

SECTION I

Form Approved
Budget Bureau No. 68-R0DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE

GRANT APPLICATION

LEAVE BLANK

TYPE	PROGRAM	NUMBER
REVIEW GROUP		FORMERLY
COUNCIL (Month, Year)		DATE RECEIVED


TO BE COMPLETED BY PRINCIPAL INVESTIGATOR (Items 1 through 7 and 16A)

1. TITLE OF PROPOSAL (Do not exceed 63 typewriter spaces) Resource Related Research - Computers and Chemistry (RR-00612 renewal)	
2. PRINCIPAL INVESTIGATOR	
2A. NAME (Last, First, Initial) Lederberg, Joshua	3. DATES OF ENTIRE PROPOSED PROJECT PERIOD (This applicable) FROM 5/1/74 THROUGH 4/30/77
2B. TITLE OF POSITION Professor and Chairman	4. TOTAL DIRECT COSTS REQUESTED FOR PERIOD IN ITEM 3. \$1,639,456
2C. MAILING ADDRESS (Street, City, State, Zip Code) Department of Genetics Stanford University Medical Center Stanford, California 94305	5. DIRECT COSTS REQUESTED FOR FIRST 12-MONTH PERIOD \$488,267
2D. DEGREE Ph.D.	6. PERFORMANCE SITE(S) (See Instructions) Department of Genetics Department of Chemistry, and Department of Computer Science Stanford University
2E. SOCIAL SECURITY NO. [REDACTED]	
2F. TELEPHONE DATA Area Code 415 TELEPHONE NUMBER 321-1200 EXTENSION 5801	
2G. DEPARTMENT, SERVICE, LABORATORY OR EQUIVALENT (See Instructions) Department of Genetics	
2H. MAJOR SUBDIVISION (See Instructions) School of Medicine	
7. Research Involving Human Subjects (See Instructions) A. <input type="checkbox"/> NO B. <input type="checkbox"/> YES Approved: _____ C. <input checked="" type="checkbox"/> YES - Pending Review _____ Date _____	8. Inventions (Renewal Applicants Only - See Instructions) A. <input checked="" type="checkbox"/> NO B. <input type="checkbox"/> YES - Not previously reported C. <input type="checkbox"/> YES - Previously reported

TO BE COMPLETED BY RESPONSIBLE ADMINISTRATIVE AUTHORITY (Items 8 through 13 and 15B)

9. APPLICANT ORGANIZATION(S) (See Instructions) Stanford University Stanford, California 94305 IRS No. 94-1156365 Congressional District No. 17	11. TYPE OF ORGANIZATION (Check applicable item) <input type="checkbox"/> FEDERAL <input type="checkbox"/> STATE <input type="checkbox"/> LOCAL <input checked="" type="checkbox"/> OTHER (Specify) Private, non-profit University
10. NAME, TITLE, AND TELEPHONE NUMBER OF OFFICIAL(S) SIGNING FOR APPLICANT ORGANIZATION(S) c/o Sponsored Projects Office Telephone Number (s) (415) 321-2300 X2883	12. NAME, TITLE, ADDRESS, AND TELEPHONE NUMBER OF OFFICIAL IN BUSINESS OFFICE WHO SHOULD ALSO BE NOTIFIED IF AN AWARD IS MADE K. D. Creighton Deputy Vice Pres. for Business and Finance Stanford University Stanford, California 94305 Telephone Number (415) 321-2300 X21
	13. IDENTIFY ORGANIZATIONAL COMPONENT TO RECEIVE CREDIT FOR INSTITUTIONAL GRANT PURPOSES (See Instructions) 01 School of Medicine
	14. ENTITY NUMBER (Formerly PHS Account Number) 458210

15. 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.

SIGNATURES (Signatures required on original copy only. Use ink, "Per" signatures not acceptable)	A. SIGNATURE OF PERSON NAMED IN ITEM 2A 	DATE APR 26 1973
	B. SIGNATURE(S) OF PERSON(S) NAMED IN ITEM 10	DATE

SECTION 1

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE

LEAVE BLANK

PROJECT NUMBER

RESEARCH OBJECTIVES

NAME AND ADDRESS OF APPLICANT ORGANIZATION

Stanford University, Stanford, California 94305

NAME, SOCIAL SECURITY NUMBER, OFFICIAL TITLE, AND DEPARTMENT OF ALL PROFESSIONAL PERSONNEL ENGAGED ON PROJECT, BEGINNING WITH PRINCIPAL INVESTIGATOR

Lederberg, Joshua,	[REDACTED]	Professor of Genetics, Department of Genetics
Djerassi, Carl,	[REDACTED]	Professor of Chemistry, Department of Chemistry
Feigenbaum, Edward,	[REDACTED]	Professor of Computer Science, Dept. of Computer Sci
Buchanan, Bruce,	[REDACTED]	Research Computer Scientist, Dept. of Computer Scien
Duffield, Alan,	[REDACTED]	Research Associate, Department of Genetics
Pereira, Wilfred,	[REDACTED]	Research Associate, Department of Genetics
Rindfleisch, Thomas	[REDACTED]	Research Associate, Department of Genetics
Smith, Dennis,	[REDACTED]	Research Associate, Department of Chemistry
Sridharan, Natesa,	0	Research Associate, Department of Computer Science
Hammerum, Steen	0	Research Associate, Department of Chemistry

TITLE OF PROJECT

Resource Related Research - Computers and Chemistry

USE THIS SPACE TO ABSTRACT YOUR PROPOSED RESEARCH. OUTLINE OBJECTIVES AND METHODS. UNDERSCORE THE KEY WORDS (NOT TO EXCEED 10) IN YOUR ABSTRACT.

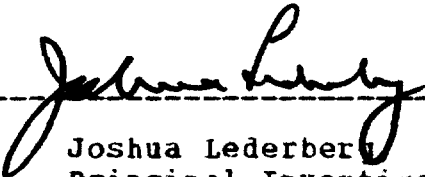
The objectives of this research program are the development of innovative computer and biochemical analysis techniques for application in medical research and closely related aspects of investigative patient care. We will apply the unique analytical capabilities of gas chromatography/mass spectrometry (GC/MS) and Carbon(13) Nuclear Magnetic Resonance Spectrometry (CMR) with the assistance of data interpreting computer programs utilizing artificial intelligence techniques, to investigate the chemical constituents of human body fluids in a variety of clinical contexts. Specific subtasks of this program include; 1) the application of artificial intelligence techniques to programs capable of interpreting mass spectra from basic principles as well as extending mass spectral theory by analysis of solved spectrum-structure examples, 2) the extension of GC/MS data systems incorporating an increasing level of automation and allowing the directed collection of specialized information, 3) the application of GC/MS techniques to analyze body fluids such as urine and blood, and to relate detected metabolic abnormalities to clinically observable disease states, and 4) the application of CMR techniques to assist in the determination of chemical structure.

LEAVE BLANK

The undersigned agrees to accept responsibility for the scientific and technical conduct of this project and for provision of required progress reports if a grant is awarded as the result of this application.

APR 26 1973

Date


Joshua Lederberg
Principal Investigator

RESOURCE-RELATED RESEARCH: COMPUTERS AND CHEMISTRY

(RR-00612 - Renewal Application)

TABLE OF CONTENTS

Introduction	P-1
Part A: Applications of Artificial Intelligence to Mass Spectrometry	P-5
Part B(i): Mass Spectrometer Data System Development	P-17
Part B(ii): Analysis of the Chemical Constituents of Body Fluids	P-28
Part C: Extension of the Theory of Mass Spectrometry by Computer	P-39
Part D: Applications of Carbon(13) Nuclear Magnetic Resonance Spectrometry to Assist in Chemical Structure Determination	P-45
Significance	P-82
Collaborative Arrangements	P-84
Facilities Available	P-86
Human Subjects	P-88
Budgets and Justification	P-90
Biographies	P-117

INTRODUCTION

INTRODUCTION

This proposal seeks a three year extension of our existing grant for Resource Related Research - Computers and Chemistry (RR-00612). Over the two years we have been supported by this grant we have made significant progress in all of the areas we initially proposed including clinical applications of body fluid analysis by gas chromatography/mass spectrometry (GC/MS), extensions to automate our GC/MS instrumentation and data systems, and the development of programs which, in specific areas, match human performance in interpreting mass spectra from first principles as well as extend mass spectral theory to new classes of compounds. Our success to date reinforces our expectations that this research will have a significant and useful impact on medical research involving studies of human biochemistry. As discussed in section B(ii) of this proposal, we have bolstered contact with real clinical problems through the Department of Pediatrics (Professor Howard Cann). We have recently encountered preliminary correlations between the amount of beta-amino isobutyric acid present in the urine of children with lymphoblastic leukemia and the state of their disease; and also between a defect in phenylalanine-tyrosine metabolism and late metabolic acidosis in premature infants.

This project is highly interdisciplinary, merging the interests of Professors Lederberg (Genetics), Djerassi (Chemistry), and Feigenbaum (Computer Science), in evolving and applying mass spectrometry as an analytical tool in medicine and in modeling aspects of scientific problem solving processes. Mass spectrometry is an ideal domain for this collaboration. On the one hand it has special importance to medical science and organic chemistry as a remarkably sensitive and analytically precise physical method for studying human biochemistry at the molecular level. On the other hand, the problems of mass spectrum interpretation are at once sufficiently complex to challenge the human intellect and sufficiently structured to be dealt with by current computer programming concepts. It is thus a rich, real-world problem domain in which to study the emulation of lower level cognitive functions, knowledge representations, and theory formation processes.

This combination of interdisciplinary interests promises both near and long term returns for the research investment. As indicated above, even with relatively crudely automated systems, a significant impact can be made on relevant medical problems. In the longer term the increasing load of body fluid analyses, which will have to be performed to be responsive to clinical needs, will require unburdening chemists from the laborious processes of reducing and interpreting the large volumes of data involved. These problems are squarely addressed by the proposed use of stored libraries of solved spectra, augmented by computer programs to extend such catalogs by "cognitive" insight.

This proposal is organized in a manner similar to the original in that the overall goals are divided into a number of subtasks. These comprise the original subtask definitions as well as one additional task proposed to explore the use of Carbon (13) nuclear magnetic resonance information as a potentially useful adjunct to mass spectral information to limit the space of candidate molecular structures. The respective proposal subtasks elaborated upon in subsequent sections include:

Part A: Applications of Artificial Intelligence to Mass Spectrometry

Part B(i): Mass Spectrometer Data System Development

Part B(ii): Analysis of the Chemical Constituents of Body Fluids

Part C: Extending the Theory of Mass Spectrometry by Computer

Part D: Applications of Carbon (13) Nuclear Magnetic Resonance Spectrometry to Assist Chemical Structure Determination

This proposal is related to several others pending, in progress, or terminating:

1) SUMEX (NIH: RR-00785, pending - Principal Investigator, J. Lederberg)-- This proposal seeks to establish a computer resource for the application of artificial intelligence in medicine as well as for the exploration of GC/MS as a tool for biomolecular characterization. The present renewal application is subsumed in the SUMEX application but is submitted independently to meet NIH renewal application deadlines which predate National Advisory Research Resources Council consideration of the SUMEX proposal. Should SUMEX be approved, this proposal will be withdrawn. Should SUMEX not be approved, this proposal seeks to continue support of our current mass spectrometry research efforts.

2) Genetics Research Center (NIH: pending - Principal Investigator, J. Lederberg)-- This proposal seeks to establish a Genetics Research Center at Stanford for research in medical genetics and the application of such research to clinical aspects of medical genetics. This proposal incorporates a significant level of cooperation between the Departments of Genetics and Pediatrics at Stanford including clinical applications of GC/MS. The Genetics Center proposal complements the present renewal application in that it concentrates on research aspects of genetic disease whereas this proposal attacks basic problems of methodology as well as developmental aspects of applying GC/MS analyses of metabolic disorders as indicators of disease states in a broader

context.

3) ACME (NIH: RR-00311, terminating, July 1973, - Principal Investigator, J. Lederberg)-- The ACME computing resource has been our major source of computing support for the reduction and analysis of mass spectral data. This support has been provided as a part of the ACME core research program without an explicit transfer of funds from the DENDRAL project. With the termination of NIH support, the ACME facility will be combined with other Medical Center computing functions on a fee-for-service basis, thereby introducing a new specific item in our budget to cover these computer costs.

4) Heuristic Programming Research in Artificial Intelligence (Advanced Research Projects Agency (ARPA): SD-183, in progress - Co-Principal Investigators, E. Feigenbaum and J. Lederberg)--This on-going research effort complements the present proposal by supporting those aspects of artificial intelligence concept and program development not directly related to medical problem areas. The present NIH-supported project benefits from this research and acts to enable the transfer of these ideas into a medically relevant context.

The current resource grant is headed by Professor E. Feigenbaum as Principal Investigator. He will shortly take a leave of absence for two years to accept the post of Deputy Director of the Information Processing Techniques Office of ARPA. During his absence, Professor Lederberg will act as Principal Investigator of the research project. Whereas Professor Feigenbaum will formally not be a member of the project during his tenure with ARPA, he will maintain his office locally, enabling him to maintain close intellectual contact with our research effort.

PART A:

APPLICATIONS OF ARTIFICIAL INTELLIGENCE
TO MASS SPECTROMETRY

Part A. Applications of Artificial Intelligence to Mass Spectrometry

OBJECTIVES:

The overall objective of part A of this proposal is to extend the reasoning power of Heuristic DENDRAL. Mass spectrometry was initially chosen as the task area in which to explore the techniques of heuristic programming for molecular structure elucidation. Much of the past and proposed future efforts will remain directed strongly to analysis of mass spectra because of the sensitivity and specificity of the technique. It is clear, however, that information available from other spectroscopic techniques, utilized routinely by chemists when sample quantities are sufficient, can and should be used where appropriate to obtain structural information which cannot be provided by mass spectrometry alone. This point is elaborated in the subsequent discussion of progress and plans.

A corollary of the overall objective is to tie the Heuristic DENDRAL program very closely to the requirements of the chemical studies outlined below (analysis of steroids from body fluids) and in Part B of the proposal (analysis of chemical constituents of urine, blood, and other body fluids). We have previously directed and will continue to direct our studies toward classes of biologically relevant molecules. Thus we have the capability of providing significant support to the chemically oriented activities as the capabilities of Heuristic DENDRAL are extended.

The overall objective encompasses several sub-tasks, outlined below, all of which represent critical steps in building a powerful program in an incremental fashion. This approach provides an operational program which can be used by chemists in a routine production mode, while extensions of the program are under development. The sub-tasks are the following:

A) Extend Heuristic DENDRAL to analysis of the mass spectra of complex molecules. This includes the assessment of the capabilities and limitations of the program in analysis of unknown compounds or mixtures of compounds. It also includes refinement of planning rules which infer compound class or molecular substructure, both being extremely important in subsequent analysis of a mass spectrum.

B) Develop the Cyclic Structure Generator to provide DENDRAL with the capabilities for generation of all isomers of a given empirical formula. Define and incorporate constraints on the generator to exclude implausible isomers. Enlarge the capacity of the cyclic generator to accept constraints of demanded or forbidden substructures (GOODLIST, BADLIST).

C) Develop the ability to incorporate information available from ancillary mass spectrometric techniques (e.g., metastable ion data, low ionizing voltage data, isotopic labelling) and other spectroscopic data (e.g., substructures from NMR) into the existing Heuristic DENDRAL program.

D) Extend the Predictor, now capable of prediction of mass spectra for limited classes of molecules, to the design of experimental strategies. Given a set of data, and partial or ambiguous structural information based on these data, specify additional experiments which may be done to effect a unique solution or minimize ambiguities.

PROGRESS:

We have, in the past two years of the existing DENDRAL grant, made significant progress in each of the areas outlined above. We feel that in some areas the progress has been particularly exciting, for example, the completion of the program for analysis of the mass spectra of complex molecules, and completion of the cyclic structure generator (unconstrained). The following represents a brief outline of accomplishments to date, keyed to the objectives A-D above.

A) Extension of Heuristic DENDRAL

Extension of Heuristic DENDRAL to the mass spectra of complex molecules dictated two important modifications in the approach used successfully for saturated, aliphatic, monofunctional (SAM) compounds. To reduce ambiguities of elemental composition inherent in low resolution mass spectra, the decision was made to extend the program to handle high resolution mass spectral data which specify the empirical composition of every ion. Although the basic strategy of Heuristic DENDRAL (plan, generate and test) was maintained, the absence of a cyclic structure generator at the time the program was written dictated that the basic skeleton, common to the class of molecules analyzed, be specified. The techniques of artificial intelligence have now been applied successfully to a problem of direct biological relevance, namely, the analysis of the high resolution mass spectra of estrogenic steroids. The performance of this program has been shown to compare favorably with the performance of trained mass spectroscopists, see Smith, et.al. (1972). The operation of this program has been detailed in this publication, a copy of which is attached. Briefly, the program was designed to emulate the thought processes of an expert as far as possible. High resolution mass spectral data are searched for evidence indicating possible substituent placements about the estrogen skeleton. Molecular structures allowed by the mass spectral data are tested against chemical constraints, and candidate solutions are proposed. Further details of the performance in analysis of more than thirty estrogen-related derivatives are presented in the above publication.

Of particular significance in this effort were, in addition to exceptional performance, the potential for analysis of mixtures of estrogens WITHOUT PRIOR SEPARATION, and for generalization of the programming approach to other classes of molecules.

Because of the structure of the Heuristic DENDRAL program it

is immaterial whether the spectrum to be analyzed is derived from a single compound or a mixture of compounds. Each component is analyzed, in terms of molecular structure, in turn, independently of the other components. This facility, if successful in practice, would represent a significant advance of the technique of mass spectrometry. Many problem areas, because of physical characteristics of samples or limited sample quantities, could be successfully approached utilizing the spectra of the unseparated mixtures. Even in combined gas chromatography/mass spectrometry (GC/MS), many overlapping peaks will be unresolved and an analysis program must be capable of dealing with these mixtures.

In collaboration with Prof. H. Adlercreutz of the University of Helsinki, we have recently completed a series of analyses of various fractions of estrogens extracted from body fluids. These fractions (analyzed by us as unknowns) were found to contain between one and four major components, and structural analysis of each major component was carried out successfully by the above program. These mixtures were analyzed as unseparated, underivatized compounds. The implications of this success are considerable. Many compounds isolated from body fluids are present in very small amounts and complete separation of the compounds of interest from the many hundreds of other compounds is difficult, time-consuming and prone to result in sample loss and contamination. We have found in this study that mixtures of limited complexity, which are difficult to analyze by conventional GC/MS techniques without derivatization (which frequently makes structural analysis more difficult), can be rationalized even in the presence of significant amounts of impurities. A manuscript on this study has been submitted to the Journal of the American Chemical Society

In the past year we have extended our library of high resolution mass spectra of estrogens to include 67 compounds. These data represent an important resource and have been included (as low resolution spectra for the moment) in a collection of mass spectra of biologically important molecules being organized by Prof. S. Markey at the University of Colorado. These data have been used extensively in developing the program strategies for Meta-DENDRAL (see Part C, below).

The Heuristic DENDRAL program for complex molecules has received considerable attention during the last year in order to generalize it from its previous emphasis on specific classes of compounds and program strategies. By removing information which is specific to estrogens, the program has become much more general. This effort has resulted in a production version of the program which is designed to allow the chemist to apply the program to the analysis of the high resolution mass spectrum of any molecule with a minimum of effort. Given the spectrum of a known or unknown compound, the chemist can supply the following kinds of information to guide analysis of the mass spectrum: a) Specifications of basic structure (superatom) common to the class of molecules. b) Specification of the fragmentation rules to be applied to the superatom, in the

form of bond cleavages, hydrogen transfers and charge placement. c) Special rules on the relative importance of the various fragments resulting from the above fragmentations. d) Threshold settings to prevent consideration of low intensity ions. e) Available metastable ion data and the way these data are subsequently used -- to establish definitive relationships between fragment ions and their respective molecular ions. f) Available low ionizing voltage data -- to aid the search for molecular ions. g) Results of deuterium exchange of labile hydrogens -- to specify the number of, e.g., -OH groups.

We have been very successful in testing the generality of the program, with particular emphasis on other classes of biologically important molecules. We have used the program in analysis of high resolution mass spectra of progesterone and some methylated analogs, a small number of androstane/testosterone related compounds, steroidal sapogenins and n-butyl-trifluoroacetyl derivatives of amino acids.

B) Cyclic Structure Generator

The cyclic structure generator has been completed after several years of effort under the continuing guidance of Professor Lederberg. The boundaries, scope and limitations of chemical structure can now be specified.

The cyclic structure generator now rests on a firm mathematical foundation such that we are confident of its thoroughness and ability to generate structures, prospectively avoiding duplicate structures. The prospective nature of the generator is a necessity for efficient implementation, as retrospective checking of each generated structure to eliminate redundancies is too time consuming. The necessary concepts have recently been transformed into an operating program. A manuscript describing the mathematical theory of the heart of the generator, the labelling algorithm, has been accepted by Discrete Mathematics (H. Brown, et.al., 1973). A companion manuscript describing the mathematical theory of the complete generator has been submitted (H. Brown and L. Masinter, 1973, submitted).

The cyclic structure generator in its entirety (encompassing acyclic and wholly cyclic structures and combinations thereof) will be described for chemists (L. Masinter et.al., in preparation). Apart from the labeling algorithm the remainder of the problem involves, first, the combinatorics of assignment of atoms to cycles or chains, and second, construction of acyclic radicals to attach to the rings using the well known principles of acyclic DENDRAL. A companion manuscript will soon be submitted describing for chemists the core of the cyclic structure generator, the labelling algorithm. This algorithm is capable of construction of all isomers, of wholly cyclic graphs, which may be formed by labelling the nodes of a cyclic skeleton with atoms (e.g., C, N, O) or labelling the atoms of the skeleton with substituents (e.g., -CH₃, -OH). Through the use of graph theory, and the symmetry-group

properties of cyclic graphs the labelling algorithm avoids construction of redundant isomers. It identifies equivalent node positions prospectively before labelling takes place. It is indicative of the precarious communication between chemists and mathematicians that it had remained unsolved (except for trivial simple cases) despite attention for over 100 years. As an indication of the complexity of chemistry in terms of numbers of possible structures, take the example of C_6H_6 . The most familiar molecule with this molecular formula is benzene. Yet there are 217 topological isomers for C_6H_6 (with valence constraints) of which only 15 are pure trees. The simple addition of one oxygen atom to the empirical formula of benzene, yielding C_6H_6O , yields 2237 isomers of the most familiar representative, phenol.

The first exercise of the generator has been to create a dictionary of carbocyclic skeletons. This time-consuming task would otherwise have to be done each time a new molecular formula is presented. The dictionary is structured to contain keys as to type of skeleton, number of rings, ring fusion, and so forth. The constraints which we wish to implement are then simple to exercise in the context of the dictionary.

C) Analysis Using Additional Data Sources

Several additional techniques are available to the mass spectroscopist other than recording the conventional mass spectrum. They provide complementary data which frequently are of great assistance in rationalization of the conventional spectrum, either in terms of structure or fragmentation mechanisms. We have designed the Heuristic DENDRAL program for complex molecules to use data from these additional techniques in much the same way as a chemist does. The following three types of data can now be used:

I) Metastable Ion (MI) Data. Metastable ions provide a means for relating fragment ions to molecular ions in a mass spectrum. This is important in two contexts. In examination of the spectrum of a known compound, the existence of a metastable ion provides strong evidence that a given fragment ion arises at least in part in a single decomposition process from an ion of higher mass (not necessarily the molecular ion). Investigations of this type are necessary to validate the fragmentation rules which guide the Heuristic DENDRAL program. (e.g., investigations of metastable ions of estrogens, Smith, Duffield and Djerassi, 1972).

The second context use is the analysis of mixtures of compounds to determine which fragment ions in a very complex spectrum are descended from which molecular parents. We have explored the analysis time and specificity of results as a function of the amount of metastable ion data available on a mixture. A 10 to 100-fold reduction in computer time is observed to arrive at single, correct solutions for various mixture components (rather than 5-20 possible solutions limited by the conventional mass spectrum alone). These results are reported in detail in the description on analysis of the estrogen mixtures (Smith, et.al., 1973

(submitted)).

Metastable ions are those which are formed by fragmentation processes occurring during the flight of an ion after formation and acceleration. These fragmentation processes may occur at any point along the flight path of ions through the mass spectrometer. Because of the complex behavior of metastable ions formed in magnetic or electric fields, they are usually studied in field-free regions. A conventional double focussing mass spectrometer possesses two field-free regions where metastable ions may be studied. One region lies between the electric sector and the magnetic sector. This region can be used to study so-called "normal" metastable ions, i.e., those metastable ions which are observed superimposed on the peaks in the conventional mass spectrum and which follow the relationship: observed mass of metastable ion = (mass of daughter)**2 / (mass of parent). The other field-free region lies between the ion source and the electric sector. Metastable ions formed in this region can be examined by de-tuning one analyzer of the instrument (defocussing). This procedure allows establishment of specific relationships between ions involved in a metastable decomposition so that the parent ion and its decomposition product, can both be identified. This technique has led to much more useful information for the Heuristic DENDRAL program, as illustrated earlier in this section.

II) Low Ionizing Voltage (LV) Data. The key to successful operation of the Heuristic DENDRAL program is correct inference of the molecular ion(s) and molecular formula(e) in a given mass spectrum. In the past, metastable ion data were used to assist the program in correct identification of molecular ions. This procedure has now been supplemented, making the program cognizant of LV data. At lower ionizing voltages, molecular ions are formed with lesser amounts of excess internal energy. Most classes of molecules (those that display significant molecular ions) can be analyzed at a sufficiently low ionizing voltage such that only molecular ions are observed, as the internal energy is not sufficient to allow fragmentation. This technique was used extensively in the analysis of estrogen mixtures and the resulting data simplify the program's task of determining molecular ions.

III) Isotopic Labeling. We have previously described how isotopic labeling of labile hydrogens with deuterium aids analysis. For example, the last phase of the analysis of spectra of complex molecules involves several "chemical" checks on the validity of proposed structures. The knowledge of the number of hydroxyl groups can be a powerful filter to reject certain candidate structures (Smith, et.al., 1972).

There are many other kinds of data available to chemists engaged in structure elucidation. The details of chemical isolation and derivitization procedures may require that only certain types of functional groups are plausible. Spectroscopic data from other techniques (e.g., proton or C13 NMR, IR, UV) may be available for a particular unknown. We have designed the Heuristic DENDRAL program for complex molecules with these additional data in mind. Specific

plans for implementation of these data as constraints on Heuristic DENDRAL are described in the Plans section below. Certain chemical information, for example, the knowledge that aromatic hydroxy functionalities have been methylated, can already be included as a constraint.

D) Extension of the Predictor Programs

The function of the Predictor in Heuristic DENDRAL has been to evaluate candidate solutions (structures) by prediction of their mass spectra, based on empirical fragmentation rules, and comparison of predicted versus observed spectra. This has been extended to high resolution mass spectra of complex molecules. Performance has been tested on estrogenic steroids and steroidal sapogenins.

There are other aspects of prediction of behavior that we have incorporated and plan to incorporate in the Predictor. We can now predict a minimum series of metastable defocussing experiments necessary to differentiate among candidate structures resulting from analysis of a mass spectrum. Other efforts are discussed in the Plans section, below. This approach amounts to design of optimum experimental strategies to effect a solution or minimize ambiguities.

We have begun to explore ways in which to predict the mass spectral behavior of molecules without the need to resort to the classical method of determining many mass spectra followed by empirical generalization. Dr. Gilda Loew has been investigating extended Huckel molecular orbital theory in an attempt at qualitative prediction of bond strength. Initial efforts on estrone will shortly appear describing these results (G. Loew, et.al., 1973). Briefly, calculated net atomic charges appear to have little bearing on subsequent fragmentation of the molecule. Bond densities (which are related to bond strengths), however, provide some indication of which bonds are likely to undergo scission in the first step of a fragmentation process.

PLANS:

As in the previous section, research plans are keyed to the objectives A-D.

A) Extension of Heuristic DENDRAL

I) We will continue use of the present program in collaborative studies with Prof. Adlercreutz concerning estrogenic steroids from, e.g., pregnancy urines. Work to date has inspired a synthetic program at Stanford University to verify conclusions of the program with regard to new estrogen metabolites. The planning program will be used extensively in analysis of the synthetic products also. As the capability for analysis of the mass spectra of other classes of steroids is developed, we hope to extend this collaboration.

II) We feel we have achieved a high level of compound-class independence in our present program. As more classes are

analyzed we expect that further "cleanup" may be necessary, but easy to carry out.

III) We are presently accumulating a large number of high resolution mass spectra of pregnanes and androstanes. For example, the first step away from estrogen analysis was initially going to be to the analysis of pregnanes, another biologically important class of steroids. A review of the mass spectrometry literature, however, revealed a paucity of information on the mass spectral fragmentation behavior of these molecules. Without fragmentation rules we cannot proceed with spectral analysis. We have, therefore, collected the high resolution mass spectra of approximately 50 pregnane related compounds. The data interpretation program (see Part C of the proposal) will be used extensively to help elucidate the fragmentation mechanisms involved. This study has already achieved the result of clarifying, through the use of high resolution data, the interpretation of mass spectra of the small number of pregnanes reported in the literature which were recorded only under low resolution conditions. Peaks have been found which have elemental compositions different from those assigned by past studies. We will investigate the performance of the program in analysis of mass spectra of urine components (see Part B of the proposal), specifically amino acid and aromatic acid derivatives.

IV) The planning program itself is extremely useful in helping build a more powerful analytical program. As new compound classes are considered the planner will be used to validate fragmentation rules developed for the class, in conjunction with the data interpretation program (see Part C of the proposal). This inspires confidence for use of the program in analysis of the spectra of related, but unknown, compounds.

V) As development of a production version of the cyclic structure generator is continued, will incorporate it into the planner. This will yield a program which more closely emulates the method originally developed for SAM compounds.

VI) Efforts in analysis of mass spectra have to this point been relatively restricted in terms of the types of structures which may be considered. As our knowledge base and the scope of the program increase it is necessary to consider general planning rules. These rules are used in initial examination of a mass spectrum to determine which compound class might be represented so that subsequent analysis utilizes rules for that class. One approach was used successfully in the past analysis of saturated aliphatic monofunctional (SAM) compounds. For more general utility, however, other approaches must be considered. The following areas will be investigated:

a) How best to exploit a version of library matching procedures to ease the computational burden on DENDRAL when dealing with routine analyses of mixtures of compounds that have previously been at least partially characterized. In this way attention can be focused on those previously uncharacterized components. This aids planning in that

effective library matching procedures frequently provide hints as to molecular structure even when the correct spectrum is absent from the library.

b) Utilize ion series spectra (Smith, 1972), an extension of the planning procedure for SAM compounds, in conjunction with the specific information embodied in a high resolution mass spectrum, which yields not only formulae but the implicit number of rings plus double bonds; both items serve as powerful limitations on compound class.

B) Cyclic Structure Generator

The present cyclic structure generator was designed to operate without constraints initially, as it must be capable of exhaustive generation of isomer. The next step in its development will be to implement constraints on the generator so that greater flexibility is possible. For example, in many cases the chemistry of a situation dictates that certain structural types may be present, or that others must be absent. The generator will use this information as constraints. We have planned a set of constraints which are useful to the chemist, for example, numbers of rings as opposed to double bonds, ring sizes, ring fusions, and so forth, and have begun developing ways to incorporate these constraints without compromising the requirements for thoroughness and non-redundancy.

We feel that the cyclic structure generator has the potential of acting as the focal point for an interactive laboratory analytical tool in addition to being a powerful addition to the existing Heuristic DENDRAL program. Constrained by inferences obtained from data (such as MS, IR, etc.) and from chemical treatments, such a generator would, under control by the chemist, be a powerful proposer of an exhaustive set of candidate solutions based on available data. We will develop this concept further as we improve both our capabilities for inference from scientific data and our techniques for using the generator.

One of the more promising spectroscopic techniques which we will exploit is C-13 NMR (see Part D of the proposal). The amount of structure specific information available is extensive, and serves in many cases to complement information available from mass spectra. Although sample requirements for the technique are prohibitively high for many applications, there is little question that this situation will improve with time. The capabilities of the structure generator as an interactive tool, or within the framework of existing Heuristic DENDRAL, will be enhanced if additional structural information can be incorporated.

C) Analysis Using Additional Data Sources

Plans under this section, i.e., extending the ability of Heuristic DENDRAL to cope with additional kinds of information, are intimately integrated with the plans for the preceding sections. When the cyclic generator is coupled with the planning program, much of this information constrains the generator to include (or exclude) some

particular substructure or functionality. We can readily deal with ancillary mass spectral data, but significant work remains on the most efficient ways (or at what point in the analysis) best to utilize, e.g., metastable ion data. We will also explore the performance of the planner when information such as C13 NMR data are available in addition to mass spectral data (see Part D of this proposal).

D) Extension of the Predictor Programs

Continuing development of the Predictor itself may prove to be an extremely interesting artificial intelligence application to chemistry. The problem facing the Predictor is the same problem faced by the chemist when available data do not yield a solution, or yield many ambiguous solutions. What additional data are needed to reach a solution? The Predictor must be made cognizant of possible measurement techniques (e.g., metastable ion data and their meaning) and which of the techniques are required. Design of an experimental strategy for further investigation of a structure problem represents a crucial link between Heuristic DENDRAL and the chemist dealing with the problem. It has important implications for more closed-loop control of instrumentation as the requisite data could as well come directly from the instrument as from the chemist by manual techniques. Thus a mechanism would exist for exploring the possibilities of "intelligent" instrument control (see Part B of the proposal). Such "control" could be exercised in a manual mode where sample quantities permit. Given the output of desired information from the Predictor, the chemist can then gather the information.

The other aspect of the Predictor, mentioned under Progress, above, is the possibility of using computational techniques to study fragmentation probabilities using, for example, molecular orbital theory rather than time consuming empirical studies. The ability to predict features of mass spectra given only a molecular structure would be an important advance both within the context of Heuristic DENDRAL and for mass spectrometry and theoretical chemistry as a whole.

PUBLICATIONS -- PART A

D.H. Smith, B.G. Buchanan, R.S. Engelmores, A.M. Duffield, A. Yeo, E.A. Feigenbaum, J. Lederberg, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference VIII. An approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids", J. Amer. Chem. Soc., 94, 5962 (1972).

D.H. Smith, B.G. Buchanan, R.S. Engelmores, H. Adlercreutz and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference IX. Analysis of Mixtures Without Prior Separation as Illustrated for Estrogens". Submitted to the Journal of the American Chemical Society.

H. Brown, L. Masinter, and L. Hjelmeland, "Constructive Graph Labelling Using Double Cosets", Discrete Mathematics, in press.

H. Brown and L. Masinter, "An Algorithm for the Construction of the Graphs of Organic Molecules", Discrete Mathematics, submitted.

L. Masinter, et.al., "Applications of Artificial Intelligence for Chemical Inference: Exhaustive Generation of Cyclic and Acyclic Isomers" (to be submitted).

L. Masinter, et.al., "An Algorithm for Constructive Labelling of Graphs Applied to Labelling Molecular Skeletons", (to be submitted).

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 Transition of Deuterated Derivatives", Org. Mass Spectrom., in press.

G. Loew, D.H. Smith and M. Chadwick, "Application of Molecular Orbital Theory to the Interpretation of Mass Spectra. Prediction of Primary Fragmentation Sites of Organic Molecules", Org. Mass Spectrom., in press.

D.H. Smith, "A Compound Classifier Based on Computer Analysis of Low Resolution Mass Spectral Data. Geochemical and Environmental Applications", Anal. Chem., 44, 536 (1972).

PART B(i)

MASS SPECTROMETER DATA SYSTEM DEVELOPMENT

PART B-(i) MASS SPECTROMETER DATA SYSTEM DEVELOPMENT

OBJECTIVES:

The large volume of data which must be reduced and interpreted from each GC/MS analysis of a body fluid sample together with the increasing number of samples which must be processed to be responsive to clinical needs, point to more and more highly automated and reliable GC/MS systems. This portion of the proposal addresses the problems of developing and applying such automated systems from several points of view. First, we propose to investigate the integration of sophisticated computer analysis programs into data reduction, data interpretation, and instrument management functions in order to progressively relieve the chemist from manually performing these tasks. Second, we will maintain the daily operation of our GC/MS systems for the on-going investigation of clinical applications and the acquisition of data necessary for the development of automated interpretation programs.

Our overall objectives for automating GC/MS systems comprise a number of specific subgoals including a) implementing highly automated and reliable systems for the acquisition and reduction of low resolution, high resolution, and metastable mass spectral data; b) implementing a data system to support combined gas chromatography/high resolution mass spectrometry; c) automating the location and identification of constituents of body fluid extracts from gas chromatogram and mass spectrum information for the routine application of these techniques to clinical problems; and d) investigating the intelligent closed loop control of mass spectrometer systems in order to optimize the data acquired relative to the task of data interpretation.

PROGRESS:

During the two years of support by this grant we have made progress toward each of the subgoals outlined above. Specific accomplishments and problems we have encountered are summarized below.

a) MASS SPECTROMETER DATA SYSTEM AUTOMATION

Funded by this grant, we have acquired a Varian-MAT 111 high resolution mass spectrometer. This instrument was formally accepted on November 5, 1971. The instrument has been used routinely in all of its operating modes including low resolution (approximately 1,000), high resolution (approximately 10,000), ultra-high resolution (approximately 80,000) peak matching, low ionizing voltage, metastable defocussing, and GC/MS operation

both at low and high resolution. It has assumed the entire burden of high resolution work for the DENDRAL project. A number of problems have arisen involving aspects of instrument alignment and operation and the mechanical, vacuum and, electronic systems. Support from Varian for resolving these problems has gotten progressively less responsive so that we have taken on most of the burden of maintenance locally.

Concentrating initially on the MAT-711 spectrometer, we have made significant 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 as well as an offset at infinite time to account for residual magnetization affecting accuracy at low masses.

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. A more detailed discussion of computing problems and plans is given under "FUTURE PLANS".

C) AUTOMATED GC/MS DATA REDUCTION

The application of GC/MS techniques to clinical problems as described in Part B(ii) of this proposal 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 problem of identification is addressed by the library search and DENDRAL mass spectrum interpretation programs discussed in Part A of this proposal. 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 chromatogram" 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 chromatograms 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 and those described in Parts A and C represent important preliminary steps toward closed-loop control.

The task of collection of 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 extremely 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 effect a solution or 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.

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 without budgeted cost as part of the core research effort. 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 have examined two alternative computing configurations to meet this transition: a) a PDP-11/45 local computer system and b) hooking up to the fee-for-service ACME follow-on machine. The trade-offs are basically as follows. The ACME follow-on option requires an expansion of the existing PDP-11/20 computer memory and a new interface to accommodate the new planned small machine interface to the 370/158. The near term costs of this approach including estimated 370/158 machine usage costs are approximately equal to the capital outlay of the PDP-11/45 amortized over 1-2 years. These 370/158 usage costs continue on a year to year basis indefinitely whereas the PDP-11/45 costs decrease to maintenance and supply levels after purchase. The ACME follow-on option requires a minimum of reprogramming since the PL-ACME language will be maintained. The PDP-11/45 option will require significant reprogramming to convert from PL-ACME to FORTRAN and Assembly Language. Once the reprogramming has been accomplished, the PDP-11/45 offers advantages of real time availability and responsiveness.

Thus the differences between these approaches revolve around short term costs versus long term flexibility. In order to minimize the impact to on-going efforts we have based this plan on the use of the ACME follow-on 370/158. Our budget estimate incorporates the preliminary expected costs for this type of operation. The rate structure for this facility is still being evolved however, and adjustments may have to be made.

a) MASS SPECTROMETER DATA SYSTEM AUTOMATION

Future efforts will include the transition of the existing ACME-based system to the new ACME follow-on configuration. We will adapt the concepts developed already 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 ACME follow-on system transition to allow routine collection and filing of sequential spectra. We will exercise 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 metastable information 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 - PART B(i)

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

Rapid Calculation of Molecular Formulas from Mass Values

The calculation of molecular compositions consistent with a given range of mass values arises particularly in mass spectrometry. Although this can be a trivial exercise on the computer, it has been vexing to do by hand. Published tables, e.g., Beynon and Williams,¹ are bulky, and nevertheless cover a limited range of atom values. The values are also awkward to search, not having been sorted.

The following approach was designed for a desk calculator that ought to be available to any student. As it involves only a few additions and subtractions, it can—*horribilis dictu*—even be done by hand. Furthermore, it lends itself to real time implementation on small computers that lack high precision “divide” instructions in their hardware.

The basis of the calculation is the table, which is an ordered list of the mass numbers of the formulas for H from 0 to 10, N from 0 to 5, and O from 0 to 11. It contains only those compositions whose masses are an integral multiple of 12. Any number of C's may then be added as required.

The use of the table is best explained by a specific example, say $m = 259.09 \pm 0.001$.

Step 1. Since $259 \equiv 7$ modulo 12, 5 H's (5.03913) will be borrowed to give $m' = m + 5H = 264.129$. This is then divided into $m' = m_i + m_j$; $m_i = 264$ ($= 22 \times 12$); $m_j = 0.129 \pm 0.001$.

Step 2. The table is searched for entries that correspond to m_j and whose mass does not exceed m_i . (m_i is expressed as $m_i/12 = C$ -equivalent.) We find none in this cycle.

Step 3. We therefore remove 12 H's (12.0939) to give $m'' = m' - 12H = 252.035 \pm 0.001$. The table now has entries at 0.034 ($H_8N_4O_8$), 0.035 ($H_{10}NO_9$) and 0.036 ($H_6N_5O_5$). These will be completed in Step 4. 12 H's are again removed until m_j falls below -0.0498 , the bottom of the table. In our example, this occurs at the next cycle.

Step 4. The table entries are now completed as follows

			Add C's to make up m''	Adjust borrowed H's	Check mass (compare 259.0900 \pm 0.0010)	
34	0.034216	$H_8N_4O_8$	$m_i = C_{16}$	C_5	$C_5H_{15}N_4O_8$	259.089
35	0.035559	$H_{10}NO_9$	$m_i = C_{14}$	C_7	$C_7H_{17}NO_9$	259.090
36	0.036895	$H_6N_5O_5$	$m_i = C_{14}$	C_8	$C_8H_{13}N_5O_5$	259.092

Step 5. Various criteria of chemical plausibility can be used to filter the list. Since the valence rules allow H's to a maximum of $2 + 2C + N$, none of these compositions is oversaturated. $C_5H_{15}N_4O_8$ however has an odd number of H's and may therefore represent a free radical.

If wider ranges of hetero atoms are contemplated, adjustments of blocks of 6 N (84.01844) and 12 O (191.9389) can be applied repetitively in a fashion similar to Step 3 so long as the adjusted mass allows.

In fact $m'' = m - 6N - 7H = 168.017 \pm 0.001$ leads to $C_6H_{11}N_8O_4$, $m = 259.090$. Further, $m - 12N - 7H = 83.999 \pm 0.001$. We read this as $m_i = 84$; $m_j = -0.001$ and find two entries in the table: -0.000826 (H_6NO_{10}) and 0.000510 ($H_2N_2O_8$), whose m_i however > 84 .

The table is arranged so as to illustrate its use in a fast computer program. A linear array with 138 cells, indexed as shown, has entries that never slip more than one position away from the value of the index. The composition values can therefore be accessed by direct lookup, obviating a table search. A card deck version of the table is available on request from the author.

This compilation is a greatly shortened form of some tables that were published some time ago.²

This work has been supported in part 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 GM-00612-01).

¹ BEYNON, J. H., AND WILLIAMS, A. E., "Mass and Abundance Tables for use in Mass Spectrometry," Elsevier, Amsterdam, 1963.

² LEDERBERG, J., "Computation of Molecular Formulas for Mass Spectrometry," Holden-Day, San Francisco, 1964.

Table of Mass Fractions for all Combinations^a of H, N, O ($H \leq 10$ $N \leq 6$ $O \leq 11$)

Index	$m_f \times 10^6$	H	N	O	=C	Index	$m_f \times 10^6$	H	N	O	=C	Index	$m_f \times 10^6$	H	N	O	=C
-49	-49787	0	2	11	17	0	0	0	0	0	0	31	31537	10	3	11	9
-45	-45765	0	0	9	12	1	510	2	5	6	14	32	32363	4	2	1	14
-38	-38554	0	4	10	18	2	1853	4	2	7	12	34	34216	8	4	8	16
-37	-37211	2	1	11	16	4	4532	2	3	4	9	35	35559	10	1	9	14
-34	-34532	0	2	8	13	5	5875	4	0	5	7	36	36895	6	5	5	13
-30	-30510	0	0	6	8	6	6385	6	5	11	21	38	38238	8	2	6	11
-25	-25978	2	3	10	17	7	7211	0	4	1	6	40	40917	6	3	3	8
-24	-24935	4	0	11	15	8	8554	2	1	2	4	41	42260	8	0	4	6
-23	-23299	0	4	7	14	10	10407	6	3	9	16	42	42770	10	5	10	20
-21	-21956	2	1	8	12	11	11750	8	0	10	14	43	43596	4	4	0	5
-19	-19277	0	2	5	9	13	13086	4	4	6	13	44	44939	6	1	1	3
-15	-15255	0	0	3	4	14	14429	6	1	7	11	46	46792	10	3	8	15
-14	-14745	2	5	9	16	15	15765	2	5	3	10	49	49471	8	4	5	12
-13	-13402	4	2	10	18	17	17108	4	2	4	8	50	50814	10	1	6	10
-10	-10723	2	3	7	13	18	18661	8	4	11	20	52	52150	6	5	2	9
-9	-9380	4	0	8	11	19	19787	2	3	1	5	53	53493	8	2	3	7
-8	-8044	0	4	4	10	20	21130	4	0	2	3	56	56172	6	3	0	4
-6	-6701	2	1	5	8	21	21640	6	5	8	17	57	57515	8	0	1	2
-4	-4022	0	2	2	5	22	22983	8	2	9	15	58	58025	10	5	7	16
-2	-2169	4	4	9	17	25	25662	6	3	6	12	62	62047	10	3	5	11
-1	-826	6	1	10	15	27	27005	8	0	7	10	64	64726	8	4	3	8
						28	28341	4	4	9	9	66	66069	10	1	3	6
						29	29684	6	1	4	7	68	68748	8	2	0	3
						30	31020	2	5	0	6	72	72980	10	5	4	12
												77	77302	10	3	2	7
												81	81324	10	1	0	2
												88	88535	10	5	1	8

(-0.049 to -0.0008)

(0 to 0.03)

(0.03 to 0.088)

^a Arranged so that the index for each entry agrees with $1000 \times m_f \pm 1.9$.

PART B(ii):

ANALYSIS OF THE
CHEMICAL CONSTITUENTS OF BODY FLUIDS

PART B-(ii) ANALYSIS OF THE CHEMICAL CONSTITUENTS OF BODY FLUIDS

OBJECTIVES:

The overall objectives of this part of the proposal are to develop the uses of gas chromatography (GC) and mass spectrometry (MS), under "intelligent" computer management, for the clinical screening, diagnosis, and study of errors of metabolism. The efficacy of these analytical tools has been demonstrated when applied to limited populations of urine samples in the research laboratory environment. We propose to enlarge the clinical investigative applications of GC/MS technology and to demonstrate its utility for the diagnosis and screening of disease states. Specifically we will apply our GC/MS analysis capabilities to larger and more diversified populations to establish better defined norms, deviations related to identifiable disease states, and control parameters required to remove ambiguities from results.

BACKGROUND AND PROGRESS:

For some time we have focussed a substantial part of our effort on exploiting the use of the mass spectrometer as an analytical instrument for biochemical purposes. Our central approach has been to integrate the mass spectrometer with the gas chromatograph on the one hand and with "intelligent" computer management on the other. Gas chromatography is a versatile and broadly applicable method for the separation of biochemical specimens into a large number of distinct but unnamed fractions. The mass spectrometer has unique power to analyze such fractions and give information relevant to their molecular structure. The computer becomes indispensable for the overall management of the system and for the reduction and interpretation of the large volume of data emanating from the analytical instruments. Our effort in instrumentation, therefore, is an integral part of this research and comprises a good deal of computational software embracing both real time instrument and data management as well as artificial intelligence. It also requires considerable effort in electronic and vacuum technology for the instrumentation hardware, and a coherent system approach for the overall integration of these components. These aspects of the effort are described in section B(i) of this proposal.

The routine screening of normal and abnormal body metabolites, as well as drugs and their metabolites, in human body fluids (ref 1) is currently the object of several research programs. Various non-specific methods, including thin layer (ref 2, 3), ion exchange (ref 4, 6), liquid (ref 5), and gas chromatography (ref 7-10), are used primarily with the goal of separating a large number of unnamed constituent materials. When used in conjunction with mass spectrometry, these methods become

specific and provide a powerful means of positive identification of metabolites in human body fluids (ref 11-13). Of these techniques, gas chromatography is the most convenient to interface to the mass spectrometer because the carrier gas can easily be removed as the analysis proceeds on a continuous flow. Based upon the references cited, as well as our own on-going programs, the ability of the GC/MS technique for the analysis of body fluids is well established. We have drawn upon the published literature in helping to design our experimental protocols.

Standard chemical procedures for extracting, derivatizing, and hydrolyzing urine and plasma are used for the GC/MS analysis (ref 13). These procedures permit separation of the following classes of substances: acids, phenols, amino acids, and carbohydrates. It is possible to detect free or conjugated compounds within these classes.

The gas chromatographic analysis of each class of compounds presents a metabolic profile. Abnormal profiles (containing either excessively large peaks from one or more components or peaks which do not correspond to metabolites usually encountered) are then assayed by mass spectrometry. The mass spectra recorded during the elution of each gas chromatographic peak then serve to identify the constituents present in that peak.

Most medical centers have access to amino acid analyzers in order to screen patients for metabolic abnormalities of the principal amino acids, but unless a special research interest exists, other errors of metabolism cannot easily be studied. At this institution the GC/MS system provides us the opportunity to detect a wide variety of errors which show accumulation of novel amino acids, fatty acids, and many other metabolites in urine, blood, and other biological fluids and tissues.

Urine is known to contain several hundred organic compounds. The separation (gas chromatography) and hence identification (mass spectrometry) of these components would be an extremely difficult task. To simplify the separation problem the urine is chemically separated into four fractions as illustrated in the following diagram.

component as beta-amino isobutyric acid from a comparison with a literature (ref. 19) spectrum of authentic material. Quantitation showed that this patient was excreting 1.2 grams per day of beta-amino isobutyric acid. After medical treatment this metabolite was no longer detected in the patient's urine thereby raising the question of whether beta-amino isobutyric acid can be used as a metabolic signature for the recognition of lymphoblastic leukemia and for the status of the disease in the course of the treatment cycle. Beta-amino isobutyric acid has been observed in the urine of 5 patients suffering from leukemia and in all instances it disappeared immediately following drug therapy. We are continuing our study of this relationship in view of the recognized excretion of elevated amounts of beta-amino isobutyric acid as the result of a genetic trait. For instance Harris et al. (ref. 14) observed daily urinary excretions of 70-300 mg of beta-amino isobutyric acid and noted that histories of high excretion levels tended to exist in particular families.

As a second 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, poor weight gain, and, as distinct from respiratory acidosis, onset after the second day of life. Its incidence is higher in infants whose birthweight is less than 1750g (one study shows 92% incidence for these children) than in infants with birthweight greater than 1750g (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 1000g and the other three were born weighing less than 1500g. 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 presence of these compounds in the urine gradually diminished and almost completely disappeared 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.

In the course of these studies, we have recognized two areas where computer analysis of the data is important in order to

handle the volume of data involved and to standardize the analyses performed. At present these operations, GC profile analysis and mass spectrum identification, are largely manual. In the case of GC profile analysis, approximately 40 peaks for each profile must be analyzed in terms of their positions, sizes, etc. relative to other peaks in the profile and instrument parameters to evaluate the presence or absence of abnormalities. For each abnormal peak, a number of mass spectra (5 to 10), each containing ion abundance measurements at approximately 500 masses, must be compared against catalogued known materials for identification. If the material is not in the catalog, the mass spectrum must be interpreted from basic principles, using high resolution spectrometry and other data sources as appropriate. These are very tedious operations requiring automation for even the proposed limited screening volume. The developmental aspects of these computer-related portions of the research program are discussed in the other sections of this proposal.

FUTURE PLANS

In the next grant period we plan to extend our efforts in applying GC/MS techniques to clinical problems both in terms of defining norms and in terms of studying identifiable disease states in collaboration with clinical investigators.

The most appropriate target material for this developmental effort is the metabolic output of NORMAL subjects under controlled conditions of diet and other intakes. The eventual application of this kind of analytical methodology to the diagnosis of disease obviously depends on the establishment of normal baselines, and much experience already tells us how important the influence of nutrient and medication intake can be in influencing the composition of urine, body fluids, and breath.

Among the most attractive subjects for such a baseline investigation are newborn infants already under close scrutiny in the Premature Research Center and the Clinical Research Center of the Department of Pediatrics at this institution. Such patients are currently, for valid medical reasons, under a degree of dietary control difficult to match under any other circumstance. Many other features of their physiological condition are being carefully monitored for other purposes as well. The examination of their urine and other effluents is therefore accompanied by the most economical context of other information and requires the least disturbance of these subjects.

Two obvious factors which could profoundly influence the excretion of metabolites detected by GC/MS are maturity and diet. We have already initiated a program for serial screening of urinary metabolite excretion in premature infants of various gestational ages and determination of changes in the pattern of excretion of various metabolites as a function of age following birth. These studies are being performed on infants admitted to

the Center for Premature Infants and the Intensive Care Nursery at Stanford, a source of some 500 premature infants per year. In addition, in conjunction with an independent study on the effects of both quality and quantity of oral protein intake on the incidence and pathogenesis of late metabolic acidosis of prematurity, we plan to measure the urinary excretion patterns of various metabolites and thereby partially assess the effect of diet on this screening method.

We shall use the analyses on blood and urine specimens from normal individuals in the final development of rapid, automated identification of compounds described by mass spectrometry. The computer will be used to match an unknown mass spectrum with reference spectra contained in computer files. Programs are also being developed which will provide the strategy for the computer to interpret an unknown mass spectrum (not contained in the library) and directly identify the compound (see Parts A and C).

Limited libraries exist for urine and plasma GC/MS analyses and will require progressive compilation (assisted by the GENERAL interpretation programs) as our clinical sampling proceeds. This will in turn speed the throughput of the system by allowing the simple identification of materials by computer library search procedures. This library will be shared freely with other investigators.

Given our ability to identify various constituents of urine and plasma and to understand normal variation, we shall apply the GC/MS system to pathology, making use of patients with already identified metabolic defects for control purposes. The main application will, of course, be diagnostic and patients with suggestive clinical manifestations, such as psychomotor retardation and progressive neurologic disease, as well as suggestive pedigrees (e.g. affected offspring of consanguineous parents or multiplex sibships) will be investigated. These patients are seen relatively frequently at any university hospital, and their presence in the various in-patient and out-patient services of the Stanford Department of Pediatrics is well documented. The GC/MS system will be helpful in diagnosing not only errors of amino acid metabolism, but also many other metabolic disorders, some of which are lactic acidemia (ref 15), Refsum's disease (a defect in the oxygenation of phytanic acid (ref 16)), methylmalonic acidemia (ref 17) and orotic aciduria (ref 18). We also recognize the potential of this methodology to define new errors of metabolism.

We will collaborate with Professor Howard Cann of the Department of Pediatrics and derive much of the clinically significant material for analysis from patients in the Premature Research Center and the Clinical Research Center of the Department of Pediatrics and the Stanford University Children's Hospital. Analyses will be performed on existing GC and MS equipment in the Departments of Genetics and Chemistry.

REFERENCES

- 1) Schwartz, M.K., "Biochemical Analysis," Anal. Chem., 44, p. 9R, (1972).
- 2) Heathcote, J.G., Davies, D.M., and Haworth, C., "The Effect of Desalting on the Determination of Amino Acids in Urine by Thin Layer Chromatography." Clin. Chim. Acta, 32, p. 457 (1971).
- 3) Davidow, B., Petri, N.L., and Quame, B., "A Thin Layer Chromatographic Screening Procedure for Detecting Drug Abuse," Amer. J. Clin. Pathol., 50, p 714, (1968).
- 4) Efron, K. and Wolf, P.L., "Accelerated Single-column Procedure for Automated Measurement of Amino Acids in Physiological Fluids," Clin. Chem., 18, p 621, (1972).
- 5) Purtsis, C.A., "The Separation of the Ultraviolet-absorbing Constituents of Urine by High Pressure Liquid Chromatography," J. Chromatog., 52, p 97, (1970).
- 6) Wilson-Pitt, W., Scott, C.D., Johnson, W.F., and Jones, G., "A Bench-top, Automated, High-resolution Analyzer for Ultraviolet Absorbing Constituents of Body Fluids," Clin. Chem., 16, p. 657 (1970).
- 7) Dalgliesh, C.E., Horning, E.C., Horning, M.G., Knose, K.L., and Yarger, K., "A Gas-Liquid Chromatographic Procedure for separating a Wide Range of Metabolites Occurring in Urine or Tissue Extracts," Biochem. J., 101, p. 792 (1966).
- 8) Teranishi, R., McN, T.R., Robinson, A.B., Cary, P., and Pauling, L., "Gas Chromatography of Volatiles from Breath and Urine," Anal. Chem., 44, p. 18, (1972).
- 9) Pauling, L., Robinson, A.B., Teranishi, R., and Cary, P., "Quantitative Analysis of Urine Vapor and Breath by Gas-liquid Partition Chromatography," Proc. Nat. Acad. Sci. USA, 68, p. 2374, (1971).
- 10) Zlatkis, A. and Liebich, H.M., "Profile of Volatile Metabolites in Human Urine," Clin. Chem., 17, 592 (1971).
- 11) Brochek, J.E., Putts, W.C., Rainey, W.T., and Burtis, C.A., "Separation and Identification of Urinary Constituents by Use of Multiple-analytical Techniques," Clin. Chem., 17, p.72 (1971).
- 12) Horning, E.C. and Horning, M.G., "Human Metabolic Profiles Obtained by GC and GC/MS," J. Chromatog. Sci., 9, p. 129, (1971)
- 13) Jellum, E., Stokke, O., and Eldjarn, L., "Combined Use of Gas Chromatography, Mass Spectrometry, and Computer in Diagnosis

and Studies of Metabolic Disorders," Clin. Chem., 18, p. 800 (1972).

14) Harris, H., "Family Studies on the Urinary Excretion of Beta-Amino Isobutyric Acid," Ann. Eugenics, Vol. 18, Page 43, (1953).

15) Haworth, J.C., Ford, J.D., and Youncszai, M.K., "Familial Chronic Acidosis due to an Error in Lactate and Pyruvate Metabolism," Canad. Med. Ass. J., 79, p. 773 (1967).

16) Herndon, J.H., Steinberg, D., and Uihendorf, B.W., "Refsum's Disease: Defective Oxidation of Phytanic Acid in Tissue Cultures Derived from Homozygotes and Heterozygotes," New England J. of Med., 281, p. 1023, (1969).

17) Morrow, G., Schwartz, R. H., Hallock, J.A., and Barnes, L.A., "Prenatal Detection of Methylmalonic Acidemia," J. Pediatrics, 77, p. 120, (1970).

18) Fallon, J.H., Smith, L.H., Graham, J.B., and Burnett, C.H., "A Genetic Study of Hereditary Orotic Aciduria," New England J. of Med., 270, p. 878, (1964).

19) Lawless, J.G. and Chadha, M.S., "Mass Spectral Analysis of C(3) and C(4) Aliphatic Amino Acid Derivatives," Anal. Biochem., 44, p. 473, (1971).

20) Reynolds, W.E., Bacon, V.A., Bridges, J.C., Coburn, T.C., Halpern, B., Lederberg, J., Levinthal, E.C., Steed, E., and Tucker, R.B., "A Computer Operated Mass Spectrometer System," Anal. Chem., 42, p. 1122, (1970).

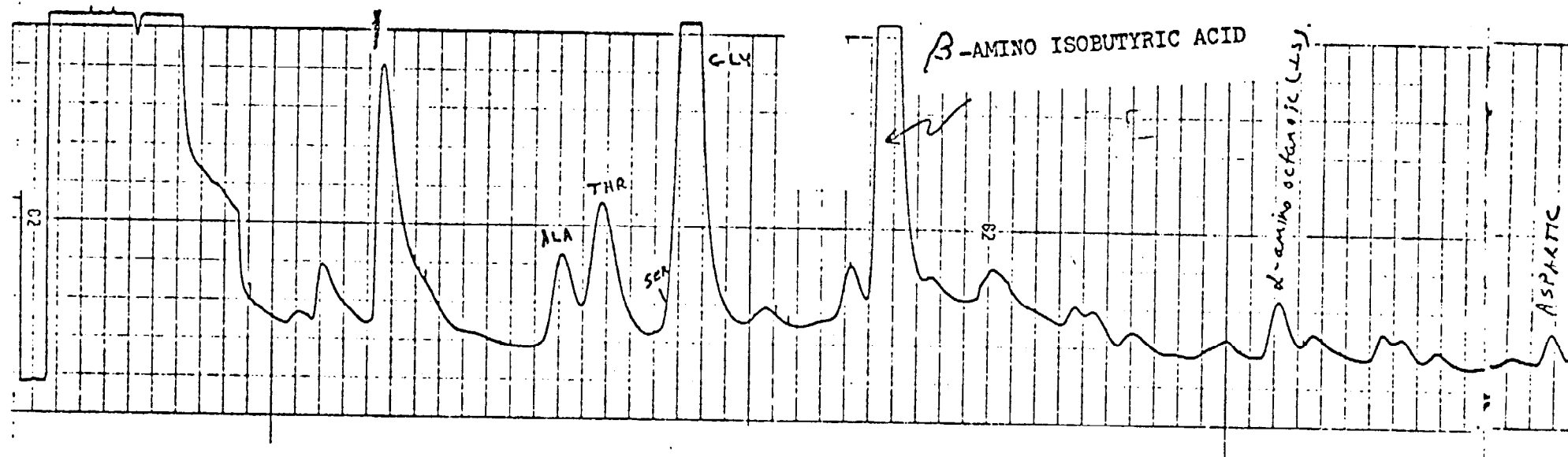


FIGURE 1

Gas Chromatogram of the Amino Acid Fraction of Urine

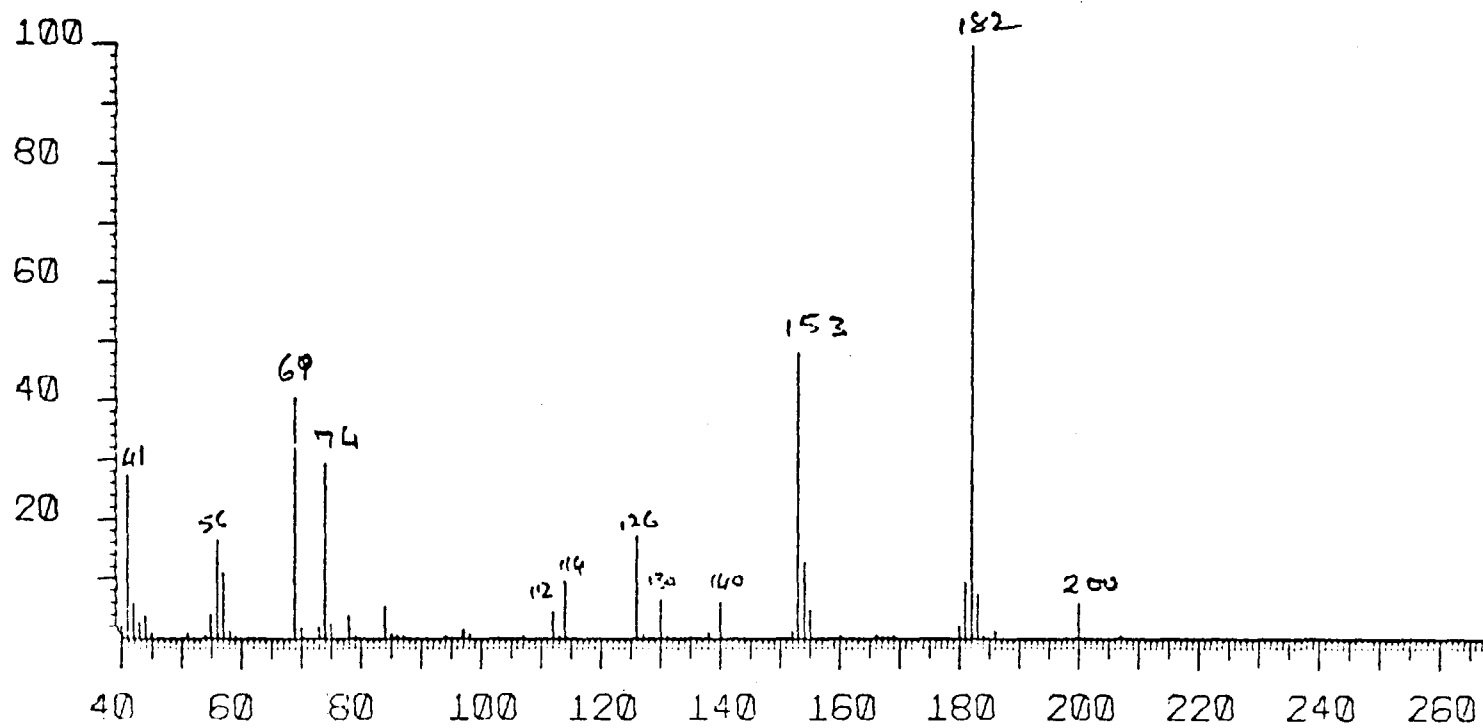


FIGURE 2

Mass Spectrum of Beta-Amino Isobutyric Acid

PART C:
EXTENSION OF THE
THEORY OF MASS SPECTROMETRY BY COMPUTER

PART C. Extending the Theory of Mass Spectrometry by a Computer (Meta-DENDRAL)

OBJECTIVES:

The Heuristic DENDRAL performance program described in Part A is an automated hypothesis formation program which models "routine", day-to-day work in science. In particular, it models the inferential procedures of scientists identifying components, such as those found in human body fluids. The power of this program clearly lies in its knowledge about various classes of compounds normally found in body fluids, which knowledge allows identification of the compounds.

The Meta-DENDRAL program described in this part is a critical adjunct to the performance program because it is designed to supply the knowledge which the performance program uses. Theory formation is essential in order to carry out the routine analyses - either by hand or by computer. However, the staggering amount of effort required to build a working theory (even for a single class of compounds) holds back the routine analyses. The goal of the Meta-DENDRAL program is to form working theories automatically (from collections of experimental data) and thus reduce the human effort required at this stage. By speeding up the time between collecting data for a class of compounds and understanding the rules underlying the data, the Meta-DENDRAL program will thus provide an improvement in the development of diagnostic procedures.

Theory formation in science is both an intriguing problem for artificial intelligence research and a problem area in which scientists can benefit greatly from any help the computer can give. While the ill-structured nature of the theory formation problem makes it more a research task than an application, we have already provided computer programs which are of definite help to the theory-forming scientist.

Mass spectrometry is the task domain for the theory formation program as it is for the Heuristic DENDRAL program. It is a natural choice for us because we have developed a large number of computer programs for manipulating molecular structures and mass spectra in the course of Heuristic DENDRAL research and because of the interest in mass spectrometry among collaborative researchers already associated with the project. This is also a good task area because it is difficult, but not impossible, for human scientists to develop fragmentation rules to explain the mass spectrometric behavior of a class of molecules. Mass spectrometry has not been completely formalized, and there still remain gaps in the theory.

Understanding theory formation enough to automate substantial parts of it will benefit all of the biomedical sciences. More directly, building a computer program which forms a theory of mass spectrometry will greatly enhance the power of mass spectrometry as a diagnostic instrument.

Detailed accounts of this research are available in the DENDRAL Project annual report to the National Institutes of Health, in several research papers already published and in manuscripts submitted for publication.

PROGRESS:

In the period covered by the initial NIH grant the Meta-DENDRAL program has moved from a set of ideas to a set of working computer programs.

The first three segments of Meta-DENDRAL have been programmed and can be used with new experimental data. These segments are first summarized and then described in more detail in subsequent sections. We described the initial design of the Meta-DENDRAL program in a paper presented to the 2nd International Joint Conference on Artificial Intelligence (London, August, 1971). And further design details and partial implementation of programs were described in a paper presented at the 7th Machine Intelligence Workshop (Machine Intelligence 7, B. Meltzer & D. Michie, eds., 1972).

Summary of Segment 1

The data interpretation and summary program (INTSUM) defines the space of mass spectrometric processes, interprets all the data in terms of these processes, and summarizes them process by process. This program is capable of a much more thorough analysis of the data than a human can perform.

Summary of Segment 2

The rule formation program starts with the interpreted and summarized results of the data. It searches the set of processes for those that meet the criteria for rules, and attempts to resolve ambiguities when several processes explain many of the same data points. The resulting rules are characteristic processes for the whole class of molecules.

Summary of Segment 3

The class separation program is an extension of the simple rule formation program just mentioned. Because the initial set of molecules may not all behave alike in the mass spectrometer, it is necessary to separate the important subclasses and formulate characteristic rules for each subclass.

SEGMENT 1. The initial segment of the theory formation program is data interpretation. After the experimental data have been collected for a large number of compounds, the program re-interprets all the data points in terms of its internal model of the experimental instrument. This part of the program has already proved useful to chemists studying the mass spectrometry of new classes of compounds. It has been described in a paper recently submitted for publication (Applications of Artificial Intelligence for Chemical

Inference X. INTSUM. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids, submitted to Tetrahedron).

The computer program for data interpretation and summary has been well developed. While it is never safe to call a program "finished", this program has reached the stage where we have turned it over to the chemists who want to look at explanatory mechanisms for the mass spectra of many compounds. Ordinarily, this is such a tedious task that chemists are forced to limit their analysis to a very few out of a total space of potentially interesting mechanisms. The computer program, on the other hand, systematically explores the space of possible mechanisms and collects evidence for each.

This program is described in the Machine Intelligence 7 paper, and the results obtained by running it with many estrogen spectra are discussed in the manuscript submitted to Tetrahedron. Mr. William C. White has been largely responsible for coding the program in LISP. The program runs in the overnight LISP system at the Medical School's ACME facility, and on the Stanford Computation Center IBM 360/67. It is currently being used by Dr. Steen Hammerum, a post-doctoral fellow in chemistry from the University of Copenhagen, to summarize the fragmentations found in the spectra of substituted progesterones, and by Dr. Dennis Smith to interpret data from other classes of steroids.

SEGMENT 2. The second segment of Meta-DENDRAL produces reasonable rules of mass spectrometry. The rule formation segment starts with the interpreted and summarized data from the first segment. It looks for the processes which are most frequent, which explain highly significant data points, and which are least ambiguous with other processes. After applying these criteria, it selects a set of processes which appear to be characteristic of the whole set of molecules initially given.

Planning before rule formation is necessary because there is so much information in the summary of possible fragmentations found in the data. It is desirable to collect all the information to avoid missing unanticipated mechanisms which occur frequently throughout the compounds in the data. But even the summary of the mechanisms is voluminous enough to obscure the "obvious" rules waiting to be found.

In a planning program implemented by Mr. Steven Reiss, the computer peruses the summary looking for mechanisms with "strong enough" evidence to call them first-order rules of mass spectrometry. Our criteria for strong evidence may well change as we gain more experience. For the moment, the program looks for mechanisms which (a) appear in almost all the compounds (80%) and (b) have no viable alternatives (where "viable alternatives" are those alternative explanations which are frequently occurring and cannot be distinguished unambiguously).

The output of this program, even though crude in many

seases, is useful to chemists who first want to see the highly reliable, unambiguous rules which can be formulated. If there are none, of course, there is little point in pressing ahead blindly. This is an indication that some modifications need to be made, for example, splitting up the original set of compounds into more homogeneous subgroups. On the other hand, if some likely rules can be found, these will serve as "anchor points" for resolving ambiguities with other sets of mechanisms and also serve as a "core" of rules to be extended and modified in the course of detailed rule formation.

SEGMENT 3. As mentioned above, class separation is important because the initial collection of compounds may not be known to behave alike in the instrument. The rule formation program must be prepared to retract its assumption of homogeneity. Mr. Steven Reiss, working with Dr. Buchanan, has written a first extension of the rule formation program which allows class separation on the basis of characteristic rules found for the subclasses.

A paper describing segments 2 and 3 - rule formation with subclass separation - has been submitted to the 3rd International Joint Conference on Artificial Intelligence.

The computer programs produced to date have already proved useful for helping to formulate mass spectrometry theory for classes of biologically relevant molecules. Chemists have used these programs as tools for rule formation. They have examined the estrogenic steroids this way, including separate studies on some equilenins, acetates and benzoates. Also, they have used the program to interpret data from several classes of pregnanes.

Plans:

In the coming period we propose to focus on three aspects of theory formation. We plan to (1) extend the capabilities of the programs, (2) make our rule formation programs more usable by chemists, and (3) continue our exploration of the more theoretical aspects of rule formation.

1. We anticipate new difficulties as the classes of molecules under study become more complex, either with respect to structural features or mass spectrometric behavior. Although we have made the programs flexible, extending the work just to new sets of data will undoubtedly introduce new problems.

Now that the usefulness of the programs has been demonstrated, we propose to couple the theory formation program more closely to data of more direct clinical relevance. For example, the mass spectrometry of amino acids and the aromatic acids frequently found in urine needs to be better understood before automatic analysis of the components of (the acid and neutral fractions of) urine is successful. Parts A and B of this proposal, in other words, can both be helped by the continuation of Part C.

The program is now limited to forming rules which are more

descriptive of the sample than explanatory. We are currently working on ways of generalizing the descriptive rules so that they are more truly general. Drs. Sridharan and Buchanan have started experimenting with computer programs which generalize the rules in various ways. Mr. Carl Farrell is currently working on a computer program for his Ph.D. thesis which allows systematic exploration of various methods of generalizing on rules. His work investigates the efficacy of different control structures as well as different inductive rules.

2. The programs are now used by chemists, but not without a fair amount of help from the programming staff. We must overcome some of the barriers to facile use before the programs can be counted as successful. For example, putting the data in the correct format can be made easier, as can defining constraints on the search space and modifying parameter values.

The programs do not now require the chemist to know LISP. However, we propose to develop easier access to control of the programs through careful design of the user interface. Depending on hardware limitations, we would also like to provide a time-shared, graphics-oriented interface.

3. The descriptive form of rules mentioned above may be inherent in the conceptual framework we have chosen for the rule formation program. The program uses a "ball and stick" model of molecular structures, so it is no surprise that situations and actions in rules are simply described. We wish to explore more sophisticated models of mass spectrometry with the hope of discovering how a program could search the space of possible models during rule formation. This is still a very challenging problem. We have so far concentrated on more practical aspects of theory formation - i.e., producing results of immediate utility. But we feel strongly that we must grapple with the outer reaches of the problem in order to arrive at meaningful solutions.

PUBLICATIONS -- PART C

B.G. Buchanan, E.A. Feigenbaum, J. Lederberg, "A Heuristic Programming Study of Theory Formation in Science", in Proceedings of Second International Joint Conference on Artificial Intelligence, Imperial College, London (September, 1971). (Also Stanford Artificial Intelligence Project Memo No. 145, Computer Science Dept. Report CS-221)

B.G. Buchanan, E.A. Feigenbaum, and N.S. Sridharan, "Heuristic Theory Formation: Data Interpretation and Rule Formation". In Machine Intelligence 7, Edinburgh University Press (1972).

B.G. Buchanan and N. Sridharan, "Rule Formation on Non-Homogeneous Classes of Objects", submitted for presentation at the Third International Joint Conference on Artificial Intelligence (Stanford, August, 1973).

PART D:

APPLICATIONS OF CARBON(13) NUCLEAR MAGNETIC
RESONANCE SPECTROMETRY TO ASSIST IN CHEMICAL
STRUCTURE DETERMINATION

PART D. CARBON-13 NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

The goal of our Heuristic DENDRAL research is to develop rapid, accurate and flexible computer techniques for identifying unknown steroids and other biologically important compounds from spectroscopic data. We have made significant progress toward this goal: Our system is currently capable of correctly analyzing high-resolution mass spectra of estrogenic steroids and mixtures thereof. As we extend our methods to the more complex problems presented by other steroid classes, and eventually by other types of biologically important molecules, we will find it necessary to have available sources of structural information other than mass spectroscopy. Carbon-13 nuclear magnetic resonance (CMR) spectroscopy is an ideal candidate.

Basically, the CMR experiment measures the extent to which each carbon nucleus in the sample molecule is shielded from an applied magnetic field. This shielding, or chemical shift, is caused by the distribution of electrons around the nucleus, and is determined by the carbon's hybridization and local chemical environment. Other investigators have determined that the shift of a carbon is strongly dependent upon the nature and placement of substituents at nearby centers, and that to a first approximation these substituent effects are additive. Thus, the CMR spectrum of a compound contains information which rather straightforwardly can be related to the possible local environments of each carbon. The structural information provided by CMR data compliments that from mass spectroscopy, and there is relatively little redundancy between the two methods. Data from the latter represent molecular fragmentations, which take place most readily near functional groups. Thus, mass spectroscopy frequently gives structural information about the environments of such groups. In CMR spectroscopy, on the other hand, the chemical shifts of carbons in large alkyl moieties, far removed from functionality, are the best understood and the most predictable. Further, the

fragmentation of large molecules such as steroids can show the general pattern of substitution in the molecule, while CMR shifts are sensitive to specific local patterns. Because the two methods "mesh" so nicely, we see the development of analytic CMR techniques as an extremely fruitful field of research. Our eventual aim is to completely define the structures of unknown compounds using only these two sources of information.

We are well equipped to study this field. In our Chemistry department, we have a Varian XL-100 (Fourier-transform) nuclear magnetic resonance spectrometer, one of the most sensitive and flexible instruments currently available for CMR work. We have competent investigators in our Chemistry and Computer Science departments who are interested in, and in fact currently working on, the project. Finally, we have had considerable experience with computerized structure analysis, and much of what we have learned can be applied to the CMR problem.

We have already begun investigating the use of CMR data in automated structure analysis, with our initial study focussed upon the acyclic amines. The analysis of low-resolution mass spectra of large amines is not capable of discerning the structures of long alkyl chains, so we felt that this class of molecules would provide a good test of CMR methods. Ms. Hanne Eggert of our group has obtained the CMR spectra of over 100 acyclic amines, and has derived an accurate set of predictive rules relating structure to chemical shifts. Dr. Raymond E. Carhart has used these rules to develop a computerized approach to the identification of amine structures from observed CMR spectra (see attached manuscript). The program, entitled AMINE, has proven to be extremely selective: The analysis of the CMR spectrum of trioctyl amine, for example, yields only seven possible structures, though the molecule has over 700 million structural isomers. In contrast, the analysis of the low-resolution mass spectrum of triheptyl amine gives nearly 2000 solutions out of a possible 38 million isomers. These results illustrate the tremendous amount of structural information which CMR spectroscopy can provide.

This source of information has, in general, been ignored in steroid-identification research, primarily because large amounts of sample (50 milligrams or more for steroids) are needed to obtain reliable CMR spectra. However, CMR spectroscopy is still a relatively new field, and the sensitivity of current instruments is far from the threshold which new technologies can provide. We expect the minimum sample size to drop to the sub-milligram level in the future, and with such sensitivity, the CMR spectrometer could be a powerful tool in biochemical and medical research. If this tool is to be utilized to its fullest extent, it is important that we begin now to develop the concepts and techniques needed in the interpretation of CMR data.

We propose, then, to study various classes of steroids in a manner analogous to the amine study, with the goal of developing a program which can 'reason out' steroid

structures from CMR data, perhaps in combination with mass-spectral data. Ms. Eggert has already collected CMR data on a variety of keto-substituted androstanes and cholestanes to assess the effect of the carbonyl group on the chemical shifts of the steroid-skeleton carbons, and has, in the process, uncovered some mistaken CMR shift assignments published in the literature. We will study a variety of functional groups in this way, deriving general rules for predicting the spectra of more complex steroids. As these rules emerge, we will couple them with the computerized heuristic-search and structure-generation techniques which we have developed in our previous mass- and CMR-spectroscopy research.

PUBLICATIONS -- PART D

R.E. Carhart and C. Djerassi, J. CHEM. SOC. (PERKIN II), submitted for publication (see attached preprint).

H. Eggert and C. Djerassi, J. Amer. Chem. Soc., in press.

Proofs (if required) by air mail to Professor Carl Djerassi
Department of Chemistry
Stanford University
Stanford, California 94305

Applications of Artificial Intelligence for Chemical Inference. XI.¹
Analysis of Carbon-13 NMR Data for Structure Elucidation of Acyclic
Amines

Raymond E. Carhart² and Carl Djerassi,* Departments of
Computer Science and Chemistry, Stanford University,
Stanford, California, 94305, U. S. A.

This paper describes a computer program, entitled AMINE, which uses a set of predictive rules to deduce the structures of acyclic amines from their empirical formulae and Carbon-13 NMR (CMR) spectra. The results, summarized in Tables 2-5, of testing the program on 102 amines indicate that AMINE is quite accurate and selective, even for large amines with many millions of structural isomers, and demonstrate that the computerized analysis of CMR data can be a powerful analytical tool. The logical structure of the program is outlined here, including a section on the general problem of spectrum matching. Generalizations of the methods used by AMINE are suggested.

I. INTRODUCTION

In recent years, there has been a substantial amount of research directed toward the computerized identification of molecular structure from mass-spectroscopic³⁻⁵, NMR,^{4,6,7} and infra-red⁷ data. Our Heuristic DENDRAL program,^{3,4} which relies primarily upon mass-spectral

data, has been shown to be quite accurate for certain classes of saturated, acyclic, monofunctional compounds, and more recently, the methods have been extended to the estrogenic steroids.^{3b} There are limitations to the information content of mass-spectral data, however, particularly when compounds are considered which have long, perhaps highly branched alkyl chains. An analysis of the mass spectrum of triheptylamine, for example, yields about 2000 solution structures,⁴ and although this is only a small fraction of the roughly 40 million (non-stereochemical) isomers of $C_{21}H_{45}N$, it is still an impractically large number. The problem is that alkyl moieties do not give characteristic fragmentation patterns, and in fact, most spectroscopic methods are relatively insensitive to their structure.

However, recent studies indicate that C-13 nuclear magnetic resonance (CMR) spectroscopy⁸ is an exception. For several classes of compounds,⁹ rules have been obtained which allow one to predict the CMR spectrum of a substance from its molecular structure, and in all cases, the rules indicate that the chemical shift of any Carbon, even one in a large alkyl chain-end, depends heavily upon branching at nearby centers. Thus, it appears that CMR spectroscopy, either alone or in combination with other methods, could be a powerful tool in the computerized analysis of molecular structure. This paper outlines the methods by which such an analysis may be carried out for the acyclic amines, and describes a FORTRAN IV computer program,¹⁰ entitled AMINE, in which these methods are implemented.

This class of compounds was chosen for two reasons. First, the recent work of Eggert and Djerassi^{9a} has yielded a detailed set of predictive rules for the acyclic amines. Secondly, for a given number of Carbon atoms, a saturated, acyclic amine has decidedly more structural possibilities than most other simple types of acyclic organic compounds (for example, stereochemistry aside, there are nearly 15 million C20-amines, but only about 6 million C20-alcohols),¹¹ and thus the structural analysis of amines represents a particularly challenging problem.

II. DEFINITION OF THE PROBLEM

A fully proton-decoupled, natural-abundance CMR spectrum⁸ typically consists of a number of sharp peaks representing the resonance frequencies, in the applied magnetic field, of the various types of Carbon atoms present in the sample. A standard compound, commonly TMS, is usually included in the sample to provide a reference frequency, and the peak positions, or chemical shifts, are measured as fractional deviations from this reference, in parts per million (ppm). Previous investigations have shown that the shift of a particular Carbon is determined by its hybridization and local environment, and thus each shift contains some structural information. There are a few ranges of shifts which are characteristic of certain functional groups, such as C=O or C=C, but aliphatic Carbons in most molecules lie in a broad

spectral region from which detailed structural information cannot be extracted readily.

For acyclic amines,^{9a} and a few other types of compounds,^{9b-k} there exist predictive rules which allow one to calculate the spectrum of a compound whose structure is known, with a typical accuracy of about 1-2 ppm in a total range of roughly 100 ppm. For these classes, the structure-identification problem could in principle be solved via the generation of all possible structures of a particular type (say, acyclic amines with a particular number of Carbon atoms), the prediction of their spectra, and the comparison of these predictions with the observed spectrum. In fact, Sasaki et al.^{6b} have used this procedure in the automated identification of a few small alkanes. For large molecules, though, the number of possible isomers can be overwhelming, and even a very efficient computer program could not carry out such an analysis in a reasonable length of time.

Program AMINE is designed to accomplish the same goal, but in a much more efficient manner. It takes, as its only input data, an observed CMR spectrum, the number of Carbons in the amine, and a goodness-of-fit criterion. The observed spectrum consists of a list of shifts, $\underline{o}=(o_1, \dots, o_n)$, measured in ppm relative to TMS. Each of these corresponds to one or more Carbons in the sample molecule. Under favorable circumstances,¹² it is possible to determine the number of Carbons corresponding to each observed shift (this will be called the tally of the shift) once the relative peak intensities and the

empirical formula are known. If the tally of a shift is known to be at least 2, 3, etc., then the shift is entered in duplicate, triplicate, etc. in the observed-shift list. These tallies are not necessary to the program's operation, but even if they are underestimated, they can add considerably to the speed and accuracy of the analysis. The number of Carbons, N_C , in the amine must be greater than, or equal to, the number of shifts in the observed spectrum. Generally, N_C cannot be determined from the CMR spectrum, but must be obtained from some other analytical method such as mass spectroscopy or elemental analysis. The goodness-of-fit criterion, DELTA, which is used in the comparison of ρ to the predicted spectra of molecules or molecular fragments, represents the maximum expected error in the predictive rules. The amine rules are derived, in part, from the alkane rules of Lindeman and Adams,^{9d} who note that 95 percent of the studied alkanes have predicted shifts within 1.5 ppm of the observed values. A similar situation exists for the amines, so a value of DELTA = 1.5 ppm has been used in most of this work.

The goal of the program is to find all acyclic, N_C -Carbon amines whose predicted spectra satisfy the following two criteria: a) Every predicted peak must lie within DELTA of one of the observed peaks; and b) Within this limit, the predicted shifts must be assignable to the observed ones in such a way that all of the latter are accounted for.

III. OVERVIEW OF PROGRAM OPERATION

The operation of program AMINE can best be viewed in terms of four interconnected processes; structure generation, pruning, filtering, and spectrum matching. The STRUCTURE GENERATOR builds a pool of increasingly large and complex alkyl chain-ends, and eventually uses these to construct amine molecules. It relies heavily upon the PRUNER to cull from the growing pool any chains which are inconsistent with the observed spectrum, and similarly upon the FILTER to test entire amine molecules. The FILTER also takes care of outputting the acceptable solution structures, and ranking them according to how well they fit the observations. Both the FILTER and the PRUNER use the spectrum MATCHER, which is responsible for the actual comparison of predicted and observed spectra. Each of these processes will be discussed in detail, below.

IV. STRUCTURE GENERATION

The structure generation scheme used in this study, which is related to the enumeration algorithm of Henze and Blair,¹³ is applicable only to saturated, acyclic, monofunctional compounds. It is an efficient approach from the standpoint of CMR structural analysis because it rapidly generates substructures which contain a relatively large number of "predictable" Carbons (*i. e.*, those near the ends of alkyl chains), and thus many of these substructures may be ruled out early in the analysis as being inconsistent with the observed data.

At any point in the generation, the STRUCTURE GENERATOR contains a pool of monovalent alkyl radicals which, through pruning (see below) have been found to be consistent with the observed CMR spectrum. The pool initially contains only the $-CH_3$ radical. By attaching one or more of these pool members (along with an appropriate number of hydrogen atoms) to a central Carbon, it constructs new radicals, each of which is passed to the PRUNER for testing. Any that agree with the observed spectrum are included in the pool, and are subsequently used to construct larger chains. In the final step of the analysis, the STRUCTURE GENERATOR similarly attaches alkyl groups to a central Nitrogen, constructing amine molecules of the proper empirical formula. These it passes to the FILTER for final testing and ranking. At all stages of the generation, tests are made which insure that no radical or amine is considered twice.

As will be discussed below, a given alkyl radical actually undergoes several different tests during pruning, with each test corresponding to a distinct chemical environment in which the chain-end might exist. The STRUCTURE GENERATOR keeps a record of these tests for each pool member, and constantly checks that it is using the radicals in a consistent fashion. If, for example, the PRUNER finds that the ethyl group is consistent with the observed spectrum only if it is attached to Nitrogen in a secondary amine, the STRUCTURE GENERATOR will never construct an n-propyl group, sec-butyl group, or any other radical which contains an ethyl group connected to Carbon. Neither will it generate

primary or tertiary amines with N-ethyl groups.

V. PRUNING

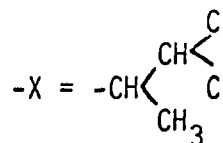
The PRUNER is the real heart of program AMINE. It is responsible for keeping the growing chain-end pool to a manageable size by weeding out alkyl radicals which are inconsistent with the observed spectrum. In testing a particular chain-end, R, shown schematically in Figure 1, the basic question considered by the PRUNER is: "Of all possible sets of CMR shifts which R could produce, is at least one consistent with the observed spectrum?" Actually, the question is somewhat more complex, but this provides a good starting point.

Now, according to the predictive rules,^{9a} a Carbon's shift is determined by the structure which surrounds it, up to four bonds away. Further, the effect of a first-row atom which is four bonds removed does not depend upon whether that atom is Carbon or Nitrogen. Thus, because X in Figure 1 must contain at least one such atom (namely Nitrogen), the shifts of C_δ and any Carbons "below" it are completely predictable and independent of the internal structure of X. The shifts of the remaining Carbons, C_β , C_γ and C_α , depend to varying degrees upon the structure of X, with C_α being the most sensitive.

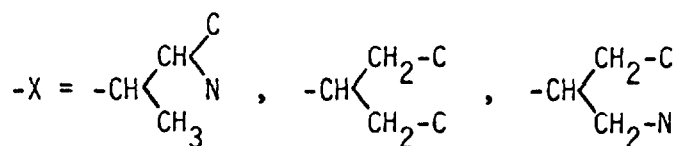
By investigating all possible X structures to a "depth" of four atoms (measured from the R-X bond), the PRUNER could generate an exhaustive list of spectra that R might produce, testing each for

inconsistency with the observed one. Usually, though, X contains enough atoms that there are several hundred of these "depth-4" structures, and the above approach proves to be rather cumbersome. Instead, the PRUNER considers only "depth-3" expansions of X, for which C_β and all Carbons below it are predictable. The shift of C_α is simply ignored, even when a reasonable estimate of its value might be made. This simplification cuts the number of unique X substructures to, at most, 94.

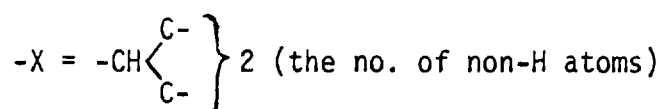
There are two factors which can reduce this number still further. First, some of the substructures may contain too many (or few) Carbons to be consistent with the known atom-count of X. Secondly, there are many cases in which a single predicted spectrum for R may result from two or more related X's. This situation arises because, according to the predictive rules,^{9a} the shifts of certain types of Carbons (specifically, those which are four or more bonds from Nitrogen, or three if the degree of the amine is known) are not sensitive to the type or distribution of first-row atoms which are four bonds away, but only to their number. Thus, in the computation of the shift of C_β ,



is equivalent to three other structures:



All four may be considered as a single entity, which can be represented as:



Once such a grouping of X substructures has been done, there remain, at most, 69 cases for the PRUNER to consider. These are summarized in Table 1, where they are further grouped into fifteen classes according to a) the type of atom directly attached to R, b) the degree of that atom and c) if that atom is Carbon, and is attached to Nitrogen, the degree of the amine. The actual purpose of the PRUNER is to consider each of these classes, determining whether at least one class member gives R a predicted spectrum consistent with the observed one, and to return the results of the fifteen class-tests to the STRUCTURE GENERATOR.

The efficiency of this class-by-class investigation can be greatly improved by the inclusion of a hierarchy of pre-tests, each of which is aimed at excluding one or more classes at once. For example, classes

1-12 in Table 1 all have one common feature: The atom to which R is attached is Carbon. Thus, as a pre-test for all twelve classes, the PRUNER treats X as a Carbon whose neighbors are unknown (schematically, $X = C-?$) and predicts as much of the spectrum of R as possible. If these predictions do not match the observations, it bypasses all further consideration of classes 1-12 and proceeds with the $X = N-?$ pre-test for classes 13-15. Otherwise, it considers a number of more detailed pre-tests, each corresponding to a possible set of neighbors to the central Carbon in $X = C-?$. The actual hierarchy is outlined in Figure 2. In each of the pre-tests, the local environment of either C_{β} or C_{γ} is known to a depth of only three atoms, and hence the corresponding shift cannot be predicted precisely. In most of these cases, the PRUNER can derive upper and lower limits for the shift from the predictive rules. These limits, which define an estimated shift, encompass a relatively small spectral region (0-5 ppm) because the shift of a Carbon is usually not very sensitive to atoms which are four bonds away. Even though the estimated shifts are not exact, they convey useful information to the MATCHER, and thus increase the overall program efficiency.

VI. FILTERING

In the final stages of the analysis, the STRUCTURE GENERATOR constructs amine molecules with N_c Carbons by attaching to a central

Nitrogen, one or more alkyl chains which have survived the pruning process. These amines are passed to the FILTER, which is responsible for calculating their total CMR spectra and, via the MATCHER, comparing these predictions with the observations. If an amine passes the test, the FILTER writes out the structure along with the predicted shifts. It then repeats the spectral comparison using progressively smaller values of DELTA until it finds the smallest value, DELMIN, for which a match still exists. In the event that several solution amines result from a particular run of AMINE, these DELMIN values can be helpful in ranking the candidates according to how well they fit the observed spectrum.

VII. SPECTRUM MATCHING

Eventually, the pruning and filtering processes reduce to problems in spectrum matching. Suppose the MATCHER receives for testing a list of m predicted shifts, some of which may be represented by small spectral regions rather than exact values. Now, the predictive rules are not precise, so each shift is actually associated with a range of acceptable values (given the generic symbol r) whose size is controlled by the input parameter DELTA. This parameter measures the maximum tolerable disagreement between predicted and observed shifts, so the range for a shift, S , extends from $S+\text{DELTA}$ to $S-\text{DELTA}$, while that for an estimated shift, bounded above and below by S_u and S_l , extends from $S_u+\text{DELTA}$ to $S_l-\text{DELTA}$. It should be noted that, in the latter case,

there are really two factors which contribute to the breadth of the range. One is the basic imprecision in the predictive rules, while the other arises because the PRUNER, in its pre-tests, sometimes calculates shifts for Carbons whose environment is not completely known. The spectrum-matching algorithm, though, makes no distinction between these; It only "knows" that a predicted spectrum consists of a list of ranges. This list will be written as $\underline{r}=(r_1, r_2, \dots, r_m)$, with u_i and l_i as, respectively, the upper and lower bounds for the range r_i .

The nature of the observed spectrum, $\underline{o}=(o_1, o_2, \dots, o_n)$, has been discussed in section II. The MATCHER takes these n shifts to be exact, because any estimated uncertainty (usually on the order of 0.1 ppm) in their measurement may be included in the tolerance DELTA. It is the task of the MATCHER to ascertain whether \underline{r} could be a subspectrum of \underline{o} , or if $m=N_c$ (N_c being the number of Carbons in the amine), whether \underline{r} could be interpreted as \underline{o} .

If, for a range r_i , there exists an observed shift o_j such that $u_i \geq o_j \geq l_i$, then it will be said that r_i can be assigned to o_j . The simplest test of agreement between \underline{r} and \underline{o} involves checking that each r_i in \underline{r} can be assigned to at least one o_j in \underline{o} . This test does not consider the important condition that, eventually, all shifts in \underline{o} must be used, and therefore a stronger test can be defined.

If every Carbon in the molecule gives a different observed shift, or if an analysis of peak intensity data gives the tally of each peak, then $n=N_c$. In this case, it is clear that no two predicted shifts can

be assigned to the same o_j . Thus, referring to Figure 3, r does not match o even though the simple test is not violated, because r_1 and r_3 must both be assigned to o_2 . In more complicated cases, each of several r_i 's may be assignable to two or more o_j 's, and vice versa, so the application of this test in an efficient manner can present difficulties. Fortunately, simple matching theory,^{14a} an outgrowth of the mathematical field of graph theory, provides a general method (see MATCHING ALGORITHM, below) of finding the maximum number, M , of ranges which can be assigned to the elements of o without duplication. Clearly r cannot match o if this number is less than m .

There may be cases, though, in which complete tally information is unavailable, which means that the number of observed shifts, n , is smaller than the number of Carbons, N_c . In such cases, there are $(N_c - n)$ "extra" shifts which lie somewhere beneath the n observed ones, but there is no way of determining where they belong. It is still possible to strengthen the simple test, but here, the additional constraint is that the predicted spectrum, once assigned, can have no more than $(N_c - n)$ "extra" peaks, either. If the simple test is passed, then every r_i can be assigned to at least one o_j . However, M is the maximum number of ranges which can be assigned to o without duplication, so $(m - M)$ must be the number of "extra" ranges in r . The condition that strengthens the simple test here, then, is $(m - M) \leq (N_c - n)$. Because M cannot exceed m , this condition reduces to the previous one ($m = M$) when $n = N_c$.

The above spectrum-matching scheme is useful not only in the current study, but for general cases in which a set of predicted CMR shifts of variable uncertainty is to be compared with an observed spectrum with, perhaps, incomplete intensity information.

VIII. MATCHING ALGORITHM

The algorithm for determining M is related to the so-called Hungarian method^{14b} of simple matching, but takes advantage of certain special features of the spectrum-matching problem. It may be described briefly as follows: Begin with $M=0$ and process the o_j 's in algebraic order, beginning with the largest. For each o_j , scan the ranges r_i looking for those which satisfy $u_i \geq o_j \geq l_i$, but which have not yet been assigned. If there are none, proceed to the next o_j . If there is just one, assign it to o_j , increment M by 1 and proceed to the next o_j . If there are several, assign the one with the largest lower limit (l_i) to o_j , increment M by 1 and proceed to the next o_j .

It is possible to prove that this gives the maximum matching between r and o , but a presentation of our proof is beyond the scope of this paper.

IX. RESULTS

Program AMINE has been implemented on the IBM 360/67 and DEC PDP-10

computers. Any mention of timing in the following discussion refers to total execution time (central processor time plus "wait time") on the former machine.¹⁵ The program requires about 35 thousand words of storage. A sample of the program's output is shown in Figure 4.

The only large set of amine CMR spectra available in the literature is that given by Eggert and Djerassi,^{9a} who used it in the derivation of predictive rules. The set consists of 102 amine spectra, including both shifts and tallies. Three of these spectra correspond to diastereomeric mixtures, and these are not suitable for testing AMINE, because the program assumes that the input spectrum corresponds to a pure compound. Neither is tridodecylamine because it exceeds the maximum number of Carbons (currently 24) allowed by the program. The remaining 98 amines were used in the testing of the program.

Some experimentation indicated that DELTA=1.5 ppm was small enough for efficient and selective program operation, yet large enough that about 95% of the test cases gave the correct solution among the output structures. Increasing DELTA by 50% to 2.25 ppm slowed the program by a factor of 2-4, but AMINE always obtained the correct structure with this higher DELTA value. Generally, shift tallies were found to be unnecessary for amines containing fifteen or fewer Carbons, but for larger molecules, the analyses proved to be excessively costly unless all of the Carbons were identified in the observed spectrum.

With DELTA=1.5 ppm, and using tallies for the amines with sixteen or more Carbons, the program obtained only one answer, the correct one,

for the 88 amines listed in Table 2. The six cases summarized in Table 3 gave from two to seven solutions, with the correct structure ranked (see section VI.) first or tied¹⁶ for first. For three of these six, the inclusion of tallies ruled out the incorrect answers. Four amines gave no solutions with DELTA=1.5 ppm. These were rerun using DELTA=2.25 ppm, and tallies were included to offset the longer running-times. As indicated in Table 4, three of these runs gave only the correct answer, while the fourth yielded two equally ranked solutions, including the correct one.

These analyses required from 0.02 to 100 seconds of computer time, with a typical 10- or 11-Carbon amine using about 1-2 seconds. In none of the runs was an incorrect solution obtained without the accompanying correct one, and in only four cases was it necessary to use the larger DELTA value. The results for the eight amines containing sixteen or more Carbons are especially encouraging: In a reasonable length of time, the program was able to select the correct structure, along with very few others, from an "isomer space" containing from about 300,000 (for $N_C=16$) to about 700,000,000 (for $N_C=24$) members.¹¹

The above results are biased to some extent because the amines used for testing the program are the same ones used by Eggert and Djerassi in the predictive-rule formation. As a test of the generality of the program, analyses have been run on the spectra of four "unknown" amines which do not appear in the original list. The results of these tests cases are summarized in Table 5. The spectra of the two 13-Carbon

amines were analyzed using DELTA=1.5 ppm, and no attempt was made to include tallies. Only the correct structure was obtained in these cases. For the two 20-Carbon amines, tallies were measured under special experimental conditions (see below). With DELTA=1.5 ppm, one of these gave two equally ranked structures, while the other gave none. A rerun of the second case with DELTA=2.25 ppm yielded five solutions with the correct one ranked as tied for second. This is the only case in which the ranking procedure favored an incorrect answer over the correct one, but here, as in most of the other multiple-result runs, the incorrect structures are sufficiently different from the correct one that they should be distinguishable by mass-spectroscopic techniques.

X. CONCLUDING COMMENTS

Two major conclusions result from this study. First, the CMR spectrum of an acyclic amine appears to be highly characteristic of the structure of the amine. For example, only one of the nearly 15 million¹¹ structural isomers of $C_{20}H_{43}N$ gives a predicted spectrum which matches the observed spectrum of N-butyl-di(2-ethylhexyl)amine. Thus it can be concluded that CMR data do indeed contain a tremendous amount of structural information. Secondly, it has been found that efficient methods for extracting this information exist, and can be implemented on the digital computer.

There is no reason to believe that these conclusions are peculiar

to the acyclic amines. The computational techniques outlined in this paper can readily be generalized to other classes of saturated, acyclic, monofunctional compounds: To do so for a particular class, one needs to obtain an accurate set of predictive rules, and, perhaps, to modify the pruning process slightly to account for special features of those rules. Such rules already exist for alkanes^{9c,d}, alkenes^{9e}, and alcohols^{9b}, and as research in CMR spectroscopy progresses, further sets should become available. Extensions to polyfunctional and/or cyclic classes would also require more sophisticated structure-generation methods, but these are available.^{3,6a,17}

In short, it appears that the computerized analysis of CMR spectra holds great promise as an accurate and selective tool in the identification of unknown compounds.

XI. EXPERIMENTAL

The four "unknown" amines were prepared, and their proton noise-decoupled CMR spectra obtained, using previously described techniques.^{9a} The spectra of the two 20-Carbon amines were also run in the presence of chromium acetylacetonate, and the integrated intensities from these were used to determine the peak tallies.¹⁸ The observed shifts for the four amines are given below, in ppm downfield of internal TMS. The estimated uncertainty in these shifts is 0.1 ppm. For the two 20-Carbon amines, tallies are included in parentheses.

N-(3-methylbutyl)-2-ethylhexylamine; 53.6, 48.6, 39.8, 39.6,
31.7, 29.3, 26.3, 24.8, 23.2, 22.8, 14.1, 11.0.

N-(3-methylbutyl)-1,5-dimethylhexylamine; 53.5, 45.6, 39.9,
39.4, 37.8, 28.1, 26.4, 24.0, 22.7, 20.6.

N-(2-ethylhexyl)-N-(3-methylbutyl)heptylamine; 59.6(1),
55.0(1), 53.1(1), 38.1(1), 36.8(1), 32.3(1), 31.8(1),
29.7(1), 29.4(1), 27.9(2), 26.5(1), 24.9(1), 23.6(2),
23.0(2), 14.3(2), 11.0(1)

N-pentyl-N-(3,3-dimethylbutyl)-3,5,5-trimethylhexylamine;
54.5(1), 52.5(1), 51.9(1), 50.1(1), 40.9(1), 31.1(1),
30.2(3), 30.0(1), 29.8(1), 29.7(3), 27.7(2), 22.9(2),
14.1(1).

XII. ACKNOWLEDGEMENTS

Thanks are due to Hanne Eggert for preparing the "unknown" amines and obtaining their spectra, and to Larry Masinter together with Drs. N. S. Sridharan and Bruce Buchanan for their helpful discussion and criticism of this work. The financial support for this project provided by the National Institutes of Health (grant RR-612) is gratefully

acknowledged.

References

1. For part X, see D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, J. Lederberg, and C. Djerassi, submitted for publication in Tetrahedron.
2. National Institutes of Health postdoctoral Fellow, 1972-1973.
3. (a) B. G. Buchanan, A. M. Duffield, and A. V. Robertson in "Mass Spectroscopy: Techniques and Applications," ed. G. W. A. Milne, Wiley and Sons, New York, 1971, p. 121;
(b) D. H. Smith, B. G. Buchanan, R. S. Engelmores, A. M. Duffield, A. Yeo, E. A. Feigenbaum, J. Lederberg, and C. Djerassi, J. Amer. Chem. Soc., 1972, 94, 5962, and previous papers in the series.
4. A. Buchs, A. M. Duffield, G. Schroll, C. Djerassi, A. B. Delfino, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, J. Amer. Chem. Soc., 1970, 92, 6831.
5. (a) H. S. Herze, R. A. Hites, and K. B. Biemann, Anal. Chem., 1971, 43, 681;
(b) L. R. Crawford and D. J. Morrison, ibid., 1971, 43, 1790;
(c) D. H. Smith, ibid., 1972, 44, 536;
(d) P. C. Jurs, ibid., 1971, 43, 1812, and references cited therein.
6. (a) S.-I. Sasaki, Y. Kudo, S. Ochiai, and H. Abe, Mikrochimica Acta [Wien], 1971, 726;
(b) S.-I. Sasaki, S. Ochiai, Y. Hirota, and Y. Kudo, Japan Analyst, 1972, 21, 916.
7. H. Abe and S.-I. Sasaki, The Science Reports of the Tohoku University, Series I, 1972, 55, 63.
8. For a general discussion of CMR spectroscopy, see "Carbon-13 Nuclear Magnetic Resonance for Organic Chemists," G. C. Levy and G. L. Nelson, Wiley-Interscience, New York, 1972.
9. (a) H. Eggert and C. Djerassi, J. Amer. Chem. Soc., in press;
(b) J. D. Roberts, F. J. Weigert, J. I. Kroschwitz, and H. J. Reich, ibid., 1970, 92, 1338;
(c) D. M. Grant and E. G. Paul, ibid., 1964, 86, 2984;
(d) L. P. Lindeman and J. Q. Adams, Anal. Chem., 1971, 43, 1245;
(e) D. E. Dorman, M. Jautelat, and J. D. Roberts, J. Org. Chem., 1971, 36, 2757;

- (f) D. K. Dalling and D. M. Grant, J. Amer. Chem. Soc., 1967, 89, 6612;
(g) M. Cristl, H. J. Reich, and J. D. Roberts, ibid., 1971, 93, 3463;
(h) D. E. Dorman, S. J. Angyal, and J. D. Roberts, ibid., 1970, 92, 1351;
(i) J. K. Crandall and S. A. Sojka, ibid., 1972, 94, 5084;
(j) W. R. Woolfenden and D. M. Grant, ibid., 1966, 88, 1496;
(k) F. J. Weigert and J. D. Roberts, ibid., 1970, 92, 1347, and references cited therein.
10. Copies of the program, along with sample input decks, may be obtained from the authors.
 11. These isomer counts were computed using the enumeration algorithm of Henze and Blair, Reference 13.
 12. G. N. La Mar, J. Amer. Chem. Soc., 1971, 93, 1040.
 13. H. R. Henze and C. M. Blair, J. Amer. Chem. Soc., 1931, 53, 3042 and 3077.
 14. (a) "The Theory of Graphs and its Applications," C. Berge, Wiley and Sons, New York, 1964, p. 92;
(b) ibid., p. 99.
 15. On the PDP-10, the program runs more slowly by a factor of about four (central processor time only).
 16. Two structures are considered to be tied when their DELMIN values differ by 0.1 ppm or less.
 17. Exhaustive, irredundant methods for the generation of cyclic structures have recently been developed by L. M. Masinter and N. S. Sridharan as part of our Heuristic DENDRAL project. A manuscript describing their work is in preparation.
 18. S. Barcza and N. Engstrom, J. Amer. Chem. Soc., 1972, 94, 1762.

Table 1. The Substructures X Considered by the PRUNER.

Class	"Depth 3" X structure(s)	Class	"Depth 3" X structure(s)
1	$R-CH_2-C-$ } 1,2,3	13	$R-NH_2$
2	$R-CH_2-NH_2$	14	$R-NH-C-$ } 0,1,2,3
3	$R-CH_2-NH-C$	15	$R-N \begin{cases} C- \\ C- \end{cases}$ } 0,1,5,6
4	$R-CH_2-N \begin{cases} C \\ C \end{cases}$		$R-N \begin{cases} CH-C \\ CH_3 \end{cases}$
5	$R-CH \begin{cases} C- \\ C- \end{cases}$ } 1,2,...,6		$R-N \begin{cases} CH_2-C \\ CH_2-C \end{cases}$
6	$R-CH \begin{cases} NH_2 \\ C- \end{cases}$ } 0,1,2,3		$R-N \begin{cases} C-C \\ C \\ CH_3 \end{cases}$
7	$R-CH \begin{cases} NH-C \\ C- \end{cases}$ } 0,1,2,3		$R-N \begin{cases} CH-C \\ CH_2-C \end{cases}$
8	$R-CH \begin{cases} N \begin{cases} C \\ C \end{cases} \\ C- \end{cases}$ } 0,1,2,3		$R-N \begin{cases} C-C \\ C \\ CH_2-C \end{cases}$
9	$R-C \begin{cases} C- \\ C- \\ C- \end{cases}$ } 1,2,...,9		$R-N \begin{cases} CH-C \\ CH-C \end{cases}$
10	$R-C \begin{cases} NH_2 \\ C-2 \\ C- \end{cases}$ } 0,1,...,6		
11	$R-C \begin{cases} NH-C \\ C- \\ C- \end{cases}$ } 0,1,...,6		
12	$R-C \begin{cases} N \begin{cases} C \\ C \end{cases} \\ C- \\ C- \end{cases}$ } 0,1,...,6		

Table 2. Cases for which AMINE obtained only the correct structure using DELTA = 1.5 ppm and, except as noted, no tallies.

Amine (prefix only)	Amine (prefix only)
methyl	N-ethyl-diisopropyl
ethyl	nonyl
propyl	N-propylhexyl
isopropyl	N- <u>sec</u> -butylpentyl
trimethyl	N- <u>sec</u> -butyl-3-methylbutyl
butyl	N- <u>tert</u> -butyl-3-methylbutyl
<u>sec</u> -butyl	N-methyl-1,1,3,3-tetramethyl- butyl
isobutyl	tripropyl
<u>tert</u> -butyl	decyl
diethyl	dipentyl
pentyl	N-butylhexyl
1-methylbutyl	N- <u>tert</u> -butylhexyl
2-methylbutyl	N- <u>sec</u> -butyl-3,3-dimethylbutyl
3-methylbutyl	di(3-methylbutyl)
2,2-dimethylpropyl	N-ethyl-dibutyl
N-methyl- <u>sec</u> -butyl	N-ethyl-dibutyl
N-methyl- <u>tert</u> -butyl	N,N-diisopropylbutyl
N-methyl-diethyl	N-pentylhexyl
hexyl	N-butyl-1-methylhexyl
1,3-dimethylbutyl	N-pentyl-1,3-dimethylbutyl
1,2,2-trimethylpropyl	N-(3,3-dimethylbutyl)pentyl
2,2-dimethylbutyl	N-butyl-1-ethylpentyl
dipropyl	N-methyl-N-butylhexyl
diisopropyl	N-propyl-dibutyl
N-ethylbutyl	N-isopropyl-dibutyl
N-ethyl- <u>sec</u> -butyl	N-(1,3-dimethylbutyl)hexyl
triethyl	tributyl
N,N-dimethyl- <u>sec</u> -butyl	N-ethyl-dipentyl
N,N-dimethyl- <u>tert</u> -butyl	N- <u>tert</u> -butyl-dibutyl
heptyl	N,N-dibutyl-3-methylbutyl
1-methylhexyl	N,N-dibutylhexyl
1-ethylpentyl	N,N-dibutyl-3,3-dimethylbutyl
1,3-dimethylpentyl	N- <u>sec</u> -butyl-dipentyl
N-methylhexyl	N,N-dipentyl-1-methylpentyl
N-isopropylbutyl	tripentyl
N-isopropyl- <u>sec</u> -butyl	tri(3-methylbutyl)
octyl	
1-methylheptyl	
2-ethylhexyl	
1,5-dimethylhexyl	Using tallies:
1,1,3,3-tetramethylbutyl	di(2-ethylhexyl)
dibutyl	N,N-dipentyl-1,3-dimethylbutyl
diisobutyl	N,N-dibutyl-1,1,3,3-tetramethyl- butyl
N-ethylhexyl	trihexyl
N,N-dimethylhexyl	N-butyl-di(2-ethylhexyl)
N,N-diethylbutyl	
N,N-diethyl- <u>sec</u> -butyl	

Table 3. Cases for which AMINE obtained two or more structures using DELTA = 1.5 ppm and, except as noted, no tallies.

Amine (prefix only)	Solutions (prefix only)	Rank
dihexyl	dihexyl N-pentylheptyl ^a	tied tied
N-pentyl-1,1,3,3-tetra- methylbutyl	N-pentyl-1,1,3,3-tetra- methylbutyl N- <u>tert</u> -butyl-1,1-dimethyl- heptyl ^a	1 2
N-(1-ethylpentyl)-1-propyl- butyl	N-(1-ethylpentyl)-1-propyl- butyl N-(1-ethylbutyl)-1-propyl- pentyl	tied tied
N,N-dibutylheptyl	N,N-dibutylheptyl N-butyl-N-pentylhexyl ^a	1 2
dioctyl ^b	dioctyl N-heptylnonyl N-hexyldecyl	tied tied tied
trioctyl ^b	trioctyl N-heptyl-N-octyl-nonyl N,N-diheptyldecyl N-hexyldinonyl N-hexyl-N-octyldecyl N-hexyl-N-heptylundecyl N,N-dihexyldodecyl	tied tied tied tied tied tied tied

a) The use of tallies excludes these structures.

b) Tallies were used in these runs.

Table 4. Cases for which AMINE found no structures using DELTA = 1.5 ppm. The correct solutions appeared when DELTA was increased to 2.25 ppm and tallies were included.

Amine

1-isopropylhexylamine

N-pentyl-1,2,2-trimethylpropylamine^a

N-butyl-N-(1,2,2-trimethylpropyl)pentylamine

N-butyl-N-pentyl(1,1,3,3-tetramethylbutyl)amine

a) A second structure, equally ranked, was found in this case: N-propyl-N-(1,2,2-trimethylpropyl)hexylamine.

Table 5. Results obtained by AMINE for the four "unknown" amines.

Amine (prefix only)	Conditions		Solutions (prefix only)	Rank
	DELTA (ppm)	Tallies used?		
N-(3-methylbutyl)-1,5-dimethylhexyl	1.5	no	N-(3-methylbutyl)-1,5-dimethylhexyl	-
N-(3-methylbutyl)-2-ethylhexyl	1.5	no	N-(3-methylbutyl)-2-ethylhexyl	-
N-heptyl-N-(3-methylbutyl)-2-ethylhexyl	1.5	yes	N-heptyl-N-(3-methylbutyl)-2-ethylhexyl	1 (tied)
			N-pentyl-N-(3-methylbutyl)-2-ethylhexyl	1 (tied)
N-pentyl-N-(3,3-dimethylbutyl)- 3,5,5-trimethylhexyl	2.25 ^a	yes	2-ethyl-1,5,5,7,7-pentamethyl-1-(2,2-dimethylpropyl)octyl	1
			N-pentyl-N-(3,3-dimethylbutyl)-3,5,5-trimethylhexyl	2 (tied)
			N,N-di(<u>tert</u> -butyl)-2-methyl-2-(2,2-dimethylpropyl)hexyl	2 (tied)
			N- <u>tert</u> -butyl-1,1,3-trimethyl-3-(2,2-dimethylpropyl)octyl	2 (tied)
			2-ethyl-1,1,5,7,7-pentamethyl-5-(2,2-dimethylpropyl)octyl	2 (tied)

a) With DELTA = 1.5 ppm, no structures were found for this amine.

Figure captions

- Figure 1. A schematic illustration of R, the alkyl chain-end to be tested by the PRUNER. The group X contains the Nitrogen atom, along with any carbons and hydrogens not included in R.
- Figure 2. The hierarchy of pre-tests used by the PRUNER. A "?" attached to an atom indicates that the neighbors of that atom are unknown at testing time.
- Figure 3. A case in which \underline{r} and \underline{o} do not match when $n = N_c$, even though the simple test is passed.
- Figure 4. Sample output from program AMINE (PDP-10 version). The solution structure is written in polish-prefix notation as described in Reference 3a.

Figure 1

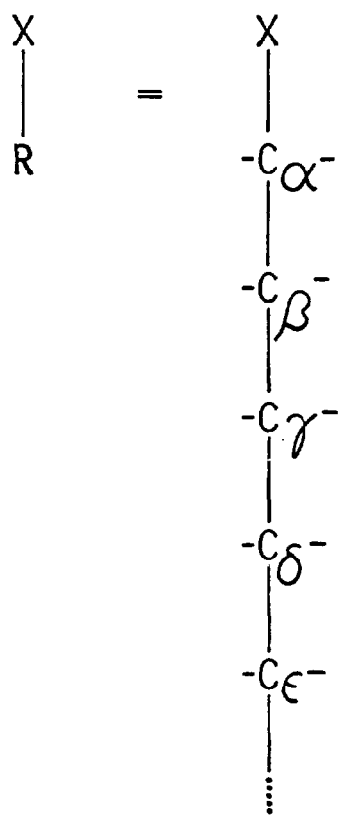


Figure 2.

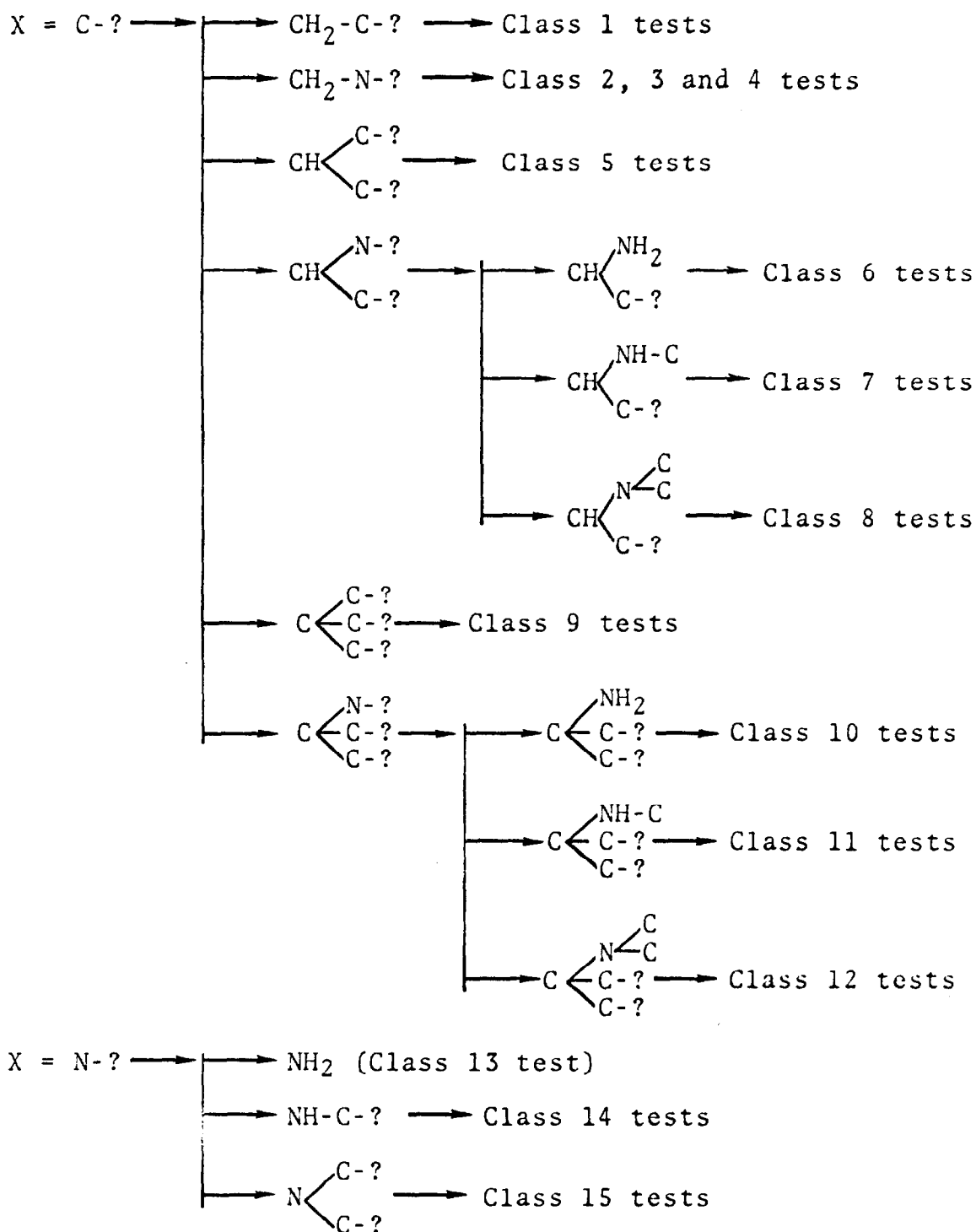
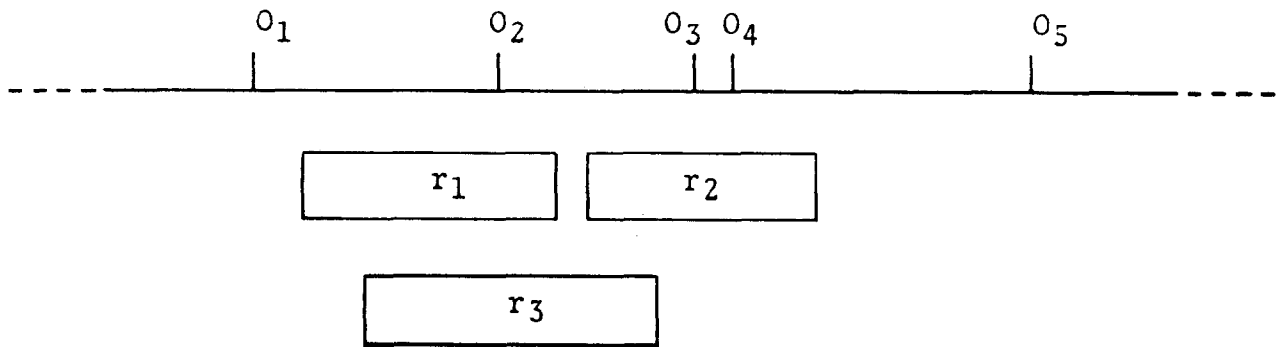


Figure 3.



CASE TITLE: N-ETYLDIPENTYLAMINE
THE AMINE HAS 12 CARBONS
GOODNESS-OF-FIT CRITERION IS 1.500
STANDARD IS TMS
INPUT SHIFTS: 54.00 27.80 30.10 23.00 14.30 47.8

SOLUTION STRUCTURES:

N...C.C.C.C.CC.C.C.C.C.C
SHIFTS: 54.299 27.881 30.239 22.960 14.210 54.29
27.881 30.239 22.960 14.210 47.667 12.86
DELMIN = 0.37

CASE FINISHED. PROCESSING TIME (IN SEC.) WAS 9.711

SIGNIFICANCE

SIGNIFICANCE

Because of the interdisciplinary character of this research, it has a significant impact in medicine, organic chemistry, and computer science. GC/MS has become one of the most powerful techniques available to the organic and biochemist. The potential applications of these techniques in medical research and in the clinic have just begun to be explored. These techniques are of unique importance to medical science since they alone of the current physical methods have sufficient sensitivity and analytical precision to study human biochemistry at the molecular level. Computer automation of these techniques, both at the instrumentation and interpretive levels, would permit the rapid, exhaustive analysis of body fluids across large populations of individuals in various medical contexts and may provide new discoveries important to public health.

In our study of errors of metabolism, accurate diagnosis of the accumulated metabolite provides insight into the biochemical pathogenesis and into therapeutic approaches to the control of such errors. In the case of inherited errors, accurate diagnosis allows reference to published data on the mode of inheritance and, thus, expresses the recurrence risk for genetic counseling purposes. The GC/MS system, with its potential for identification of any metabolites, provides the diagnostic accuracy necessary for a clinical program. GC/MS also provides the methodology for detecting previously unrecognized metabolic errors.

From the point of view of computer science, mass spectrometry is an advantageous environment in which to investigate the concepts necessary for the emulation of lower-level cognitive and manipulative functions as well as for the study of various forms of knowledge representation and automatic theory formation. These concepts will be common in some form to all "intelligent" systems and must be more fully developed from their present primitive state. Mass spectrometry is ideal as a milieu for this research in that it has tremendous practical importance to medicine, is sufficiently complex to challenge the human intellect, and is structured to an extent amenable to computer program formulation within the current state-of-the-art.

COLLABORATIVE ARRANGEMENTS

COLLABORATIVE ARRANGEMENTS

This project is an interdisciplinary research effort involving day-to-day collaboration between Professor J. Lederberg (Department of Genetics), Professor C. Djerassi (Department of Chemistry), Professor E. Feigenbaum (Department of Computer Science), Professor H. Cann (Department of Pediatrics), Dr. B. Buchanan (Computer Science), Dr. A. Duffield (Genetics), Dr. D. Smith (Chemistry), Dr. N. Sridharan (Computer Science), Dr. S. Hammerum (Chemistry), and the Instrumentation Research Laboratory of the Department of Genetics. We are also soliciting additional participation of clinical research interests of the Departments of Medicine and Psychiatry as well as other members of the Department of Genetics (Professors Cavalli-Sforza and Herzenberg). The proximity of these people and facilities in a medical environment offers a highly unique opportunity for collaborative interaction.

FACILITIES AVAILABLE

FACILITIES AVAILABLE

We will derive much of the clinically significant material for analysis from patients in the Premature Research Center and the Clinical Research Center of the Department of Pediatrics at Stanford. Analyses will be performed on existing gas chromatograph and mass spectrometer instrumentation. We have available a GC-coupled Finnigan 1015 quadrupole instrument in the Department of Genetics and a GC-coupled Varian-MAT 711 instrument in the Department of Chemistry. Also available in the Department of Chemistry are MS-9 and Varian-MAT Ch-4 instruments.

We will derive our computing resources from existing PDP-11/20 mini-computer systems which interface the mass spectrometer instruments as well as from the ACME follow-on 370/156 computer at Stanford for data reduction and graphics support. Artificial intelligence program development will be carried out on the Stanford Computation Center IBM 360/67 and machines available over the ARPA computer network. GC/MS data will be interfaced to these programs through standard communication links.

HUMAN SUBJECTS

HUMAN SUBJECTS

As a part of this research project, GC/MS analysis techniques will be applied to human body fluids in collaboration with clinical investigators and blood and urine specimens will be collected from human subjects. Collection of VOIDED URINE SPECIMENS presents no risk to the patient. Collection of 5-10 ml of blood by venepuncture is a procedure attended by minimal risk; infection is a remote possibility, especially from deep venepuncture (e.g. femoral tap). However, superficial veins are usually used in children, and even infants. It is only the occasional infant that requires a femoral tap and this procedure would be deferred for this project unless the specimen was essential for diagnosis.

BUDGETS AND JUSTIFICATION

In the following budget estimates, the abbreviations listed below are used to denote departmental affiliation or professional specialty:

G - Genetics

CS - Computer Science

Ch - Chemistry

E - Electrical Engineering

BUDGET - PART A

APPLICATIONS OF ARTIFICIAL INTELLIGENCE
TO MASS SPECTROMETRY

SUBSTITUTE DETAILED BUDGET FOR FIRST 12-MONTH PERIOD		PERIOD COVERED		GRANT NUMBER
		FROM 5/1/74	THROUGH 4/30/75	
1. PERSONNEL <i>(List all personnel engaged on project)</i>			TIME OR EFFORT	AMOUNT REQUESTED (Omit cents)
NAME <i>(Last, first, initial)</i>	TITLE OF POSITION	%/HRS.		TOTAL
Lederberg, Joshua	G Principal Investigator or Program Director	4	PART A	
Feigenbaum, Edward A. (1)	CS Co-Principal Invest.	10		
Buchanan, Bruce G. (1,2)	CS Associate Invest.	50		
Duffield, Alan	Ch Associate Invest.	25		
Smith, Dennis	Ch Research Associate	100		
Hammerum, Steen	Ch Research Associate	50		
Sridharan, Natesa	CS Research Associate	50		
Reiss, Steve	CS Computer Programmer	50		
Hjelmeland, Larry	CS Research Assistant	100		
Masinter, Larry	CS Research Assistant	50		
Stefik, Mark	CS Research Assistant	50		
Wharton, Kathy	Admin. Assistant	25		
Larson, Dee	Secretary	25		
(1) See Budget Notes				
(2) In first year only 9/1/74-4/30/75 covered				
			TOTAL →	\$ 80,624
2. CONSULTANT COSTS <i>(Include Fees and Travel)</i>				\$ 1,100
3. EQUIPMENT <i>(Itemize)</i>				-
4. SUPPLIES				
Office supplies				350
5. STAFF TRAVEL <i>(See Instructions)</i>				\$ 1,400
a. DOMESTIC				\$
b. FOREIGN				\$
6. PATIENT COSTS <i>(Separate Inpatient and Outpatient)</i>				\$ -
7. ALTERATIONS AND RENOVATIONS				\$ -
8. OTHER EXPENSES <i>(Itemize per Instructions)</i>				
Telephone, postage, etc. \$ 200				
Publication costs 700				
Computer terminal rent 3,200				
Computer usage costs 36,000				
				\$ 40,100
9.			Subtotal - Items 1 thru 8 →	\$ 123,574
FOR TRAINING GRANTS ONLY	10. TRAINEE EXPENSES <i>(See Instructions)</i>			
	a. STIPENDS	PREDOCTORAL	No. Proposed _____	\$
		POSTDOCTORAL	No. Proposed _____	\$
		OTHER <i>(Specify)</i>	No. Proposed _____	\$
		DEPENDENCY ALLOWANCE		\$
				TOTAL \$TIPEND EXPENSES →
b. TUITION AND FEES				\$
c. TRAINEE TRAVEL <i>(Describe)</i>				\$
11.			Subtotal - Trainee Expenses →	\$
12. TOTAL DIRECT COST <i>(Add Subtotals, Items 9 and 11, and enter on Page 1)</i>			→	\$ 123,574

BUDGET ESTIMATES FOR ALL YEARS OF SUPPORT REQUESTED FROM PUBLIC HEALTH SERVICE DIRECT COSTS ONLY (Omit Cents)							
DESCRIPTION	1ST PERIOD (SAME AS DE- TAILED BUDGET)	ADDITIONAL YEARS SUPPORT REQUESTED <i>(This application only)</i>					
		2ND YEAR	3RD YEAR	4TH YEAR	5TH YEAR	6TH YEAR	7TH YEAR
PERSONNEL COSTS	80,624	95,175	100,320				
CONSULTANT COSTS <i>(Include fees, travel, etc.)</i>	1,100	1,200	1,300				
EQUIPMENT	-	-	-				
SUPPLIES	350	400	450				
TRAVEL	DOMESTIC	1,400	1,600	1,800			
	FOREIGN						
PATIENT COSTS	-	-	-				
ALTERATIONS AND RENOVATIONS	-	-	-				
OTHER EXPENSES	40,100	45,450	50,000				
TOTAL DIRECT COSTS	123,574	143,825	153,870				
TOTAL FOR ENTIRE PROPOSED PROJECT PERIOD <i>(Enter on Page 1, Item 4)</i> →					\$ 421,269		
<p>REMARKS: <i>Justify all costs for the first year for which the need may not be obvious. For future years, justify equipment costs, as well as any significant increases in any other category. If a recurring annual increase in personnel costs is requested, give percentage. (Use continuation page if needed.)</i></p> <p style="text-align: center;">See attached budget justification notes.</p>							

BUDGET - PARTS B (i) AND B (ii)

MASS SPECTROMETER DATA SYSTEM DEVELOPMENT

AND

ANALYSIS OF THE CHEMICAL CONSTITUENTS OF BODY FLUIDS

SUBSTITUTE DETAILED BUDGET FOR FIRST 12-MONTH PERIOD		PERIOD COVERED		GRANT NUMBER
		FROM 5/1/74	THROUGH 4/30/75	
1. PERSONNEL (List all personnel engaged on project)			TIME OR EFFORT %/HRS.	AMOUNT REQUESTED (Omit cents)
NAME (Last, first, initial)	TITLE OF POSITION			TOTAL
Lederberg, Joshua	G Principal Investigator or Program Director	3	PART B (i) and (ii)	
Duffield, Alan	Ch Associate Investig.	25		
Pereira, Wilfred	Ch Research Associate	50		
Summons, Roger	Ch Post Doctoral Fellow	100		
Rindfleisch, Thomas	E Research Associate	100		
Veizades, Nicholas	E Research Engineer	100		
Reynolds, Walter	E Research Engineer	20		
Tucker, Robert	CS Computer Programmer	75		
Wegmann, Annemarie	Ch Sr. Research Assist.	100		
Steed, Ernest	E Research Engineer	10		
Pearson, Dale	E Electronics Tech.	60		
DeFrancisci, Richard	Machinist	20		
Allan, Muriel	Secretary	25		
				TOTAL
2. CONSULTANT COSTS (Include Fees and Travel)				\$ -
3. EQUIPMENT (Itemize) Computer Terminal				\$ 3,000 *
4. SUPPLIES Office supplies-\$750; chemicals, glassware, and lab apparatus-\$2,500; GC supplies (gases, phases, columns, etc.)-\$950; dry ice and liq. nitrogen-\$1,500; electronic supplies and parts-\$3,500; GC/MS data recording media-\$2,100; mini-computer supplies-\$1,500; mass spec. repairs and parts-\$7,600				\$ 20,400
5. STAFF TRAVEL (See Instructions)	a. DOMESTIC	east coast (\$500); 1 mid-west (\$350); 1 west coast (\$150)		\$ 1,000
	b. FOREIGN			\$ -
6. PATIENT COSTS (Separate Inpatient and Outpatient)				\$ -
7. ALTERATIONS AND RENOVATIONS Mass spectrometer laboratory air conditioning and power modifications				\$ 2,500
8. OTHER EXPENSES (Itemize per instructions) Telephone and data communications - \$1,200; Publication costs - \$1,000; Mini-computer maintenance contract - \$4,600; computing costs from ACME follow-on - \$64,000				\$ 70,800
9. Subtotal - Items 1 thru 8				\$ 237,530
FOR TRAINING GRANTS ONLY	10. TRAINEE EXPENSES (See Instructions)			
	a. STIPENDS	PREDOCTORAL	No. Proposed _____	\$
		POSTDOCTORAL	No. Proposed _____	\$
		OTHER (Specify)	No. Proposed _____	\$
		DEPENDENCY ALLOWANCE		\$
				TOTAL STIPEND EXPENSES
	b. TUITION AND FEES			\$
	c. TRAINEE TRAVEL (Describe)			\$
11. Subtotal - Trainee Expenses				\$
12. TOTAL DIRECT COST (Add Subtotals, Items 9 and 11, and enter on Page 1)				\$ 237,530

BUDGET ESTIMATES FOR ALL YEARS OF SUPPORT REQUESTED FROM PUBLIC HEALTH SERVICE							
DIRECT COSTS ONLY (Omit Cents)							
DESCRIPTION	1ST PERIOD (SAME AS DE- TAILED BUDGET)	ADDITIONAL YEARS SUPPORT REQUESTED <i>(This application only)</i>					
		2ND YEAR	3RD YEAR	4TH YEAR	5TH YEAR	6TH YEAR	7TH YEAR
PERSONNEL COSTS	139,830	148,066	156,775				
CONSULTANT COSTS <i>(Include fees, travel, etc.)</i>	-	-	-				
EQUIPMENT	3,000	3,000	3,000				
SUPPLIES	20,400	21,050	22,250				
TRAVEL	DOMESTIC	1,000	1,000	1,000			
	FOREIGN						
PATIENT COSTS	-	-	-				
ALTERATIONS AND RENOVATIONS	2,500	-	-				
OTHER EXPENSES	70,800	75,000	79,500				
TOTAL DIRECT COSTS	237,530	248,116	262,525				
TOTAL FOR ENTIRE PROPOSED PROJECT PERIOD <i>(Enter on Page 1, Item 4)</i> →					\$ 748,171		
<p>REMARKS: <i>Justify all costs for the first year for which the need may not be obvious. For future years, justify equipment costs, as well as any significant increases in any other category. If a recurring annual increase in personnel costs is requested, give percentage. (Use continuation page if needed.)</i></p> <p>See attached budget justification.</p>							

BUDGET - PART C

EXTENSION OF THE THEORY OF
MASS SPECTROMETRY BY COMPUTER

SUBSTITUTE DETAILED BUDGET FOR FIRST 12-MONTH PERIOD		PERIOD COVERED		GRANT NUMBER
		FROM 5/1/74	THROUGH 4/30/75	
1. PERSONNEL <i>(List all personnel engaged on project)</i>			TIME OR EFFORT %/HRS.	AMOUNT REQUESTED <i>(Omit cents)</i> TOTAL
NAME <i>(Last, first, initial)</i>	TITLE OF POSITION			
Lederberg, Joshua	G Principal Investigator or		3	PART C
Feigenbaum, Edward A. (1)	CS Program Director		10	
Buchanan, Bruce G. (1,2)	CS Co-Principal Invest.		50	
Sridharan, Natesa	CS Research Associate		50	
Hammerum, Steen	Ch Research Associate		50	
White, William	CS Computer Programmer		50	
Farrell, Carl	CS Research Assistant		100	
Wharton, Kathy	Admin. Assistant		25	
Larson, Dee	Secretary		25	
(1) See budget notes				
(2) Covers 9/1/74-4/30/75 in year 1				
TOTAL →				\$ 48,521
2. CONSULTANT COSTS <i>(Include Fees and Travel)</i>				\$ -
3. EQUIPMENT <i>(Itemize)</i>				\$ -
4. SUPPLIES				\$ 350
5. STAFF TRAVEL <i>(See Instructions)</i>	a. DOMESTIC			\$ 1,400
	b. FOREIGN			\$ -
6. PATIENT COSTS <i>(Separate Inpatient and Outpatient)</i>				\$ -
7. ALTERATIONS AND RENOVATIONS				\$ -
8. OTHER EXPENSES <i>(Itemize per instructions)</i>				
Telephone, postage, etc.		\$ 200		
Publication costs		700		
Computer terminal rental		1,600		
Computer usage		21,000		
				\$ 23,500
9. Subtotal - Items 1 thru 8 →				\$ 73,771
FOR TRAINING GRANTS ONLY	10. TRAINEE EXPENSES <i>(See Instructions)</i>			
	a. STIPENDS	PREDOCTORAL	No. Proposed _____	\$
		POSTDOCTORAL	No. Proposed _____	\$
		OTHER <i>(Specify)</i>	No. Proposed _____	\$
		DEPENDENCY ALLOWANCE		\$
	TOTAL \$TIPEND EXPENSES →			\$
b. TUITION AND FEES				\$
c. TRAINEE TRAVEL <i>(Describe)</i>				\$
11. Subtotal - Trainee Expenses →				\$
12. TOTAL DIRECT COST <i>(Add Subtotals, Items 9 and 11, and enter on Page 1)</i> →				\$ 73,771

SECTION II - PRIVILEGED COMMUNICATION

BUDGET ESTIMATES FOR ALL YEARS OF SUPPORT REQUESTED FROM PUBLIC HEALTH SERVICE DIRECT COSTS ONLY (Omit Cents)							
DESCRIPTION		1ST PERIOD (SAME AS DE TAILED BUDGET)	ADDITIONAL YEARS SUPPORT REQUESTED (<i>This application only</i>)				
			2ND YEAR	3RD YEAR	4TH YEAR	5TH YEAR	6TH YEAR
PERSONNEL COSTS		48,521	61,194	64,655			
CONSULTANT COSTS (Include fees, travel, etc.)		-	-	-			
EQUIPMENT		-	-	-			
SUPPLIES		350	400	450			
TRAVEL	DOMESTIC	1,400	1,600	1,800			
	FOREIGN						
PATIENT COSTS		-	-	-			
ALTERATIONS AND RENOVATIONS		-	-	-			
OTHER EXPENSES		23,500	27,650	30,450			
TOTAL DIRECT COSTS		73,771	90,844	97,355			
TOTAL FOR ENTIRE PROPOSED PROJECT PERIOD (<i>Enter on Page 1, Item 4</i>) →					\$ 261,970		
<p>REMARKS: <i>Justify all costs for the first year for which the need may not be obvious. For future years, justify equipment costs, as well as any significant increases in any other category. If a recurring annual increase in personnel costs is requested, give percentage. (Use continuation page if needed.)</i></p> <p>See attached budget justification notes.</p>							

BUDGET - PART D

APPLICATIONS OF CARBON(13) NUCLEAR MAGNETIC
RESONANCE SPECTROMETRY TO ASSIST IN CHEMICAL
STRUCTURE DETERMINATION

SUBSTITUTE DETAILED BUDGET FOR FIRST 12-MONTH PERIOD		PERIOD COVERED		GRANT NUMBER
		FROM	THROUGH	
		5/1/74	4/30/75	
1. PERSONNEL <i>(List all personnel engaged on project)</i>			TIME OR EFFORT %/HRS.	AMOUNT REQUESTED <i>(Omit cents)</i>
NAME <i>(Last, first, initial)</i>	TITLE OF POSITION			TOTAL
Djerassi, Carl ⁽¹⁾	Ch Principal Investigator or Program Director		3	
Carhart, Ray ⁽²⁾	Ch Post Doctoral Fellow		100	
Unnamed	Ch Post.Doc.Res.Assoc.		100	
Van Antwerp, Craig	Ch Research Assistant		50	
PART D				
TOTAL →				\$ 33,592
2. CONSULTANT COSTS <i>(Include Fees and Travel)</i>				\$ -
3. EQUIPMENT <i>(Itemize)</i>				\$ -
4. SUPPLIES Chemical supplies				\$ 900
5. STAFF TRAVEL <i>(See Instructions)</i>		a. DOMESTIC 1 east coast trip		\$ 500
		b. FOREIGN		\$ -
6. PATIENT COSTS <i>(Separate Inpatient and Outpatient)</i>				\$ -
7. ALTERATIONS AND RENOVATIONS				\$ -
8. OTHER EXPENSES <i>(Itemize per instructions)</i>				
		Publication costs and reproduction services	\$ 100	
		NMR instrument usage (25 hrs/month @ \$25/hour)	7,500	
		Computer usage	10,800	
				\$ 18,400
9. Subtotal - Items 1 thru 8 →				\$ 53,392
FOR TRAINING GRANTS ONLY	10. TRAINEE EXPENSES <i>(See Instructions)</i>			
	a. STIPENDS	PREDOCTORAL	No. Proposed _____	\$
		POSTDOCTORAL	No. Proposed _____	\$
		OTHER <i>(Specify)</i>	No. Proposed _____	\$
		DEPENDENCY ALLOWANCE		\$
	TOTAL STIPEND EXPENSES →			\$
b. TUITION AND FEES				\$
c. TRAINEE TRAVEL <i>(Describe)</i>				\$
11. Subtotal - Trainee Expenses →				\$
12. TOTAL DIRECT COST <i>(Add Subtotals, Items 9 and 11, and enter on Page 1)</i> →				\$ 53,392

SECTION II - PRIVILEGED COMMUNICATION

BUDGET ESTIMATES FOR ALL YEARS OF SUPPORT REQUESTED FROM PUBLIC HEALTH SERVICE DIRECT COSTS ONLY (Omit Cents)							
DESCRIPTION	1ST PERIOD (SAME AS DE- TAILED BUDGET)	ADDITIONAL YEARS SUPPORT REQUESTED <i>(This application only)</i>					
		2ND YEAR	3RD YEAR	4TH YEAR	5TH YEAR	6TH YEAR	7TH YEAR
PERSONNEL COSTS	33,592	53,178	56,176				
CONSULTANT COSTS <i>(Include fees, travel, etc.)</i>	-	-	-				
EQUIPMENT	-	-	-				
SUPPLIES	900	1,000	1,100				
TRAVEL	DOMESTIC	500	500	500			
	FOREIGN						
PATIENT COSTS	-	-	-				
ALTERATIONS AND RENOVATIONS	-	-	-				
OTHER EXPENSES	18,400	20,000	22,200				
TOTAL DIRECT COSTS	53,392	74,678	79,976				
TOTAL FOR ENTIRE PROPOSED PROJECT PERIOD <i>(Enter on Page 1, Item 4)</i> →					\$ 208,046		
<p>REMARKS: Justify all costs for the first year for which the need may not be obvious. For future years, justify equipment costs, as well as any significant increases in any other category. If a recurring annual increase in personnel costs is requested, give percentage. (Use continuation page if needed.)</p> <p>See attached budget justification notes.</p>							

COMPOSITE BUDGET -

PARTS A + B + C + D

SUBSTITUTE DETAILED BUDGET FOR FIRST 12-MONTH PERIOD		PERIOD COVERED		GRANT NUMBER	
		FROM 5/1/74	THROUGH 4/30/75		
1. PERSONNEL <i>(List all personnel engaged on project)</i>		TIME OR EFFORT %/HRS.	AMOUNT REQUESTED <i>(Omit cents)</i>		
NAME <i>(Last, first, initial)</i>	TITLE OF POSITION		TOTAL		
Lederberg, Joshua	G Principal Investigator or Program Director	10	COMPOSITE BUDGET		
Feigenbaum, Edward	CS Co-Principal Inves.	20			
Djerassi, Carl	Ch Co-Principal Inves.	3			
Buchanan, Bruce	CS Associate Inves.	100			
Duffield, Alan	Ch Associate Inves.	50			
Smith, Dennis	Ch Research Associate	100			
Sridharan Natesa	CS Research Associate	100			
Hammerum, Steen	Ch Research Associate	100			
Pereira, Wilfred	Ch Research Associate	50			
Rindfleisch, Thomas	E Research Associate	100			
Carhart, Ray	Ch Post Doctoral Fellow	100			
Summons, Roger	Ch Post Doctoral Fellow	100			
Unnamed	Ch Post Doc.Res.Assoc.	100			
See attached sheet					
TOTAL →			\$ 302,567		
2. CONSULTANT COSTS <i>(Include Fees and Travel)</i>			\$ 1,100		
3. EQUIPMENT <i>(Itemize)</i>					
Computer Terminal			\$ 3,000		
4. SUPPLIES					
See attached sheet			\$ 22,000		
5. STAFF TRAVEL <i>(See Instructions)</i>					
a. DOMESTIC		\$ 4,300			
b. FOREIGN		\$ -			
6. PATIENT COSTS <i>(Separate Inpatient and Outpatient)</i>			\$ -		
7. ALTERATIONS AND RENOVATIONS					
Mass spectrometer laboratory air conditioning and power modifications			\$ 2,500		
8. OTHER EXPENSES <i>(Itemize per instructions)</i>					
Telephone, data communications, postage, etc.			\$ 1,600		
Publication costs			\$ 2,500		
Mini-computer maintenance contract			\$ 4,600		
NMR Instrument usage			\$ 7,500		
Computer terminal rental			\$ 4,800		
Computer usage (ACME follow-on, Campus 360/67, and ARPANET)			\$131,800	\$ 152,800	
9. Subtotal - Items 1 thru 8 →			\$ 488,267		
FOR TRAINING GRANTS ONLY	10. TRAINEE EXPENSES <i>(See Instructions)</i>				
	a. STIPENDS	PREDOCTORAL	No. Proposed _____	\$	
		POSTDOCTORAL	No. Proposed _____	\$	
		OTHER <i>(Specify)</i>	No. Proposed _____	\$	
	DEPENDENCY ALLOWANCE			\$	
	TOTAL STIPEND EXPENSES →			\$	
b. TUITION AND FEES			\$		
c. TRAINEE TRAVEL <i>(Describe)</i>			\$		
11. Subtotal - Trainee Expenses →			\$ -		
12. TOTAL DIRECT COST <i>(Add Subtotals, Items 9 and 11, and enter on Page 1)</i>			\$ 488,267		

PERSONNEL (Continued)

<u>Name</u>		<u>Title of Position</u>	<u>Time or Effort</u>
Veizades, Nicholas	E	Research Engineer	100
Reynolds, Walter	E	Research Engineer	20
Steed, Ernest	E	Research Engineer	10
White, William	CS	Computer Programmer	50
Tucker, Robert	CS	Computer Programmer	75
Reiss, Steve	CS	Computer Programmer	50
Wegmann, Annemarie	Ch	Senior Research Assistant	100
Pearson, Dale	E	Electronics Technician	60
Hjelmeland, Larry	CS	Research Assistant	100
Masinter, Larry	CS	Research Assistant	50
Stefik, Mark	CS	Research Assistant	50
Farrell, Carl	CS	Research Assistant	100
Van Antwerp, Craig	Ch	Research Assistant	50
Wyche, Margaret		Laboratory Technician	50
DeFrancisci, Richard		Machinist	20
Wharton, Kathy		Administrative Assistant	50
Larson, Dee		Secretary	50
Allan, Muriel		Secretary	25

SUPPLIES

Office supplies	\$ 1,450
Chemicals, glassware, and laboratory apparatus	3,400
GC supplies (gases, phases, columns, etc.)	950
Dry ice and liquid nitrogen	1,500
Electronic supplies and parts	3,500
GC/MS data recording media (chart paper, Calcomp, etc.)	2,100
Mini-computer supplies (paper, ribbons, tapes, disks, etc.)	1,500
Mass spectrometer repairs and replacement parts	7,600
	<u>\$22,000</u>

DO NOT TYPE IN THIS SPACE-BINDING MARGIN

SECTION II - PRIVILEGED COMMUNICATION

BUDGET ESTIMATES FOR ALL YEARS OF SUPPORT REQUESTED FROM PUBLIC HEALTH SERVICE							
DIRECT COSTS ONLY (Omit Cents)							
DESCRIPTION	1ST PERIOD (SAME AS DE- TAILED BUDGET)	ADDITIONAL YEARS SUPPORT REQUESTED (This application only)					
		2ND YEAR	3RD YEAR	4TH YEAR	5TH YEAR	6TH YEAR	7TH YEAR
PERSONNEL COSTS	302,567	357,613	377,926				
CONSULTANT COSTS (Include fees, travel, etc.)	1,100	1,200	1,300				
EQUIPMENT	3,000	3,000	3,000				
SUPPLIES	22,000	22,850	24,250				
TRAVEL	DOMESTIC	4,300	4,700	5,100			
	FOREIGN	-	-	-			
PATIENT COSTS	-	-	-				
ALTERATIONS AND RENOVATIONS	2,500	-	-				
OTHER EXPENSES	152,800	168,100	182,150				
TOTAL DIRECT COSTS	488,267	557,463	593,726				
TOTAL FOR ENTIRE PROPOSED PROJECT PERIOD (Enter on Page 1, Item 4) →					\$ 1,639,456		
REMARKS: Justify all costs for the first year for which the need may not be obvious. For future years, justify equipment costs, as well as any significant increases in any other category. If a recurring annual increase in personnel costs is requested, give percentage. (Use continuation page if needed.)							

BUDGET DETAIL AND JUSTIFICATION

The budgets for the DENDRAL Project are presented in four parts, corresponding to the four proposal sections; A, B(i) and (ii), C, and D. Parts A and C represent the portions concerned with Heuristic and Meta-DENDRAL; Part B deals with the data system automation and instrument maintenance functions as well as the development aspects of GC/MS analysis of body fluids; and Part D is an extension of DENDRAL methodology to Carbon(13) nuclear magnetic resonance spectrometry.

As a general note, Professor Lederberg will devote a total of 10% of his time to this research as the Principal Investigator. His time is budgeted as follows: 4% on Part A, 3% on Part B, and 3% on Part C.

The narrative comments on Parts A and C have been combined below because the personnel and computer resources overlap to a large extent.

BUDGET EXPLANATION - PARTS A & C

PERSONNEL:

a) The personnel on the DENDRAL staff constitute its most valuable resource. All of the people listed in the proposal are now working on the DENDRAL Project. All are necessary to support the high level of scientific activity in Chemistry (A. Duffield, D. Smith, S. Hammerum, and L. Hjelmeland) and Computer Science (R. Feigenbaum, B. Buchanan, N. Sridharan, W. White, S. Reiss, M. Stefik, L. Masinter, and C. Farrell).

Mr. Mark Stefik's status will have changed to Research Assistant for Part A from his current status as Computer Programmer on Part B.

Mr. Steve Reiss' salary has been increased in order to properly compensate him for the duties he performs. Recent changes in draft board policies allow Conscientious Objectors to receive higher compensation to reflect actual job duties. Specific University approval has been requested for this increase but has not yet been received.

Mr. Larry Masinter has previously been paid from other funds, but is essential to the NIH-related work.

b) Salary figures are increased annually by 5% for merit increases and promotions. Fringe benefits are budgeted at the standard University rates of 17% through 8/74 and are increased annually per University projections to 18.3% in 9/74, 19.3% in 9/75, and 20.4% in 9/76.

No new personnel are added in Year 2. However, the salary budget increases by more than the rates noted above because all of Dr. Buchanan's salary is covered (see c) below) and Professor Feigenbaum returns from his leave of absence (see d) below).

c) Bruce Buchanan currently has an NIH Career Development Award through 8/31/76. However, because of recent NIH budget cutbacks, there is a strong probability that this award will be cancelled before that date. Dr. Ferguson of NIH stated on 2/8/73 that the award could only be guaranteed through 8/74.

d) As noted in the Introduction to this proposal, Dr. Feigenbaum will be on leave of absence with ARPA for a period of two years. This overlaps the term of this grant application such that no salary is budgeted for Dr. Feigenbaum during the first grant year. His salary is budgeted starting in the second grant year when he will formally return to his position in this research project.

EQUIPMENT:

No equipment purchases are required for Parts A and C.

SUPPLIES AND TRAVEL:

Office supplies are budgeted based on our experience over the past year. The travel budget covers expected costs for attending professional meetings and maintaining contact with related work at other locations. Because Artificial Intelligence is a rapidly expanding field, it is essential to maintain a high degree of personal interaction in order to assimilate new developments. These budget items are increased and rounded at 10% per year.

OTHER EXPENSES:

Telephone costs include connections and usage for computer terminals. Publication costs are budgeted at a nominal rate based on past experience and are increased by 10% per year. In the category of Computer Terminal Rent, the budget for Part A includes the lease cost of 2 portable Texas Instruments terminals. An additional terminal is added in 5/75 to accommodate increased use of the programs by personnel and a larger community of Stanford users. The Part C budget covers the continued lease of one T.I. terminal and an additional terminal starting in 5/75.

Computer time is budgeted according to current rate structures based on our on-going experience in utilizing the Stanford (SCC) 360/67 and machines available via the ARPANET. We will not make use of the ACME follow-on machine (370/158) for Parts A and C because of the availability of superior LISP facilities on these other machines. Instrument data will be communicated from the 370/158 (see Part B) to the LISP programs for analysis.

BUDGET EXPLANATION - PARTS B(i) AND B(ii)

This budget covers instrumentation maintenance, data system development, and research into applications of GC/MS analysis of body fluids as described in Parts B(i) and B(ii) of this proposal. This budget represents a significant increase over that submitted for Part B of the DENDRAL grant currently in-progress (current budget \$80,000 per year). The major reasons for this increase are twofold: a) Increases in required personnel support because of corresponding decreases in support from other sources and b) The need to implement our computing support from a source other than the ACME 360/50 for which NIH funding is terminating. We have rigorously attempted to keep these increases to an absolute minimum consistent with maintaining the viability of our unaugmented research program.

We have previously received substantial support for our GC/MS research from NASA. Because of shifting federal priorities, however, NASA support has declined substantially and we project will terminate in the first year of this renewal. At the same time, our research has been moving to emphasize more and more heavily GC/MS applications in clinically related aspects of metabolic indicators of disease. Thus it is reasonable, as well as necessary, that support for this continued research shift to NIH.

As mentioned in the Introduction to this proposal, we have an application pending with NIH-GMS for support in applying these techniques to aspects of genetic disease. These proposals are complementary in goals and it is assumed in this budget that the Genetics Center proposal will provide support for a major fraction (approximately 50%) of the low resolution GC/MS laboratory (Finnigan 1015 instrument) including personnel, supplies, etc. There is, however, a small amount of operational manpower overlap between the two proposed efforts. If both proposals are funded, a savings will result through common operational support which will be negotiated with NIH at the appropriate time.

As discussed under future plans for Part B(i) of this proposal, we have had to plan an alternative source of computing to support this research because NIH subsidy of the ACME facility terminates in July 1973. We have chosen to use the Stanford-sponsored follow-on to ACME, mounted on an IBM 370/158, since our computer programs will operate with a minimum of modification. This facility will operate on a fee-for-service basis. Whereas its rate structure is still evolving, we have estimated, on the basis of available information, the cost of transferring our computing to that facility as reflected in our budget (\$64,000 per year). It should be noted that this rate structure does not include indirect charges at this time. As the rate structure becomes better defined, the indirect cost may be

included in the usage rates. This would necessitate a slight modification of the budget as will be negotiated with NID as appropriate.

The following gives a detailed description of the various components of the Part B budget:

PERSONNEL:

The personnel budgeted for GC/MS applications, laboratory operations, and data system development are necessary to achieve our research goals and are currently active in the GC/MS programs. Chemistry support for the interpretation of body fluid analyses in cooperation with our clinical collaborators include Drs. A. Duffield (25%), W. Periera (50%), and R. Summons (100%). M. Wyche provides laboratory and instrument operation support for the low resolution GC/MS laboratory. Messers Rindfleisch, Veizades, Reynolds, and Tucker are essential to the data system development effort and provide hardware and software maintenance support as well. Messers Rindfleisch (100%) and Tucker (75%) are primarily responsible for the software system design, implementation, and maintenance. Mr. Veizades (100%) is primarily concerned with the hardware maintenance and development aspects of the high resolution MAT-711 instrument and Mr. Reynolds (20%) with the Finnigan 1015 low resolution instrument. Ms. A. Wegmann (100%) is responsible for the operation of the high resolution GC/MS instrument (MAT-711). Mr. Steed (10%) provides necessary glasswork development and maintenance, Mr. Pearson (60%) supports the fabrication and repair of electronic hardware for both instruments, and Mr. DeFrancisci (20%) provides necessary machinist support for mechanical repairs and fixtures. Ms. Allan (25%) provides required secretarial support for the above Instrumentation Research Laboratory personnel.

This manpower complement is carried into the future years as shown. Salaries are increased by 5% per year and staff benefits are applied at standard University rates. These start at 17% in fiscal year 1974 (9/73 - 8/74) and increase to 18.3% in 9/74, 19.3% in 9/75, and 20.4% in 9/76 based on University projections.

EQUIPMENT:

Our request for additional equipment is minimal. We budget for the purchase of a computer terminal in the first year for \$3,000. This replaces a currently rented terminal integral to the GC/MS data system and saves \$5,280 over the three year

grant period by purchasing instead of continued rental.

In the second year we budget for an event counter necessary for proper equipment maintenance for which we are assuming responsibility. We already maintain the Finnigan 1015 instrument and will take over the MAT-711 because of progressively poorer performance by VARIAN Associates in maintaining that instrument over the past year. This equipment is also needed to implement experimental control functions on the mass spectrometer.

In the third year, replacement of outdated test equipment will be required. \$3,000 are budgeted for this purpose.

SUPPLIES:

Supplies are budgeted based on our actual operating experience and are minimized consistent with a viable research effort. Office supplies include stationery supplies, postage, reproduction services, etc. and are budgeted at \$63 per month. The budget for chemicals, glassware, and laboratory apparatus (\$2,500) provides the necessary materials for derivatizing and analyzing body fluid samples. GC supplies (\$950) and dry ice and liquid nitrogen (\$1,500) are necessary for instrument operation and are based on past experience. The largest part of the liquid nitrogen budget is used for the high resolution instrument. Electronics supplies and parts (\$3,500) include circuit boards, semi-conductors, etc. needed for mass spectrometer control electronics such as for the metastable acquisition system as well as for maintaining our existing test equipment (oscilloscopes, voltmeters, power supplies, etc.). GC/MS data recording media (\$2,100) include chart and Calcomp plotter papers of various types (including 6V-sensitive paper for the MAT-711) for the purpose of recording mass spectrometer and gas chromatograph effluent data. The budgeted amount reflects our usage over the past year. Similarly, mini-computer supplies (\$1,500) include Teletype and line printer paper and ribbons, magnetic tapes (DEC tape and IBM compatible tape), and disk cartridges based on previous usage history. The budget for mass spectrometer repairs and replacement parts (\$7,600) covers our maintenance of these instruments based in part on predictable replacements (filaments, multipliers, etc.) and in part on an estimate from previous experience of unscheduled problems (power supplies, valves, pumps, etc.).

The supplies budget for future years covers these same items with 6% added for increased usage and inflation.

TRAVEL:

We have budgeted for travel to attend professional meetings and to visit other GC/MS laboratories on the basis of 1 east coast trip (\$500), 1 mid-west trip (\$350), and 1 west coast trip (\$150).

ALTERATIONS AND RENOVATIONS:

We have had problems with thermal overloads on the high resolution mass spectrometer instrument and associated electronics during the summer months. In addition, because of the modified computing configuration required by the ACME transition, we will locate a disk and printer equipment in the same laboratory to support the mini-computer interfacing the MAT-711. These conditions require an augmentation to existing air-conditioning and power facilities in the laboratory estimated at \$2,500.

OTHER EXPENSES:

We budget for telephone and data communications service based on our current experience (\$100 per month). In addition, \$1,000 is budgeted for publication costs and \$4,600 for mini-computer maintenance. This maintenance is an extension of our current contract with Digital Equipment Corporation and includes the prevailing 10% discount in the Stanford/DEC contract.

We budget for data reduction and storage computing costs on the ACME follow-on machine (370/158) as follows, based on our ACME experience and current information on the follow-on system rate structure. We consume approximately 300,000 page-minutes of computing per month on ACME for development and production computing. At a rate of \$.02 per page-minute, this comes to \$6,000 per month. In addition, we use approximately \$2,000 per month for data storage (20,000 blocks at \$.10 per block per month). This gives a total of \$96,000 per year and applying a projected 30% discount rate for high volume usage, leaves an estimated net cost of \$64,000 per year.

These estimates are increased by 6% in succeeding years for increased usage and inflation.

BUDGET EXPLANATION - PART D

This budget covers the portion of the research program which extends the DENDRAL methodology to Carbon(13) Nuclear Magnetic Resonance Spectrometry.

PERSONNEL:

The personnel budget includes a salary for Dr. R. Carhart after the expiration of his NIH Fellowship in 8/74, one Post Doctoral Research Associate (to be added to the staff), and one half-time Research Assistant (Mr. Van Antwerp). No funding is requested for Dr. Carl Djerassi's time (3%). A Computer Programmer (to be added to the staff) is budgeted in 1975 to assume the additional anticipated programming duties.

Salaries are increased by 5% per year and staff benefits are applied at standard University rates. These start at 17% in fiscal year 1974 (9/73 - 8/74) and increase per University projections to 18.3% in 9/74, 19.3% in 9/75, and 20.4% in 9/76.

SUPPLIES:

We budget \$900 for chemical supplies for the preparation of test samples.

TRAVEL:

We budget \$500 to cover one east coast trip.

OTHER EXPENSES:

Other expenses include \$100 for publication and reproduction costs and \$7,500 for usage of the existing NMR instrument in the Department of Chemistry. This NMR usage is budgeted at standard rates covering 25 hours of usage per month at \$25 per hour. In addition, we budget for use of the Stanford (SCC) 360/67 computer where CMR analysis programs, at the current level of development, are run. These costs are computed on the

basis of 1.5 hours of usage per month at approximately \$600 per hour.

BIOGRAPHIES

NAME	TITLE	BIRTHDATE (Mo., Day, Yr.)
LEDERBERG, JOSHUA	Professor and Executive Head, Department of Genetics	5-23-25
PLACE OF BIRTH (City, State, Country)	PRESENT NATIONALITY (If non-U.S. citizen, Indicate kind of visa and expiration date)	SEX
Montclair, New Jersey	U.S.A.	<input checked="" type="checkbox"/> Male <input type="checkbox"/> Female

EDUCATION (Begin with baccalaureate training and include postdoctoral)

INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Columbia College, New York College of Physicians & Surgeons, Columbia University, New York (1944-46)	B.A.	1944	
Yale University	Ph.D.	1947	Microbiology

HONORS

1957 - National Academy of Sciences
1958 - Nobel Prize in Medicine

MAJOR RESEARCH INTEREST	ROLE IN PROPOSED PROJECT
Molecular Genetics; Artificial Intelligence	PRINCIPAL INVESTIGATOR

RESEARCH SUPPORT (See instructions)

SEE ATTACHMENTS:

RESEARCH AND/OR PROFESSIONAL EXPERIENCE (Starting with present position, list training and experience relevant to area of project. List all or most representative publications. Do not exceed 3 pages for each individual.)

1961- Stanford University
Director, Kennedy Laboratories for Molecular Medicine

1959- Professor, Genetics and Biology, and Executive Head, Department of
Genetics, Stanford University

1957-1959 University of Wisconsin
Chairman, Department of Medical Genetics

1957 Melbourne University, Australia
Fullbright Visiting Professor of Bacteriology

1950 University of California, Berkeley
Visiting Professor of Bacteriology

1947-1959 University of Wisconsin
Professor of Genetics

1946-1947 Yale University. Research Fellow of the Jane Coffin Childs Fund for
Medical Research

1945-1946 Columbia University. Research Assistant in Zoology

Professional Activities:

1967- NIMH: National Mental Health Advisory Council

1961-1962 President (Kennedy)'s Panel on Mental Retardation

1960- NASA Committees: Lunar and Planetary Missions Board

1958- National Academy of Sciences: Committees on Space Biology

1950- President's Science Advisory Committee panels: National Institutes
of Health, National Science Foundation study sections (genetics)

RESEARCH SUPPORT SUMMARY FOR JOSHUA LEDERBERG

Grant Number	Grant Title	Current Year	Total Award	Grant Term	Budgeted % Time
1) NASA:NGR-05-020	Cytochemical Studies of Planetary Micro-organisms	\$ 180,000	\$3,800,000	9/60-8/73 (Future support dubious)	4%
2) NIH:AI-05160	Genetics of Bacteria	60,000	280,000	9/68-8/73 (Renewal pending)	15%
3) NIH:RR-00311	Advanced Computer for Medical Research (ACME) Stanford Medical School Facility	362,632	2,612,632 (yrs 4-7)	1966-7/73 (see #5)	25%
4) NIH:GM-	Genetics Research Center (J. Lederberg, Principal Investigator)	547,035	2,609,383	9/73-8/78 (Pending)	10%
5) NIH:RR-00785	Stanford University Medical Experimental Computer Facility (SUMEX) Successor to #3	884,660	5,960,417	9/73-8/78 (Pending)	20%
6) NIH: Computer Laboratory Health Care Resource Program	Large Scale Screening of Body Fluids for Metabolic Signs of Disease with Computer-managed Gas Chromatography and Mass Spectrometry	159,881	900,238	9/73-8/78 (Pending, Program funds impounded)	10%
7) NIH:GM00295	Training Grant in Genetics	143,964	756,650	7/69-6/73 (Renewal pending)	15%

SELECTED LIST OF PUBLICATIONS

Lederberg, J., 1959

A View of Genetics

Les Prix Nobel en 1958: 170-89.

Buchs, A., A. B. Delfino, A. M. Duffield, C. Djerassi, B. G. Buchanan,
E. A. Feigenbaum, and J. Lederberg, 1970.

Applications of Artificial Intelligence for Chemical Inference.

VI. Approach to a general method of interpreting low resolution
mass spectra with a computer. Helveticia Chimica Acta 53 (6): 1394-1417.

Feigenbaum, E. A., B. G. Buchanan, J. Lederberg, 1971

On generality and problem solving: a case study using the DENDRAL
program in Machine Intelligence 6, (B. Meltzer and D. Michie, eds.),
Edinburgh University Press, P. 165-190.

Reynolds, W. E., V. A. Bacon, J. C. Bridges, T. C. Coburn, B. Halpern,
J. Lederberg, E. C. Levinthal, E. Steed, R. B. Tucker, 1970

A Computer Operated Mass Spectrometer System.

Analytical Chem. 42:1122-1129, September 1970.

Lederberg, J.

"Use of Computer to Identify Unknown Compounds: The Automation of
Scientific Inference" in Biochemical Applications of Mass Spectrometry
(G. R. Waller, ed.). John Wiley & Sons, New York (in press).

BIOGRAPHICAL SKETCH

(Give the following information for all professional personnel listed on page 3, beginning with the Principal Investigator. Use continuation pages and follow the same general format for each person.)

NAME	TITLE	BIRTHDATE (Mo., Day, Yr.)
Carl DJERASSI	Professor of Chemistry	October 29, 1923
PLACE OF BIRTH (City, State, Country)	PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date)	SEX
Vienna, Austria	U.S.A.	<input checked="" type="checkbox"/> Male <input type="checkbox"/> Female

EDUCATION (Begin with baccalaureate training and include postdoctoral)

INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Kenyon College	A.B. (summa cum laude)	1942	Chemistry, Biology
University of Wisconsin	Ph.D.	1945	Organic chemistry, Biochemistry (minor)

HONORS Hon. D.Sc., Natl. Univ. of Mexico (1953), Kenyon College (1958), Worcester Polytechnic Institute (1972); Hon. Prof., Fed. Univ. Rio de Janeiro (1969). Member U.S. National Academy of Sciences, American Academy of Arts and Sciences, foreign member, Royal Swedish Academy of Sciences, German Academy of Natural Scientists (Leopoldina), Brazilian Academy of Sciences, (cont. below)

MAJOR RESEARCH INTEREST	ROLE IN PROPOSED PROJECT
Nat. prod. chemistry (steroids, alkaloids, terpenoids, antibiotics) and chem. applications of physical methods (mass spec., optical rotatory dispersion, circular dichroism).	Principal Investigator

RESEARCH SUPPORT (See instructions)	Grant	Title	Period	Current Year	Total Budgeted	% Time Effort
	NIH AM 04257	Mass Spectrometry in Organic and Biochemistry	10/1/70 to 9/30/75	\$56,833	\$316,016	10%
	NIH GM AM 06840-15	Marine Chemistry with special emphasis on steroids	1/1/73 to 12/31/77	112,550	578,180	18%

This is a pending application which, if approved, will represent a renewal of my current NIH Grants No. GM 06840 and No. AMCA-12785, both of which expire in 1973.

RESEARCH AND/OR PROFESSIONAL EXPERIENCE (Starting with present position, list training and experience relevant to area of project. List all or most representative publications. Do not exceed 3 pages for each individual.)

Academic Experience:

Professor of Chemistry, Stanford University, 1959-present.
Associate Professor (1952-1954) and Professor (1954-1959), Wayne State University.

Industrial Research Experience:

Ciba Pharmaceutical Co., Summit, N.J.: Research Chemist, 1942-1943 and 1945-1949.
Syntex Corporation: Associate Director of Chemical Research (Mexico City) 1949-1952, Research Vice President (Mexico City) 1957-1960; (Palo Alto, California) 1960-1968, President, Syntex Research 1968-present.

Editorial Boards:

(Current) Journal of the American Chemical Society, Steroids, Tetrahedron, Organic Mass Spectrometry.

(continued on next page)

Honors (cont.)

Mexican Academy for Scientific Investigation. Hon. Fellow of Phi Lambda Upsilon. Amer. Academy of Pharmaceutical Sciences, British Chemical Society and Mexican Chemical Society, Phi Beta Kappa. Numerous hon. lectureships including 1964 Centenary Lecturer (The British Chemical Society) and 1969 Annual Chemistry Lecturer, Royal Swedish Academy of Engineering. American Chemical Society Award in Pure Chemistry (1958), Baekeland Medal (1959), Fritzsche Award (1960). Intra-Science Research Foundation Award (1969). Freedman Patent Award of American Institute of Chemists (1971). Foreign Member, Royal Swedish Academy of Sciences (1972). D.Sc. (hon.), Worcester Polytechnic Institute (1972). Scheele-Lecturer, Pharmaceutical Society of Sweden (1972); American Chemical Society's Award for Creative Invention (1973).

RESEARCH AND/OR PROFESSIONAL EXPERIENCE (cont.)

Miscellaneous:

Chairman of the AAAS Gordon Research Conferences on Steroids and Natural Products (1952-1954); Member of American Pugwash Committee (1968 to present); Chairman of Latin America Science Board of National Academy of Sciences (1966-1968); Chairman of National Academy's Board on Science and Technology for International Development.

PUBLICATIONS

Author or co-author of 750 publications and six books. Approximately 150 papers and one book deal with various applications of chiroptical methods in organic and biochemistry.

DO NOT TYPE IN THIS SPACE-BINDING MARGIN

BIOGRAPHICAL SKETCH

(Give the following information for all professional personnel listed on page 3, beginning with the Principal Investigator. Use continuation pages and follow the same general format for each person.)

NAME Feigenbaum, Edward A.	TITLE Principal Investigator, DENDRAL Project	BIRTHDATE (Mo., Day, Yr.) 1-20-36	
PLACE OF BIRTH (City, State, Country) Weehawken, New Jersey	PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) U.S. Citizen	SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female	
EDUCATION (Begin with baccalaureate training and include postdoctoral)			
INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Carnegie Institute of Technology Pittsburgh, Pennsylvania	B.S. Ph.D.	1956 1959	Electrical Engineering Behavioral Sciences.
HONORS and memberships: American Psychological Association; Association for Computing Machinery (Member of the National Council 1966-68); American Association for the Advancement of Science.			
MAJOR RESEARCH INTEREST Artificial Intelligence	ROLE IN PROPOSED PROJECT Principal Investigator		
RESEARCH SUPPORT (See instructions)			

RESEARCH AND/OR PROFESSIONAL EXPERIENCE (Starting with present position, list training and experience relevant to area of project. List all or most representative publications. Do not exceed 3 pages for each individual.)

1965- Stanford University, Computer Science Department Faculty
 1965-1968 Stanford University, Director, Computation Center
 1963 Summer Research Training Institute in Computer Simulation of Cognitive Processes (National Science Foundation)
 1962 Carnegie Corporation. Summer Research Training Institute in Heuristic Programming. Faculty member.
 1960-1964 University of California, Berkeley
 Research-Center for Research in Management Science, 1960-1964
 Research-Center for Human Learning, 1961-1964
 Assistant and Associate Professor, School of Business Administration, 1960-64
 1957-1960 The RAND Corporation, Santa Monica, California
 1956 IBM Scientific Computing Center, New York

Selected Publications:

"Applications of Artificial Intelligence for Chemical Inference I. The Number of Possible Organic Compounds. Acyclic Structures Containing C, H, O and N", J. Am. Chem. Soc., 91, 2973 (1969). (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference II. Interpretation of Low Resolution Mass Spectra of Ketones", J. Am. Chem. Soc., 91, 2977 (1969). (Co-Author).

Publications of Edward Feigenbaum

"Applications of Artificial Intelligence for Chemical Inference III. Aliphatic Ethers Diagnosed by their Low Resolution Mass Spectra and Nuclear Magnetic Resonance", J. Am. Chem. Soc., 91, 7440 (1969). (Co-Author).

"Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry", in Machine Intelligence 4, Edinburgh University Press, 1969. (Co-Author).

"Toward an Understanding of Information Processes of Scientific Inference in the Context of Organic Chemistry", in Machine Intelligence 5, Edinburgh University Press, 1970. (Co-Author).

"A Heuristic Program for Solving a Scientific Inference Problem: Summary of Motivation and Implementation", Stanford Artificial Intelligence Project Memo No. 104, November 1969. (Co-Author).

"Applications of Artificial Intelligence For Chemical Inference IV. Saturated Amines Diagnosed by Their Low Resolution Mass Spectra and Nuclear Magnetic Resonance Spectra", Journal of the American Chemical Society, 92, 6831 (1970). (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference V. An Approach to the Computer Generation of Cyclic Structures. Differentiation Between All the Possible Isomeric Ketones of Composition C₆H₁₀O", Organic Mass Spectrometry, 4, 493 (1970). (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference VI. Approach to a General Method of Interpreting Low Resolution Mass Spectra with a Computer", Chem. Acta Helvetica, 53, 1394 (1970). (Co-Author).

"On Generality and Problem Solving: A Case Study Using the DENDRAL Program", in Machine Intelligence 6, Edinburgh University Press (1971). (Co-Author).

"A Heuristic Programming Study of Theory Formation in Science", in proceedings of the Second International Joint Conference on Artificial Intelligence, Imperial College, London (September 1971). (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference VIII. An Approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids", Journal of the American Chemical Society, 94, 5962-5971 (1972). (Co-Author).

"Heuristic Theory Formation: Data Interpretation and Rule Formation", in Machine Intelligence 7, Edinburgh University Press (1972). (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference X. Datsum. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids", to be submitted. (Co-Author).

BIOGRAPHICAL SKETCH

(Give the following information for all professional personnel listed on page 3, beginning with the Principal Investigator.
Use continuation pages and follow the same general format for each person.)

NAME Buchanan, Bruce G.	TITLE Research Computer Scientist	BIRTHDATE (Mo., Day, Yr.) 7-7-40	
PLACE OF BIRTH (City, State, Country) St. Louis, Missouri	PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) U.S. Citizen	SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female	
EDUCATION (Begin with baccalaureate training and include postdoctoral)			
INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Ohio Wesleyan University	B.A.	1961	Mathematics
Michigan State University	M.A., Ph.D.	1966	Philosophy
HONORS Recipient of National Institutes of Health Career Development Award (1971-1976) Invited Speaker at 1972 National Institutes of Health Symposium on Numerical Methods in Chemistry (Washington)			
MAJOR RESEARCH INTEREST	ROLE IN PROPOSED PROJECT Associate Investigator		
RESEARCH SUPPORT (See instructions)			

RESEARCH AND/OR PROFESSIONAL EXPERIENCE (Starting with present position, list training and experience relevant to area of project. List all or most representative publications. Do not exceed 3 pages for each individual.)

1972-present Research Computer Scientist, Stanford University
1966-1971 Research Associate, Stanford Artificial Intelligence Project

Publications:

"On the Design of Inductive Systems: Some Philosophical Problems". British Journal for the Philosophy of Science 20 (1969), 311-323. (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference II. Interpretation of Low Resolution Mass Spectra of Ketones". Journal of the American Chemical Society, 91, 2977-2981 (1969). (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference I. The Number of Possible Organic Compounds: Acyclic Structures Containing C, H, O and N". Journal of the American Chemical Society, 91, 2973-2976 (1969). (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference III. Aliphatic Ethers Diagnosed by Their Low Resolution Mass Spectra and NMR Data". Journal of the American Chemical Society, 91, 7440-45 (1969). (Co-Author).

"Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry". Machine Intelligence 4, Edinburgh University Press (1969). (Co-Author).

Publications of Bruce Buchanan:

"Toward an Understanding of Information Processes of Scientific Inference in the Context of Organic Chemistry". Machine Intelligence 5, Edinburgh University Press (1969). (Co-Author).

"On Generality and Problem Solving: A Case Study Using the DENDRAL Program". Machine Intelligence 6, Edinburgh University Press (1969). (Co-Author).

"Some Speculation About Artificial Intelligence and Legal Reasoning". Stanford Law Review, Vol. 23, No. 1, November 1970. (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference VI. Approach to a General Method of Interpreting Low Resolution Mass Spectra with a Computer". *Chemica Acta Helvetica*, 53, 1394 (1970). (Co-Author).

"An Application of Artificial Intelligence to the Interpretation of Mass Spectra". *Mass Spectrometry Techniques and Appliances* (1970).

"Applications of Artificial Intelligence for Chemical Inference IV. Saturated Amines Diagnosed by Their Low Resolution Mass Spectra and Nuclear Magnetic Resonance Spectra". *Journal of the American Chemical Society*, 93, 6831 (1970). (Co-Author).

"The Heuristic DENDRAL Program for Explaining Empirical Data". *Proceedings of IFIP Congress 1971, Ljubljana, Yugoslavia*. (Co-Author).

"A Heuristic Programming Study of Theory Formation in Science". *Proceedings of Second International Joint Conference on Artificial Intelligence, Imperial College, London* (1971). (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference VIII. An Approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids". *Journal of the American Chemical Society*, 1972. (Co-Author).

"Heuristic Theory Formation: Data Interpretation and Rule Formation". *Machine Intelligence 7, Edinburgh University Press* (1972). (Co-Author).

"Review of Hubert Dreyfus' 'What Computers Can't Do: A Critique of Artificial Reason'", *Computing Reviews* (January, 1973).

"Applications of Artificial Intelligence for Chemical Inference IX. Analysis of Mixtures Without Prior Separation as Illustrated for Estrogens". Submitted to the *Journal of the American Chemical Society*. (Co-Author).

"Applications of Artificial Intelligence for Chemical Inference X. Datsum. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids". To be submitted. (Co-Author).

Memberships

Association for Computing Machinery (ACM)

Philosophy of Science Association

American Association for Advancement of Science (AAAS)

NAME Alan M. DUFFIELD		TITLE Research Associate		BIRTH DATE (MM/DD/YYYY) December 16 1935	
PLACE OF BIRTH (City, State, Country) Perth, Western Australia		PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) Australian, Permanent resident Immigrant Visa		SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female	
EDUCATION (Begin with baccalaureate training and include postdoctoral)					
INSTITUTION AND LOCATION		DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD	
University of Western Australia		B. Sc (1st Class Hons)	1958	Organic Chemistry	
University of Western Australia		Ph.D.	1962	Organic Chemistry	
HONORS					
MAJOR RESEARCH INTEREST Applications of mass spectrometry to Biology and Biomedical Problems			ROLE IN PROPOSED PROJECT Organic Chemist/mass spectroscopist		
RESEARCH SUPPORT (See instructions) N/A					

RESEARCH AND/OR PROFESSIONAL EXPERIENCE (Starting with present position, list training and experience relevant to area of project. List all or most representative publications. Do not exceed 3 pages for each individual.)

- 1970 - Research Associate, Department of Genetics, Stanford University School of Medicine
- 1969 - Head of the Mass Spectrometry Laboratory, Chemistry Department Stanford University
- 1965 - 69 Research Associate, Department of Chemistry, Stanford University
- 1963 - 65 Postdoctoral Fellow, Department of Chemistry, Stanford University
- 1962 - 63 Postdoctoral Fellow, Department of Biochemistry, Stanford University School of Medicine.

PUBLICATIONS SINCE 1971

1. An Application of Artificial Intelligence to the Interpretation of Mass Spectra. Mass Spectrometry, B.W.G. Milne, Ed., John Wiley and Sons, New York, 1971, pp. 121-178
By B. G. Buchanan, A. M. Duffield and A. V. Robertson

2. Mass Spectrometry in Structural and Stereochemical Problems. CCIV. Spectra of Hydantoins.II. Electron Impact Induced Fragmentation of some Substituted Hydantoins.
Org. Mass Spectr., 5, 551 (1971)
By R. A. Corral, O. O. Orazi, A. M. Duffield and C. Djerassi
3. Electron Impact Induced Hydrogen Scrambling in Cyclohexanol and Isomeric Methylcyclohexanols.
Org. Mass Spectr., 5, 383 (1971)
By R. H. Shapiro, S. P. Levine and A. M. Duffield
4. Derivatives of 2-Biphenylcarboxylic Acid.
Rev. Roumain. Chem., 16, 1095 (1971)
By A. T. Balaban and A. M. Duffield
5. Alkaloide aus Evonymus europaea L.
Helv. Chim. Acta, 54, 2144 (1971)
By A. Klásek, T. Reichstein, A. M. Duffield and F. Santavý
6. Studies on Indian Medicinal Plants. XXVIII. Sesquiterpene Lactones of Enhydra Fluctuans Lour. Structures of Enhydrin, Fluctuanin and Fluctuadin.
Tetrahedron, 28, 2285 (1972).
By E. Ali, P. P. Ghosh Dastidar, S. C. Pakrashi, L. J. Durham and A. M. Duffield
7. The Electron Impact Promoted Fragmentation of Aurone Epoxides.
Org. Mass Spectr., 6, 199 (1972)
By B. A. Brady, W. I. O'Sullivan and A. M. Duffield
8. The Determination of Cyclohexylamine in Aqueous Solutions of Sodium Cyclamate by Electron Capture Gas Chromatography.
Anal. Letters, 4, 301 (1971)
By M. D. Solomon, W. E. Pereira and A. M. Duffield
9. Computer Recognition of Metastable Ions. Nineteenth Annual Conference on Mass Spectrometry, Atlanta, 1971, p. 63
By A. M. Duffield, W. E. Reynolds, D. A. Anderson, R. A. Stillman, Jr. and C. E. Carroll
10. Spectrometrie de Masse. VI. Fragmentation de Dimethyl-2,2-dioxolanes-1,3-Insatures.
Org. Mass Spectr., 5, 1409 (1971)
By J. Kossanyi, J. Chucho and A. M. Duffield
11. Chlorpromazine Metabolism in Sheep. II. In vitro Metabolism and Preparation of 3H-7-Hydroxychlorpromazine.
Journées D'Agressologie, 12, 333 (1971)
By L. G. Brooks, M. A. Holmes, I. S. Forrest, V. A. Bacon, A. M. Duffield and M. D. Solomon
12. 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.
Canadian J. Chem., 50, 2776 (1972)
By Y. M. Sheikh, R. J. Liedtke, A. M. Duffield and C. Djerassi

NAME Wilfred E. PEREIRA	TITLE Research Associate	BIRTHDATE (Mo., Day, Yr.) June 23 1936	
PLACE OF BIRTH (City, State, Country) Madras, S. India	PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) Indian, Permanent Resident Immigrant Visa	SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female	
EDUCATION (Begin with baccalaureate training and include postdoctoral)			
INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Madras Medical College, Madras, India	B. Pharm	1960	Pharmaceutical Chemistry
Saugar Univ, Madhya Pradesh, India	M. Pharm	1962	Pharm. Chem & Chem of Natur
U.C. Med. Center, San Francisco, Calif	Ph.D.	1968	Pharm. Chem & Pharmacology

HONORS

MAJOR RESEARCH INTEREST Identification of Metabolites & drug metabolites in Biological fluids	ROLE IN PROPOSED PROJECT Organic chemist
--	---

RESEARCH SUPPORT (See instructions)

RESEARCH AND OR PROFESSIONAL EXPERIENCE (Starting with present position, list training and experience relevant to area of project. List 3 or most representative publications. Do not exceed 3 pages for each individual.)

1968 - 1970 Post Doctoral Fellow, Dept. of Genetics Stanford University Med. School
 1970 - present Research Associate same institution
 During these four years I have been involved with peptide synthesis, amino acid analysis and synthetic organic chemistry. I helped develop methods for the separation of diastereoisomers by gas chromatography and have been involved with the routine use of gas chromatography mass spectrometry for the identification of urinary metabolites in normal and pathological urine and serum samples. My applications of mass spectrometry have included the development of mass fragmentography for the determination of the amino acid contents of soil and ~~plasma~~ serum. My present project involves the screening of urine from leukemic patients for abnormal metabolites and to investigate the metabolic fate of anti-leukemic chemotherapeutic agents in the body.

PUBLICATIONS

1. Transesterification with an Anion-exchange Resin;
W. Pereira, V. Close, W. Patton and B. Halpern,
J. Org. Chem. 34:2032 (1969).
2. Alcoholysis of the Merrifield-type Peptide-polymer Bond with an Anion Exchange Resin;
W. Pereira, V. A. Close, E. Jellum, W. Patton and B. Halpern,
Australian J. of Chem. 22:1337 (1969).

A. M. Duffield
Publications

13. Thermal Fragmentation of Quinoline and Isoquinoline N-Oxides in the Ion Source of a Mass Spectrometer.
Acta Chem. Scand., 26, 2423 (1972).
By A. M. Duffield and O. Buchardt
14. Applications of Artificial Intelligence for Chemical Inference. VII. An Approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids.
J. Amer. Chem. Soc., 94, 5962 (1972)
By D. H. Smith, B. G. Buchanan, R. S. Englemore, A. M. Duffield, A. Yeo, E. A. Feigenbaum, J. Lederberg and C. Djerassi
15. Mass Spectrometry in Structural and Stereochemical Problems. CCXIX. Identification of a Unidirectional Quadruple Hydrogen Transfer Process in 7-Phenyl-hept-3-en-2-one O-Methyl Oxime Ether.
Org. Mass Spectr., 6, 1271 (1972).
By R. J. Liedtke, Y. M. Sheikh, A. M. Duffield and C. Djerassi
16. An Automated Gas Chromatographic Analysis of Phenylalanine in Serum.
Clinical Biochem., 5, 166 (1972)
By E. Steed, W. Pereira, B. Halpern, M. D. Solomon and A. M. Duffield
17. Pyrrolizidine Alkaloids. XIX. Structure of the Alkaloid Erucifoline.
Coll. Czech. Chem. Commun., (1972)
By P. Sedmera, A. Klásek, A. M. Duffield and F. Santavý.
18. 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 Spectr., 7, (1973)
By D. H. Smith, A. M. Duffield and C. Djerassi
19. 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
20. A Study of the Electron Impact Fragmentation of Promazine Sulphoxide and Promazine using Specifically Deuterated Analogues.
Austral. J. Chem., 26, (1973).
By M. D. Solomon, R. Summons, W. Pereira and A. M. Duffield
21. Spectrometric de Masse. VIII. Elimination d'eau Induite par Impact Electronique dans le Tétrahydro-1,2,3,4-naphtalenediol-1,2.
Org. Mass. Spectrom., 7 (1973).
By P. Perros, J. P. Morizui, J. Kossanyi and A. M. Duffield
22. The Determination of Phenylalanine in Serum by Mass Fragmentography
Clinical Biochem., submitted for publication (1973).
By W. E. Pereira, V. A. Bacon, Y. Hoyano, R. Summons and A. M. Duffield

3. The Action of Nitrosyl Chloride on Phenylalanine Peptides;
W. Patton, E. Jellum, D. Nitecki, W. Pereira and B. Halpern,
Australian J. of Chem. 22:2709 (1969).
4. Abnormal Circular Dichroism of α -Amino Acid Esters;
J. Cymerman Craig and W. E. Pereira,
Tet. Let. 18:1563 (1970).
5. The Use of (+)-2,2,2-Trifluoro-1-Phenylethylhydrazine in the Optical
Analysis of Asymmetric Ketones by Gas Chromatography;
W. E. Pereira, M. Solomon and B. Halpern,
Australian J. of Chem. 24:1103 (1971).
6. The Microsomal Oxygenation of Ethyl Benzene. Isotopic, Stereochemical,
and Induction Studies;
R. E. McMahon, H. R. Sullivan, J. Cymerman Craig and W. E. Pereira,
Arch. Biochem. Biophys. 132:575 (1969).
7. The Steric Analysis of Aliphatic Amines with Two Asymmetric Centers
by Gas-liquid Chromatography of Diastereoisomeric Amides,
W. E. Pereira and B. Halpern,
Australian J. Chem. 25:667 (1972).
8. Optical Rotatory Dispersion and Absolute Configuration -XVII.
 α -Alkylphenylacetic Acids;
J. Cymerman Craig, W. E. Pereira, B. Halpern and J. W. Westley,
Tetrahedron 27:1173 (1971).
9. The Optical Rotary Dispersion and Circular Dichroism of α -Amino and
 α -Hydroxy Acids;
J. Cymerman Craig and W. E. Pereira
Tetrahedron 26:3457 (1970).
10. The Determination of Cyclohexylamine in Aqueous Solutions of Sodium
Cyclamate by Electron-capture Gas Chromatography;
M. D. Solomon, W. E. Pereira and A. M. Duffield,
Anal. Let. 4:301 (1971).

Publications continued-

11. Chlorination Studies. I. The Reaction of Aqueous Hypochlorous Acid with Cytosine; ^{ACCA}
W. Patton, V. Brown, A. M. Duffield, B. Halpern, Y. Hoyano, W. Pereira and J. Lederberg,
Biochem. Biophys. Res. Commun. 48:880 (1972).
12. The Use of R-(+)-1-Phenylethylisocyanate in the Optical Analysis of Asymmetric Secondary Alcohols by Gas Chromatography;
W. Pereira, V. A. Bacon, W. Patton, B. Halpern, and G. E. Pollock,
Anal. Let. 3:23 (1970).
13. A Rapid and Quantitative Gas Chromatographic Analysis for Phenylalanine in Serum;
B. Halpern, W. E. Pereira, M. D. Solomon and E. Steed,
Anal. Biochem. 39:156 (1971).
14. Electron-impact Promoted Fragmentation of Alkyl-N-(1-Phenylethyl)-Carbamates of Primary, Secondary and Tertiary Alcohols;
W. E. Pereira, B. Halpern, M. D. Solomon and A. M. Duffield,
Org. Mass Spectrometry 2:157 (1971).
15. Peptide Sequencing by Low Resolution Mass Spectrometry;
V. Bacon, E. Jellum, W. Patton, W. Pereira and B. Halpern,
Biochem. Biophys. Res. Commun. 37:878 (1969).
16. A Gas Liquid Chromatographic Method for the Determination of Phenylalanine in Serum;
E. Jellum, V. A. Close, W. Patton, W. Pereira and B. Halpern,
Anal. Biochem. 31:227 (1969).
17. Quantitative Determination of Biologically Important Thiols and Disulfides by Gas Liquid Chromatography;
E. Jellum, W. Patton, V. A. Bacon, W. E. Pereira and B. Halpern,
Anal. Biochem. 31:339 (1969).
18. A Study of the Electron Impact-promoted Fragmentation of Promazine Sulfoxide and Promazine Using Specifically Deuterated Analogues;
M. D. Solomon, R. Summons, W. Pereira and A. M. Duffield,
Australian J. Chem. (1973, in press).
19. The Determination of Phenylalanine in Serum by Mass Fragmentography;
W. Pereira, V. A. Bacon, Y. Hoyano, R. Summons and A. M. Duffield,
Clin. Biochem. (In press).
20. Chlorination Studies II. The Reaction of Aqueous Hypochlorous Acid with α -Amino Acids and Dipeptides;
W. E. Pereira, Y. Hoyano, R. Summons, V. A. Bacon and A. M. Duffield,
Biochem. et Biophys. Acta (In press).

BIOGRAPHICAL SKETCH

(Give the following information for all professional persons listed on page 3, beginning with the Principal Investigator. Use continuation pages and follow the same general format for each person.)

NAME Thomas C. Rindfleisch	TITLE Research Associate	BIRTHDATE (Mo., Day, Yr.) 12-10-41	
PLACE OF BIRTH (City, State, Country) Oshkosh, Wisconsin, USA	PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) USA	SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female	
EDUCATION (Begin with baccalaureate training and include postdoctoral)			
INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Purdue University, Lafayette, Ind. California Institute of Technology, Pasadena, CA	B.S M.S Ph.D	1962 1965 Thesis to be completed. All course work and examinations completed.	Physics Physics completed. All course work and examinations completed.
HONORS Purdue University, Graduated with Highest Honors, Sigma Xi.			
MAJOR RESEARCH INTEREST Space sciences, computer science and image processing		ROLE IN PROPOSED PROJECT Technical Support	
RESEARCH SUPPORT (See instructions)			

RESEARCH AND/OR PROFESSIONAL EXPERIENCE (Starting with present position, list training and experience relevant to area of project. List all or most representative publications. Do not exceed 3 pages for each individual.)

- 1971-Present** Stanford University Medical School, Department of Genetics, Stanford, CA.
Research Associate - Mass Spectrometry, Instrumentation research.
- 1962-1971** Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.
Relevant Experience:
1969-1971: Supervisor of Image Processing Development and Applications Group.
1968-1969: Mariner Mars 1969 Cognizant Engineer for Image Processing
1962-1968: Engineer - design and implement image processing computer software.
1. Rindfleisch, T. and Willingham, D., "A Figure of Merit Measuring Picture Resolution," JPL Technical Report 32-666, September 1, 1965.
 2. Rindfleisch, T. and Willingham, D., "A Figure of Merit Measuring Picture Resolution," Advances in Electronics and Electron Physics, Volume 22A, Photo-Electronic Image Devices, Academic Press, 1966.

Thomas C. Rindfleisch
PUBLICATIONS (cont'd)

3. Rindfleisch, T., "A Photometric Method for Deriving Lunar Topographic Information," JPL Technical Report 32-786, September 15, 1965.
4. Rindfleisch, T., "Photometric Method for Lunar Topography," Photogrammetric Engineering, March 1966.
5. Rindfleisch, T., "Generalizations and Limitations of Photoclinometry," JPL Space Science Summary Volume III, 1967.
6. Rindfleisch, T., "The Digital Removal of Noise from Imagery," JPL Space Science Summary 37-62 Volume III, 1970.
7. Rindfleisch, T., "Digital Image Processing for the Rectification of Television Camera Distortions," Astronomical Use of Television-Type Image Sensors, NASA Special Publication SP-256, 1971.
8. Rindfleisch, T., Dunne, J., Frieden, H., Stromberg, W., and Ruiz, R., "Digital Processing of the Mariner 6 and 7 Pictures," Journal of Geophysical Research, Volume 76, Number 2, January 1971.
9. Rindfleisch, T., "Digital Image Processing," To be published, IEEE Special Issue, July 1972.

BIOGRAPHICAL SKETCH

(Give the following information for all professional personnel listed on page 3, beginning with the Principal Investigator.
Use continuation pages and follow the same general format for each person.)

NAME Dennis H. Smith	TITLE Research Associate	BIRTHDATE (Mo., Day, Yr.) 11/12/42	
PLACE OF BIRTH (City, State, Country) New York	PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) USA	SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female	
EDUCATION (Begin with baccalaureate training and include postdoctoral)			
INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Massachusetts Inst. of Technology Cambridge, Mass.	S.B.	1964	Chemistry
University of California, Berkeley Berkeley, California	Ph.D.	1967	Chemistry
HONORS Alfred P. Sloan Foundation Scholarship NASA Predoctoral Traineeship Phi Lambda Upsilon, Sigma Xi			
MAJOR RESEARCH INTEREST Mass Spectrometry and A.I. in Chemistry	ROLE IN PROPOSED PROJECT Research Associate		

RESEARCH SUPPORT (See instructions)

N/A

RESEARCH AND/OR PROFESSIONAL EXPERIENCE (Starting with present position, list training and experience relevant to area of project. List all or most representative publications. Do not exceed 3 pages for each individual.)

1971-Present Research Associate, Stanford University, Stanford, Ca.
 1970-1971 Visiting Scientist, University of Bristol, Bristol, England
 1967-1970 Assistant Research Chemist, University of Calif. at Berkeley, Berkeley, Ca.
 1965-1967 NASA Pre-Doctoral Traineeship, University of Calif. at Berkeley, Berkeley, Ca.

Publications: See attached list.

Publications:

1. H. G. Langer, R. S. Gohlke, and D. H. Smith, "Mass Spectrometric Differential Thermal Analysis," Anal. Chem., 37, 433 (1965).
2. S. M. Kupchan, J. M. Cassady, J. E. Kelsey, H. K. Schnoes, D. H. Smith, and A. L. Burlingame, "Structural Elucidation and High Resolution Mass Spectrometry of Gaillardin, a New Cytotoxic Sesquiterpene Lactone," J. Amer. Chem. Soc. 88, 5292 (1966).
3. D. H. Smith, Ph.D. Thesis, "High Resolution Mass Spectrometry: Techniques and Applications to Molecular Structure Problems," Dept. of Chemistry, University of California, Berkeley, California (1967).
4. H. K. Schnoes, D. H. Smith, A. L. Burlingame, P. W. Jeffs, and W. Döpke, "Mass Spectra of Amaryllidaceae Alkaloids: The Lycorenine Series," Tetrahedron, 24, 2825 (1968).
5. A. L. Burlingame, D. H. Smith, and R. W. Olsen, "High Resolution Mass Spectrometry in Molecular Structure Studies, XIV. Real-time Data Acquisition, Processing and Display of High Resolution Mass Spectral Data," Anal. Chem., 40, 13 (1968).
6. A. L. Burlingame and D. H. Smith, "High Resolution Mass Spectrometry in Molecular Structure Studies II. Automated Heteroatomic Plotting as an Aid to the Presentation and Interpretation of High Resolution Mass Spectra Data," Tetrahedron, 24, 5749 (1968).
7. W. J. Richter, B. R. Simoneit, D. H. Smith, and A. L. Burlingame, "Detection and Identification of Oxocarboxylic and Dicarboxylic Acids in Complex Mixtures by Reductive Silylation and Computer-Aided Analysis of High Resolution Mass Spectral Data," Anal. Chem., 41, 1392 (1969).
8. The Lunar Sample Preliminary Examination Team, "Preliminary Examination of Lunar Samples from Apollo 11," Science, 165, 1211 (1969).
9. S. M. Kupchan, W. K. Anderson, P. Bollinger, R. W. Doskotch, R. M. Smith, J. A. Saenz Renauld, H. K. Schnoes, A. L. Burlingame, and D. H. Smith, "Tumor Inhibitors, XXXIX. Active Principles of Acnistus arborescens. Isolation and Structural and Spectral Studies of Withaferin A and Withacnistin," J. Org. Chem., 34, 3858 (1969).
10. A. L. Burlingame, D. H. Smith, T. O. Merren, and R. W. Olsen, "Real-time High Resolution Mass Spectrometry," in Computers in Analytical Chemistry (Vol. 4 in Progress in Analytical Chemistry series), C. H. Orr and J. Norris, Eds., Plenum Press, New York, 1970, pp. 17-38.
11. The Lunar Sample Preliminary Examination Team, "Preliminary Examination of Lunar Samples from Apollo 12," Science, 167, 1325 (1970).
12. D. H. Smith, R. W. Olsen, F. C. Walls, and A. L. Burlingame, "Real-Time Mass Spectrometry: LOGOS--A Generalized Mass Spectrometry Computer System for High and Low Resolution, GC/MS and Closed-Loop Applications," Anal. Chem., 43, 1796 (1971).
13. A. L. Burlingame, J. S. Hauser, B. R. Simoneit, D. H. Smith, K. Biemann, N. Mancuso, R. Murphy, D. A. Flory, and M. A. Reynolds, "Preliminary Organic Analysis of the Apollo 12 Cores," Proceedings of the Apollo 12 Lunar Science Conference, E. Levinson, Ed., M.I.T. Press, Cambridge, Mass. 1971, p. 1891.

DO NOT TYPE IN THIS SPACE-BINDING MARGIN

14. D. H. Smith, "A Compound Classifier Based on Computer Analysis of Low Resolution Mass Spectral Data," Anal. Chem., 44, 536 (1972).
15. D. H. Smith and G. Eglinton, "Compound Classification by Computer Treatment of Low Resolution Mass Spectra—Application to Geochemical and Environmental Problems," Nature, 235, 325 (1972).
16. D. H. Smith, N. A. B. Gray, C. T. Dillinger, B. J. Kimble, and G. Eglinton, "Complex Mixture Analysis - Geochemical and Environmental Applications of a Compound Classifier Based on Computer Analysis of Low Resolution Mass Spectra," "Advances in Organic Geochemistry 1971," M. R. v. Gaertner and M. Weher, Ed., Pergamon Press, Oxford, New York, Toronto, Sydney and Braunschweig, 1972, p.249.
17. D. H. Smith, B. G. Buchanan, R. S. Engelmores, A. M. Duffield, A. Yeo, E. A. Feigenbaum, J. Lederberg, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference, VIII. An Approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids," J. Amer. Chem. Soc., 94, 5962 (1972).
18. 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., in press.
19. B. R. Simoneit, D. H. Smith, G. Eglinton, and A. L. Burlingame, "Applications of Real-Time Mass Spectrometric Techniques to Environmental Organic Geochemistry, II. San Francisco Bay Area Waters," Arch. Env. Contam. and Tox., in press.
20. D. H. Smith, B. G. Buchanan, R. S. Engelmores, 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., submitted for publication.
21. D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, J. Lederberg, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference X, INTSUM. A Data Interpretation and Summary Program as Applied to the Collected Mass Spectra of Estrogenic Steroids," Tetrahedron, submitted.
22. D. H. Smith, "Mass Spectrometry," Chapter X in Guide to Modern Methods of Instrumental Analysis, T. H. Gow, Ed., Wiley-Interscience, New York, 1972.

DO NOT TYPE IN THIS SPACE-BINDING MARGIN

BIOGRAPHICAL SKETCH

(Give the following information for all professional personnel listed on page 3, beginning with the Principal Investigator.
Use continuation pages and follow the same general format for each person.)

NAME	TITLE	BIRTHDATE (Mo., Day, Yr.)	
Sridharan, Natesa S.	Research Associate	10-2-46	
PLACE OF BIRTH (City, State, Country)	PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date)	SEX	
Madras, India	India; pending permanent residence	<input checked="" type="checkbox"/> Male <input type="checkbox"/> Female	
EDUCATION (Begin with baccalaureate training and include postdoctoral)			
INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Indian Institute of Technology, Madras, India	Bachelor of Technology	1967	Electrical Engineering
State University of New York, Stony Brook	M.S.	1969	Computer Science
	Ph.D.	1971	Computer Science
HONORS			
University Fellow	1968-1971	SUNY Stony Brook	
Graduate Assistant	1967-1968	SUNY Stony Brook	
Siemens' Award (awarded for top rank in Electrical Engineering)	1967	ITT Madras	
National Merit Scholarship	1963-1967	ITT Madras	
MAJOR RESEARCH INTEREST	ROLE IN PROPOSED PROJECT		
Computer Applications in Chemistry and Medicine	Research Associate		
RESEARCH SUPPORT (See instructions)			

RESEARCH AND/OR PROFESSIONAL EXPERIENCE (Starting with present position, list training and experience relevant to area of project. List all or most representative publications. Do not exceed 3 pages for each individual.)

1971-present Research Associate, Heuristic Programming Project, Stanford University
1970-1971 Consultant, IAC Computer Company, Long Island, N.Y.

"Heuristic Theory Formation: Data Interpretation and Rule Formation". Machine Intelligence, Volume VII, 1972. (Co-Author).

"An Application of Artificial Intelligence to Organic Chemical Synthesis" Doctoral Dissertation, SUNY Stony Brook, August, 1971.