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 chromtography/mass spectrometery (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 artifical 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 data, 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 biologicaly 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 C6H6. The most familiar molecule with this molecular formula is benzene. Yet there are 217 topological isomers for C6H6 (with valence constraints) of which only 15 are pure trees. The simple addition of one oxygen atom to the empirical formula of benzene, yielding C6H6O, 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 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 volatges, 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 Universty 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 spctra 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 spctra. 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 preceeding 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 expirical 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. Engelmore, 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. Engelmore, 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 obectives 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 711 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 chromatograpms 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 breifly 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.