University Medical Center. To date over 50 samples have been analyzed; the majority (35) being

urine, while amniotic fluid (10), blood (6), and cerebrospinal fluid (6) were also analyzed. It has been and will continue to be our practice to analyze aliquots of fluid samples in collaboration with clinical investigators obtained for valid diagnostic purposes completely divorced from this research on GC/MS analysis techniques. This investigation is not intended to serve as a screening program for a large population but rather to focus on those individuals who exhibit suggestive clinical manifestations such as psychomotor retardation and progressive neurologic disease as well as suggestive pedigrees.

In the case of amniotic fluid the hope is to be able to monitor the condition of the fetus in those pregnancies which might be considered at risk. To date we have investigated specimens from normal pregnancies in order to establish the catalog of compounds to be observed in amniotic fluid. From this base it could prove possible to identify materials which might identify the health of the fetus.

We have been able to confirm the presence of orotic acid in a urine from a person found to have orotic aciduria while another urine sample was used to demonstrate our ability to identify the characteristic metabolites present in isovaleric acidemia. The following description refers to a urine from a child with hypophosphatasia.

A child died 33 hours after birth in Fresno, California, with the classical signs of hypophosphatasia. This genetic defect is marked by high phosphoethanolamine (PEA) concentrations in urine of affected homozygotes and unaffected heterozygotes. After derivatization (in this instance the TMS ethers of the water soluble carbohydrate fraction were prepared) we were able to detect by GC/MS large concentrations of ethanolamine and phosphoric acid but not PEA itself. The derivatization procedure we used most likely hydrolyzed PEA. We were able to quantitate for this compound in the infant's urine using an amino acid analyzer, and PEA excretion was extremely high (over 200 times normal values for infants) confirming the diagnosis. Next we examined urine samples from the child's parents, presumed heterozygotes, by GC/MS and by the amino acid analyzer. Again, no PEA was detected by the former method although the presence of ethanolamine and phosphoric acid was demonstrated. We determined the following excretion levels of PEA by amino acid analyzer:

Newborn infant:	94 micromoles per 100 ml.	(Normal 0.21-0.33)
Father:	269 micromoles per 24 hours	(normal 17-99)
Mother:	32 micromoles per 24 hours	(normal 17-99)

It is of interest that in this family the affected infant and his unaffected father both show subnormal serum alkaline phosphatase activity. The mother, who did not excrete increased amounts of PEA, was found to have normal activity of this enzyme in her serum. The following table summarizes the serum phosphatase activity measurements:

Newborn infant:	0.2 units*	(normal 2.8-6.7)
Father:	0.7 units	(normal 0.8-2.3)
Mother:	3.4 units	(normal 0.8-2.3)

(* - 1 unit is that phosphatase activity which will liberate 1 millimole of p-nitrophenol per hour per liter of serum)

E. Drug Analysis Service Using GC/MS

We were recently contacted by physicians to rapidly identify a drug self-administered by a patient in the Stanford University Hospital. From the mass spectrum the drug was identified as pentazocaine within the hour. Although not part of the formal DENDRAL proposal we expect that similar cases may arise in the future and we intend to respond positively to such requests.

Development of Library Search Routines for Mass Spectrum Identification

The analysis of a single body fluid fraction produces between 600 and 750 mass spectra. In order to cope with the interpretation of the daily production of mass specta (about 8 body fluid fractions for a total of between 4,800 and 6,000 mass spectra) we have begun the implementation of library search routines. Concurrent with the analysis of body fluids for metabolic content we have been recording the mass spectra of many reference compounds. This collection represents the beginning of the construction of a library of reference spectra. Late in 1973 we expect to receive from Dr. S. Markey, University of Colorado Medical Center, a more comprehensive library which he has collated from contributors (including our own laboratory) in the field of biological applications of gas chromatography/mass spectrometry.

Prior to the demise of the ACME computer faciliity at Stanford University, we ran library search routines on data collected from urine fractions. Because of the ACME system being heavily loaded, our programs took about one minute per compound identification. However, the experience gained will be used to implement library search routines on our current PDP-11 GC/MS data system. In addition we have sent mass spectra from several urine analyses to Dr. S. Grotch, Jet Propulsion Laboratory, Pasadena, California, in order that he could use his library search routines on real data. In this instance the limiting factor for efficient compound identification was the library content which was limited to a few compounds of biological significance. In addition those compounds of interest that were present in the library were often in a derivatized form different from that used in our analytical methodology.

Application of GC/HRMS to Body Fluid Analysis

We reported in the last annual report of the DENDRAL project that the Varian MAT 711 mass spectrometer was interfaced with a gas chromatograph for the recording of low resolution mass spectra. We have now used this system for the recording of HRMS of gas chromatographic fractions from urine analyses. We were able to record HRMS scans over several gas chromatographic peaks of interest in a number of urine fractions. The high resolution results were found to be of a high quality in mass measurement accuracy. When using the MAT 711 instrument for GC/HRMS the sensitivity of the ion source was a limiting factor in that less intense gas chromatographic peaks often lacked sufficient material to generate acceptable high resolution mass spectra. Notwithstanding this limitation the HRMS data recorded on different urine fractions was used to confirm the identification of several metabolites. If by chance the metabolite of concern was available only in quantities insufficient for direct GC/HRMS, preparative GC would be used to concentrate the component of interest for subsequent HRMS.

RESOURCE OPERATION

Over the term of this grant our mass spectrometry laboratories have provided support to numerous research projects in addition to the DENDRAL computer program development project funded under this grant. These cover a variety of applications at Stanford, in the United States, and abroad. Included are problems in the study of human metabolites, biochemistry, and natural product chemistry. Samples have been run in collaboration with outside people both on the MAT-711 GC/High Resolution Mass Spectrometer system and the Finnigan 1015 GC/Low Resolution Quadrupole Mass Spectrometer system. The low resolution system has also been supported by a NASA research grant.

The following tables summarize the support rendered in terms of numbers of samples run through various types of analysis:

I. MAT-711 High Resolution System (Period covered 11/71 - 6/73).

	Batch High Resol. MS	Batch Low Resol. MS	GC/High Resol. MS	GC/Low Resol. MS
DENDRAL program devel.	317	3		
Stanford Genetics (Body fluid analysis)	39	17	13	
Stanford Chemistry (non-DENDRAL - Dr. Djerassi's group)	91	112		50
Stanford Chemistry (non-DENDRAL - Drs. Vantamelen, Johnson, Mosher, Collman, Altman, Goldstein)	29	23		4
Stanford Surgery (Dr. Fair)	8			
Dr. Adlerkreutz (Finland)	10			
Dr. Venien (France)	26			
Dr. Gilbert, Mors, Baker (Brazil)	40	44		
Dr. Orazi (Argentina)	19	1		
Dr. Subramanian (India)	10	5		
Dr. Khastgir (India)		5		
Dr. O'Sullivan (Ireland)		5		
Dr. Badr (Libya)	30			
Dr. Mital (India)	5			

624 samples	215 samples	13 samples	54 samples

II) FINNIGAN 1015 Low Resolution System (period covered 8/72-8/73)

Note the samples run are specified by fluid type. Each fluid is extracted and derivatized as described in Part B (ii) and therefore may represent several GC/LRMS analyses. Specific discussions of the results of various of the analyses run are discussed earlier in Part B(ii).

	GC/Low Resolution MS
Stanford Pediatrics (Drs. Cann, Sunshine and Johnson)	141 urines
	7 Amniotic Fluids
	6 bloods
	2 cerebrospinal fluids
Stanford Oncology (Dr. Rosenberg)	20 urines
Stanford Psychiatry - Genetics (Drs. Brodie and Cavalli-Sforza)	4 cerebrospinal fluids
Stanford Respiratory Medicine (Dr. Robin)	2 urines
	2 bloods
Stanford Pharmacology (Dr. Kalman)	2 extracts
Stanford Biochemistry (Dr. Stark)	4 extracts
Stanford Children's Hospital (Drs. Wilbur and Long)	24 urines
UC San Francisco Medical School - Dermatology (Dr. Banda)	2 urines
Menlo Park V.A. Hospital (Dr. Forrest)	13 extracts
Palo Alto V.A. Hospital (Drs. Hollister and Green)	7 extracts
University of Puerto Rico School of Medicine (Dr. Garcia-Castro)	7 urines
	243 samples

PART B PUBLICATIONS

The following summarizes the publications resulting from research in the low resolution mass spectrometry laboratory over the past year, including body fluid analysis. This laboratory has been jointly supported by NIH (DENDRAL) and NASA. The listed publications include research relevant to both sponsors.

The Determination of Phenylalanine in Serum by Mass Fragmentography.
Clinical Biochem., 6 (1973)
By W.E. Pereira, V.A. Bacon, Y. Hoyano, R. Summons and A.M. Duffield.

The Simultaneous Quantitation of Ten Amino Acids in Soil Extracts by Mass Fragmentography
Anal. Biochem., 55, 236 (1973)
By. W.E. Pereira, Y. Hoyano, W.E. Reynolds, R.E. Summons and A.M. Duffield.

An Analysis of Twelve Amino Acids in Biological Fluids by Mass Fragmentography.
Anal. Chem.,
By R.E. Summons, W.E. Pereira, W.E. Reynolds, T.C. Rindfleisch and A.M. Duffield.

The Quantitation of B-Amino isobutyric Acid in Urine by Mass Fragmentography.
Clin. Chim. Acta, in press
By W.E. Pereira, R.E. Summons, W.E. Reynolds, T.C. Rindfleisch and A.M. Duffield.

The Determination of Ethanol in Blood and Urine by Mass Fragmentography.
Clin. Chim. Acta
By W.E. Pereira, R.E. Summons, T.C. Rindfleisch and A.M. Duffield.

A Study of the Electron Impact Fragmentation of Promazine Sulphoxide and Promazine using Specifically Deuterated Analogues.
Austral. J. Chem., 26, 325 (1973)
By M.D. Solomon, R. Summons, W. Pereira and A.M. Duffield.

Mass Spectrometry in Structural and Stereochemical Problems. CCXXXVII. Electron Impact Induced Hydrogen Losses and Migrations in Some Aromatic Amides
Org. Mass Spectry., in press.
By A.M. Duffield, G. DeMartino and C. Djerassi.

Spectrometrie de Masse. IX. Fragmentations Induites par Impact Electronique de Glycols- -En Serie Tetraline
Bull Soc. Chim. France, 2105 (1973).

Spectrometric de Masse VIII. Elimination d'can Induite par Impact Electronique dans Le Tetrahydro-1,2,3,4-Napthtal-ene-diol-1,2.
Org. Mass Spectre., 7, 357 (1973).
By P. Perros, J.P. Morizur, J. Kossanyi and A.M. Duffield.

Chlorination Studies I. The Reaction of Aqueous Hypochlorous Acid with Cytosine.
Biochem. Biophys. Res. Commun., 48, 880 (1972)
By W. Patton, V. Bacon, A.M. Duffield, B. Halpern, Y. Hoyano, W. Pereira and J. Lederberg.

Chlorination Studies II. The Reaction of Aqueous Hypochlorous Acid with -Amino Acids and Dipeptides.
Biochim. et Biophys. Acta, 313, 170 (1973).
By. W.E. Pereira, Y. Hoyano, R. Summons, V.A. Bacon and A.M. Duffield.

Chlorination Studies IV. The Reaction of Aqueous Hypochlorous Acid with Pyrimidine and Purine Bases.
Biochem. Biophys. Res. Commun., 53, 1195 (1973).
By Y. Hoyano, V. Bacon, R.E. Summons, W.E. Pereira, B. Halpern and A.M. Duffield.

Part C. EXTENSION OF THE THEORY OF MASS SPECTROMETRY BY COMPUTER

OBJECTIVES:

Part C of the DENDRAL effort, termed Meta-DENDRAL, aims at providing theory formation help for chemists interested in the mass spectrometric behavior of new classes of compounds. Our goals are necessarily long-range because theory formation by computer is itself an exciting, unsolved problem in computer science. We have chosen to explore this problem in the context of mass spectrometry in order to make frontier computer research results available to working scientists.

The problem of finding judgmental rules for use in a computer program is common to many biomedical computing projects, such as medical diagnosis and therapy recommendation programs. <See, for example, Shortliffe, et.al.> In order to give these programs the knowledge that makes them perform at acceptable levels, a medical expert is often asked to summarize his own knowledge of the problem area in rules that the program can use. The Meta-DENDRAL theory formation program is a paradigm for the kind of assistance that computers can give to the medical experts in this role. Programs of this sort can, first of all, provide the expert with an interpreted summary of a large collection of "hard" empirical data. Second, the program can suggest to the expert plausible rules that appear to explain major features of the data. Thus, the expert is able to assimilate large collections of data in the rules given to the computer. We believe that the meta-DENDRAL work is a useful model on which fruitful work in other biomedical problems can be based.

The over-all strategy of this research is to model the theory formation activity of scientists. We start with a set of empirical data which are known molecular structures and their associated mass spectra. By exploring the possible mechanistic explanations of each mass spectrum, the program is able to find a set of mechanisms that appear to be characteristic for the class of molecules. These characteristic mechanisms constitute the general mass spectrometry rules for the class, or a first-level theory for the class. Further refinements of the rules give more sophisticated restatements of the theory.

We have designed the programs in such a way as to provide useful results from the intermediate steps. The progress section discusses several sets of results that have been obtained, even though the entire program has not yet been completed.

PROGRESS:

In the past ten months (since January, 1973) the theory formation programs have seen significant application and significant new extensions. In addition, the work has been described in publications for both chemists and computer scientists.

Applications of Existing Programs.

The INTSUM program, for interpreting and summarizing the mass spectra of many known compounds of one class, was described in the previous annual report as essentially finished. In this last period we have used this program to help understand the mass spectrometry of several

classes of compounds, including estrogens, equilenins and other estrogenic steroids, androstanes, alkyl pregnanes, vinyl quinazalones, amino acids and aromatic acids. An article written for mass spectroscopists and soon to appear in Tetrahedron (Smith, et.al, enclosed) describes this program and its usefulness in understanding the previously unreported mass spectrometry of the equilenins. The amino acid and aromatic acid results are useful for interpreting the mass spectra taken from those fractions of urine (see Part B).

The INTSUM program is available to anyone who requests it, as stated in the article soon to appear. Because of the complexity of the program, we recommend that mass spectroscopists use this program on a network computer after they have collected a number of mass spectra from a class of compounds whose fragmentation mechanisms they wish to investigate.

Recent Extensions to Meta-DENDRAL.

In this last period significant progress has been made on the theory formation programs that use the interpreted summary of the data provided by the INTSUM program. A simple rule formation program, described previously (HI7), finds the characteristic mass spectrometry mechanisms for a class of compounds, assuming that the compounds exhibit regular behavior as a class. Recent work has removed the restriction that the compounds must behave as a class - important classes can be found by the program within the set of given compounds. The procedure was described in a paper for the Third International Joint Conference on Artificial Intelligence, which is enclosed. At the same time that the rule formation program looks for characteristic mechanisms, the class separation procedure refines the class of molecules that appear to behave uniformly (i.e., appear to exhibit most of the characteristic mechanisms).

Another important extension of the theory formation program makes the rule descriptions more general and less specific to the class of compounds studied. The mechanisms in the rules are now described generally in terms of the kinds of bonds that break, and not in terms of the precise relations of the bonds to the skeletal structure common to the class. For example, a rule is now stated as "Any bond that is the second bond from a nitrogen atom is likely to break", rather than "In the skeleton R1-C2-N3-C4-R5 the bond between atoms 1 and 2 and the bond between atoms 4 and 5 are both likely to break".

These general descriptions will allow much more freedom in the kinds of interpretations that can be placed on the INTSUM results. It is possible, for example, to alter the set of predicates used to describe bonds without altering the program.

The program can be conceptualized as a search program through the space of possible combinations of predicates. Some predicates describe the type of bond (e.g., 'single'), others describe the atoms joined by the bond (e.g., 'nitrogen', 'secondary'), and others describe the bonds and atoms next away from the bond that breaks. Some a priori heuristics limit consideration of complex predicates to chemically meaningful combinations, for example, by forbidding consideration of a single atom as both carbon and nitrogen. Other heuristics guide the process of expansion by forbidding a new predicate to be added to a description if its addition reduces the explanatory power of the existing description. For example, if a high average intensity is associated with breaking the