ANNUAL REPORT FOR RESEARCH PROJECT: "RESOURCE RELATED RESEARCH -- COMPUTERS AND CHEMISTRY"

Table of Contents

I. RESOURCE IDENTIFICATION

Project Title: Resource-Related Research Computers and Chemistry

Principal Investigator: Dr. Edward A. Feigenbaum
Telephone (415) 321-2300 Ext. 4878

Department: Computer Science

School: Humanities and Sciences

Site: Stanford University, Stanford, California

Total Project Period: May 1, 1971 through April 30, 1974

Business and Administrative Official: K. D. Creighton
Telephone (415) 321-2300 Ext. 2251

NATIONAL INSTITUTES OF HEALTH

DIVISION OF RESEARCH RESOURCES

BIOTECHNOLOGY RESOURCES BRANCH

SECTION I - RESOURCE IDENTIFICATION

Report Period:	Grant No. 5 R24 RR00612-02
From: 1/1/72 To: 12/31/72 mo/day/year mo/day/year	September 30, 1972 Date of Report Preparation

Name of resource	Resource Address	Resource Telephone No.
Resource-Related Research Computers and Chemistry	Computer Science Dept. Stanford University Stanford, Calif. 94305	(415) 321-2300 Ext. 4878
Principal Investigator	Title	Academic Dept.
Dr. Edward A. Feigenbaum	Professor	Computer Science
Grantee Institution	Type of Institution (Private Univ., State Univ., Hosp., etc.)	Investigator's Telephone No.
Stanford University Stanford, Calif. 94305	Private University	(415) 321-2300 Ext. 4878

Name of Institution's Biotechnology Resource Advisory Committee:

N/A

Membership of Biotechnology Resource Advisory Committee:
(Indicate Chairman and those who have reviewed this report)

Name	Title	Department	Institution

Typed Name & Title of Principal Investigator Dr. Edward A. Feigenbaum Professor	Signature Edward A. Feigenbaum	Date Oct. 24, 1972
Typed Name & Title of Grantee Institution Official Kathleen Butler Assistant Research Administrator	Signature	Date

II. RESOURCE OPERATIONS

A. Description of Progress

1. Overview

The Heuristic DENDRAL Project at Stanford University is an interdisciplinary research effort. The task area is a many-faceted problem of interest to medicine, chemistry, and computer science. Because the actual work has been divided into separate sub-problems along lines of scientific expertise, an overview is given here to establish the context of progress in each area, details of which are described in subsequent sections.

Following the organization of the original proposal, the progress and plans are organized into Parts A, B, and C, representing the different research efforts included within the scope of the proposal. Part A is aimed at enhancing the reasoning power of the existing Heuristic DENDRAL performance program so that eventually it may become a useful working tool for mass spectrum analysts. The goals of Part B include the closed-loop control of a mass spectrometer in realtime by a version of the Heuristic DENDRAL program; and the development of mass spectrum analysis techniques for certain classes of biologically important compounds. Part C concerns the development of the Meta-DENDRAL program, an attempt to achieve automatic theory formation in the area of mass spectrometry.

During this year we have made continued progress in each of the three major project areas. The following sections describe our progress and plans in detail. The highlights are:

Part A: (1) Analysis of high resolution mass spectra of estrogens and estrogen mixtures.
(2) Completion of the algorithm for generating cyclic structures.

Part B: (1) Development of hardware and software for routine data acquisition on the Varian-MAT 711 mass spectrometer, sending data to the IBM 360/50 computer at the Medical School's ACME facility.
(2) Preliminary work on analysis of the chemical components of urine. An initial application of this work for analyzing the urine of premature infants.

Part C: (1) Completion of the data interpretation program, the first part of automatic theory formation. Application of this program to new sets of data.
(2) Continued work on rule formation, the second main stage of theory formation.

The problem we have chosen to work on - the application of artificial intelligence to mass spectrometry - remains a richly varied problem domain. Its interest to medicine, analytic chemistry, and computer science have not diminished. We have discovered aspects of the problem which are more difficult than we initially thought. On the other hand, we have made more progress with other aspects in the last year than we would have predicted.

Interpretation of mass spectra requires the judicious application of a very large body of knowledge, whether it is done by a chemist or by a computer. Part of our work centers on acquiring new knowledge of mass spectrometry and codifying old knowledge. This means running and analyzing the mass spectra of unstudied classes of compounds as well as putting mass spectrometry rules into the computer program. These tasks have required the development of artificial intelligence techniques necessary to apply the chemical knowledge efficiently.

Part A. APPLICATIONS OF ARTIFICIAL INTELLIGENCE TO MASS SPECTROMETRY

Objectives:

The overall objective of this part of the research is extension of the Heuristic DENDRAL program to analysis of the mass spectra of complex organic molecules. This overall objective encompasses several sub-tasks, all of which represent critical steps in building a powerful program in an incremental fashion. Thus the current status of the program permits operation to continue in a routine, production mode wherein problem areas within the scope of the program can be investigated while extensions of the program are under development. The following specific objectives reflect both applications of the existing program and ongoing program development:

I) Assess the capabilities and limitations of the programming techniques for estrogenic steroids analyzed as unknown compounds and mixtures of compounds.

II) Generalize the programming techniques to ensure a high level of compound class independence.

III) Apply the techniques to other classes of steroids, alkaloids, and amino acids.

IV) Develop the cyclic structure generator for inclusion into the Heuristic DENDRAL program and explore the potential of the generator as an analytical aid of general utility.

V) Refine planning rules to infer compound classes or molecular substructures to minimize structures considered by the DENDRAL algorithm.

VI) Exploit ancillary information which can be obtained from other mass spectral techniques such as metastable ion spectra, low ionizing voltage spectra and mass spectral pattern shifts in isotopically or substituent labeled molecules.

VII) Design experimental strategies to collect, using the techniques of part VI, only those ancillary data required by DENDRAL to effect a solution or minimize ambiguities.

VIII) Structure the programs to utilize and/or request data from other spectroscopic techniques (e.g., proton magnetic resonance (PMR), carbon-13 magnetic resonance (CMR), infra-red (IR) or chemical techniques, such as isotopic labeling with deuterium).

IX) Explore the theoretical bases for mass spectral fragmentation processes to improve existing mass spectral theory.

X) Implement production analysis programs on the ACME computer facility to permit closer integration with the mass spectral data acquired and reduced on this facility.

Progress:

The following discussion of this task area of the proposal is keyed to the sub-task objectives described above:

I) The techniques of artificial intelligence have been applied successfully for the first time 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. The last topic is discussed in more detail in (II) and (III) following.

Because of the structure of the Heuristic DENDRAL program for estrogens, 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), (see proposal section Part B-2 below), many mixture components will be unresolved and an analysis program must be capable of dealing with these mixtures.

We have, in collaboration with Prof. H. Adlercreutz of the University of Helsinki, recently completed a series of analyses of various fractions of estrogens extracted from bodily fluids and supplied to us by Prof. Adlercreutz. 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 bodily 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 some complexity (<10 components), 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 will be submitted shortly. Because of the potential generality of this technique we will continue our investigations of estrogens and begin studies on mixtures of other steroids.

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 will tentatively be 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 are being used extensively in developing the program strategies for Meta-DENDRAL (see Part C, below).

II) The Heuristic DENDRAL program for complex molecules has received considerable attention during the last year in order to remove compound class specific information or program strategies. By removing information which is specific to estrogens, the program has become much more general. This effor+ 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) Specification 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 (see VI, below).

f) Available low ionizing voltage data -- to aid the search for molecular ions (see VI, below).

g) Results of deuterium exchange of labile hydrogens -- to specify the number of, e.g., -OH groups (see VI, below).

In the case of a known compound this procedure may be used to validate fragmentation rules developed on other, related compounds. This mode will be used extensively in testing the output of the data interpretation program (see Part C, below).

In the case of unknown compounds, rules with known generality for related, known structures may be used to determine the structure of the unknown. This mode has been used extensively for estrogens and will be extended to other classes (see III, below).

III) 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, below) 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 have also collected a total of 26 spectra of three classes of quinazolone and quinolone alkaloids for which mass spectra have not been previously recorded. As fragmentation mechanisms are developed for these classes, they will be tested against the known structures, and in the case of the quinazolone alkaloids tested against a set of nine compounds for which spectra have not been determined and which then can be treated as unknowns.

In connection with the goals of Part B-2 (see below) we will shortly commence a study of derivatized amino acids (Ntrilouoracetyl-O-lautyl esters). These are derivatives of choice for GC/MS analysis of amino acids whether derived from, e.g., bodily fluids or geological samples. This will be an important first step in integration of the data analysis programs with GC/HRMS data on urine extracts, as essentially no high resolution mass spectral studies have been carried out on constituents of urines.

IV. The cyclic structure generator now rests on a firm mathematical foundation such that we are confident of its thoroughness and ability to generate structures with PROSPECTIVE elimination of 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 algorithm.

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. Mr. Larry Masinter, Dr. N. S. Sridharan, and Mr. Larry Hjelmeland have been key personnel in bringing the algorithm to completion and implementing it.

A 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, group theory, and the symmetry properties of cyclic graphs the labelling algorithm avoids construction of redundant isomers by identification of equivalent node positions on the graph structure before labelling takes place. It is indicative of the complexity of this problem and the importance of its solution to both chemists and mathematicians that it has remained unsolved (until now) despite attention for over 100 years. A manuscript describing the underlying mathematical theory has been submitted to the DISCRETE MATHEMATICS.

The cyclic structure generator in its entirety (encompassing acyclic, wholly cyclic and combinations thereof) will be described separately. 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. Manuscripts describing the mathematical and chemical aspects of the structure generator are in preparation.

Over the summer we were fortunate to have the help of Prof. Harold Brown, a visitor to Stanford from the Dept. of Mathematics at Ohio State University. He brought to the problem a depth of mathematical analysis which was important for finishing the design of the algorithm and working out details of its implementation. He was largely responsible for the manuscripts describing the graph theory of the labeling algorithm and the graph theory of the structure generation algorithm.

The cyclic structure generator makes it possible to define the boundaries, scope and limitations of organic chemistry as a whole, rather than simply the acyclic part of it. 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 more than 200 topological isomers for C6H6 (with valence constraints) of which only 15 are totally acyclic.

The first use 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, so that the constraints mentioned previously are simple to exercise in the context of the dictionary.

We feel that the cyclic structure generator has the potential of acting as the focal point for an interactive laboratory analytical tool. 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. This concept will certainly be developed further as we improve both our capabilities for inference from scientific data and our techniques for using the generator.

V) 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 are presently under investigation:

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. Mr. Larry Hjelmeland and Mr. Mark Stefik have been investigating library matching procedures which fit our needs.

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.

c) For complex molecules which may contain several functional groups we have explored and are continuing exploration of incorporation of molecular substructures into the planning scheme. Thus rather than infer a class or particular skeleton, inferences are made about specific functional groups (e.g., -NH2, OH) or substructures (e.g., -CH2-CH2-CH3). This is the form in which information from other spectroscopic techniques is available, and we plan to extend our present capabilities for planning based on this information (see VIII, below).

VI) There are several additional techniques available to the mass spectroscopist other than recording the conventional mass spectrum. These techniques are used routinely in everyday research as they provide considerable complementary data which frequently are of great assistance in rationalization of the conventional spectrum, either in terms of structure or fragmentation mechanisms. We have modeled the Heuristic DENDRAL program for complex molecules to use data from these additional techniques in much the same way as a chemist does. We have the capability of determining the following three types of data on our mass spectrometers and using them in the program.

a) Metastable Ion (MI) Data. Metastable ions provide a means for relating fragment ions to molecular ions in a mass spectrum. This information is extremely 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 establish that a set of fragmentation processes which are to be used as rules to guide the Heuristic DENDRAL program are in fact viable processes and occur in a known manner. An example of the utility of these observations has been investigations of metastable ion data in the mass spectra of estrogens (Smith, Duffield and Djerassi, 1972).

The second context is, in the case of analysis of mixtures of compounds, a determination of which fragment ions in a very complex spectrum are related to which molecular ions. We have explored the analysis time and specificity of results as a function of the amount of metastable ion data available on a mixture and noted one to two orders of magnitude reduction in computer time 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 will be reported in detail in the description on analysis of the estrogen mixtures (see I, above).

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 of a mass spectrometer. Earlier work was directed at ions formed in a fieldfree region just prior to entering a magnetic field (mass analysis). This is the only method available for metastable ion studies for a single focussing mass spectrometer: The metastable ions formed in this region appear as diffuse peaks superimposed on the normal mass spectrum. The mass positions of these metastable ions, however, satisfy (mathematically) several relationships of pairs of normal ions. This lack of specificity and frequent difficulties in accurately determining the mass positions has caused us to turn our attention to studies of so-called "defocussed" metastable ions. A conventional double focussing mass spectrometer possesses two field-free regions where metastable ions may be studied. One field-free region lies between the electric sector and the magnetic sector. This region can be used to study metastable ions of the type discussed above. 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 the instrument (defocussing) so that normal ions are not observed, but metastable ions are. This procedure allows establishment of specific relationships between ions involved in a metastable decomposition so that the original ion which decomposes during flight, and its decomposition product, can be identified. This technique has let to much more useful information for the Heuristic DENDRAL program, as illustrated earlier in this section.

b) 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 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.

c) 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. Isotopically labeled molecules have permitted a detailel examination of fragmentation processes of complex molecules utilizing comparisons of metastable ion spectra of labeled and unlabeled molecules (Smith, Duffield and Djerassi, 1972).

Future work will involve suggestions by a program of likely sites of hydrogen transfer in the course of fragmentation. Elucidation of fragmentation processes is a part of the Meta-DENDRAL effort (Part C, below). More detailed specification of these processes can be effected by isotopic or substituent labeling of molecules and we feel that a program is capable of suggesting the necessary experiments.

In addition, we are exploring the feasibility of using C13 NMR data to complement mass spectrometry data. Its initial use will be to determine the branching structure of alkyl chains away from the heteroatom in aliphatic monofunctional compounds. Dr. Ray Carhart, an NIH post-doctoral fellow, is working on this problem together with Ms. Hanne Eggert, a visiting scholar from the University of Copenhagen, Denmark. Substantial work on the C13 NMR theory of amines has been described in a manuscript (by Eggert & Djerassi) to be submitted soon.

VII) Designs of experimental strategies represent a crucial link between the Heuristic DENDRAL program and the instrument control aspects of this proposal (see Part B-1, below). We have begun planning ways in which the program, cognizant of intermediate results, can suggest additional collection of data that will be required for an unambiguous determination of structure, or at least to minimize ambiguities. These suggestions can ultimately be translated into control parameters sent back to the mass spectrometer. In any real-time data collection scheme involving small amounts of sample, time is of the essence. It is crucial to select those data which are necessary and sufficient and to avoid collection of redundant or spurious data. We feel an "intelligent" program can supervise the data collection and analysis to fulfill this goal and can accomplish the task in real-time.

VIII) The Heuristic DENDRAL program for SAM molecules is already structured to accept additional spectroscopic data in the forms of GOODLIST and BADLIST specifying molecular substructures which are present or absent. We have deferred implementation of this more general approach to the Heuristic DENDRAL program for complex molecules until the cyclic structure generator is ready. Up until now, any such data from other techniques have been used retrospectively to check candidate structures for the requisite functional groups or substructures. Now that the structure generator is available, we will begin implementation of the GOODLIST and BADLIST for cyclic molecules.

IX) 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 generalizations. Quantum mechanics may be capable of providing this information. With Dr. Gilda Loew, we have been investigating extended Huckel molecular orbital theory in an attempt to predict some qualitative indications of the propensity of bonds to fragment. Our initial efforts have been aimed at the estrogenic steroid estrone, and a manuscript will shortly be submitted describing these results. 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. We are attempting to extend these results to other molecules, specifically, amino acids.

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.

X) A version of Stanford 360/LISP has been mounted on the Medical School's ACME computer system. This version, available to us in the overnight batch processing operation, has proven useful for running production versions of programs. Because our mass spectral data are acquired and reduced via ACME, this facility has removed the need for transferring data from ACME to the campus facility. We regret to report, however, that this version of LISP is not available to us in the time sharing mode during the day when mass spectral data are collected. Thus, although routine data analysis is facilitated, there is no immediate prospect for integration of DENDRAL into the real-time aspects of the problem. For the near future these activities will be simulated through batch processing to enable us to develop the necessary techniques for real-time interaction.

Plans:

In most cases, the plans for future work are embodied in and dictated by the progress we have made so far. Many of the plans, therefore, are outlined in the Progress section, above. As a brief summary then we plan the following activities, again keyed to the sub-task objectives:

I) We plan to continue with analyis of additional estrogen mixtures from bodily fluids in view of the excellent performance of the program so far.

II) We feel we have achieved a high level of 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) Extend Heuristic DENDRAL for complex molecules to the classes for which spectral data are or shortly will be available, pregnanes, cholestores, the above alkaloids and amino acid derivatives.

IV) Constraints will be developed for the cyclic generator that are easily understood by chemists and easily implemented in the computer program.

V) Planning rules for compound class determination will receive considerable attention as Heuristic DENDRAL is extended.

VI, VII) We understand how to use this additional information. Work needs to be done on algorithms to determine which experiments to do and how best to do them to minimize consumption of valuable samples.

VIII) As the structure generator is developed, we plan to implement it in Heuristic DENDRAL so that constraints imposed by spectroscopic data may be used effectively.

IX) We plan to analyze amino acids using molecular orbital theory to extend the theoretical basis for prediction of mass spectra.

X) We plan to simulate in as much detail as possible the interaction between Heuristic DENDRAL and the mass spectrometer to direct data collection in an intelligent fashion.

Part B-i. EXTENSIONS OF THE COMPUTER-MASS SPECTROMETER SYSTEM.

Objectives:

Data acquisition in real-time from the Varian-MAT 711 mass spectrometer with analysis of these data by Heuristic DENDRAL is the primary objective of this section of the research. We ultimately seek a substantial degree of control by computer program over the acquisition of data from the mass spectrometer. With sufficient computer power it is possible to accomplish the control within the time scale of GC/MS operation. A rationale of this approach and our efforts toward devising suitable programs to achieve this goal are described above under Part A.

The following operational parameters of the mass spectrometer are desirable and amenable to control: magnetic scan speed and mass range of scan, slit widths (to adjust to high or low resolution operation, ion optical stops (to increase resolution in the metastable defocussed mode), accelerating or electrostatic sector voltages, ionizing voltage (to switch from normal to low ionizing voltage), and rate and temperature of probe heating when the direct insertion probe is used to introduce samples into the mass spectrometer. Control of GC conditions is also possible.

Progress:

The Varian-MAT 711 mass spectrometer was formally accepted by Stanford University on Nov. 5, 1971. Prior to this time the instrument installation and performance tests went extremely smoothly. Shortly after acceptance, however, a series of electronic and mechanical malfunctions occurred which necessitated a visit of an engineer from Germany for a period of several weeks. Since that time the instrument has been used routinely in all its operating modes including ultra-high resolution peak matching, scanning at high resolution for accurate mass measurement; GC/MS operation, low ionizing voltages, and metastable defocussing. This instrument has now assumed the entire burden of data acquisition for DENDRAL related activities.

There are two activities related to the goals of this Part area which have proceeded in parallel with gaining familiarity with the new instrument. These activities are improving the software (programming) for data acquisition and reduction, and developing new hardware the initial efforts toward instrument control.

Software.

Great advances have been made in the programming for data acquisition and reduction, particularly since the arrival of Mr. Tom Rindfleisch, who helps direct the Instrumentation Research Laboratory's efforts in the DENDRAL mass spectrometry area. The following items indicate these advances.

a) 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. Proof of this theory with actual data would greatly relieve the burden of high data rates on the computer system, particularly as imposed by GC/MS operation, and permit considerably more data reduction to be accomplished in real-time. Statistical and observed models of peak profiles have suggested certain design changes in the hardware (see below).

b) Instrument Evaluation. A high resolution mass spectrometer operating in a dynamic scanning mode is a complex beast. There are many things that can go wrong which yield effects which may be invisible to the operator. Furthermore mode changes during closed loop operation require instrument adjustments which must be computer controlled. It is, therefore, necessary that the computer 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. To ensure that the instrument is operating properly and high quality data are being gathered, we have devoted some time to development of a program which monitors the state of the mass spectrometer. This preliminary program checks the following items:

i) Data acquisition parameters, i.e., the threshold, specifically determined peak width and intensity criteria, the number of peaks and the data storage utilized.

ii) Calibration of the mass/time scale, storage of same to be used as a model for subsequent spectra, output of mass range over which 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.

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

All output and warnings to the operator are provided on a CRT adjacent to the mass spectrometer immediately after a scan. Although this program works for the present time only with the calibration compound, PFK (no additional sample), it provides a basis for a general mechanism to monitor data quality to prevent wasting valuable samples when the instrument malfunctions.

The program contains many interactive features which permit the operator to examine selected features of the data at his leisure. He may display any selected peak profiles, obtain listings of calculated masses, plot a spectrum from the data and so forth.

In the longer term as more quantitative experience is gained with operating the MAT 711 in various modes and as instrument control hardware is completed, models relating instrument parameters to control functions and interactions will be developed. These will allow strategies to be planned for automated mode switching and performance optimization needed for intelligent control of data collection and reduction processes.

c) Data Resolution. A program has been written which allows automatic reduction of high resolution data based on the results of the prior instrument evaluation spectrum. This program uses parameters supplied by the operator prior to running the sample. Calibration of the mass/time curve is effected by mapping each spectrum into the calibration model developed previously. Separation of reference compound peaks (PFK) from unknown sample peaks is accomplished 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.

Mass calculation is accomplished with an algorithm based on a detailed evaluation of the behavior of the mass/time curve as a function of mass. Determination of elemental compositions proceeds utilizing a new, rapid and efficient algorithm developed by Prof. Lederberg. This program has made a previously onerous task (much human intervention) into an automatic one. This is an important step toward fully automatic data acquisition and reduction.

Hardware.

The gas chromatograph has been successfully interfaced to the mass spectrometer. An oscilloscope has also been incorporated with the spectrometer to supplement the strip chart recorder, to simplify initial adjustment of the instrument and to monitor every spectrum.

New interfaces for mass spectrometer operation and control have been developed. They have been designed around the PDP-11 computer as this computer represents our means of real-time interaction with the mass spectrometer. The interfaces can handle (through an analog multiplexer) several analog inputs and outputs which require that the computer be relatively near the mass spectrometer. This move has recently been accomplished, as the computer used to reside in a separate building. 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 important 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 should lessen ion statistical uncertainties in intensity values and thus ultimately improve peak position determinations in time and mass.

Plans.

As in Part A, many of the plans are mentioned in the above Progress sections. Again, a brief summary would include the following:

I) Continue improvement of the high resolution data acquisition and reduction programs. Pay particular attention to increased speed and tasks which may be carried out in real-time in the small computer, leaving ACME for those tasks requiring large compute power.

II) Develop a data acquisition and reduction system to be used in initial studies of the GC/MS system. Initially this system will operate at low resolution to avoid sensitivity problems in the time constraints imposed by GC operation. The real goal is high resolution operation of the system as we solve sensitivity problems. Some programming and experiments have already been done in this area.

III) Explore the GC/MS system and its interface for optimum conditions for the urine samples and related mixtures extracted from other bodily fluids (see Part B-ii, below).

IV) Develop additional hardware to exercise specific control functions as necessary for on-line mode changes and instrument performance optimization.

V) Develop better analytical models for the behavior of the mass spectrometer to yield more accurate data (masses and intensities).

VI) Finish study of ion signal treatment and related digitization rate requirements.

VII) Develop software communication between DENDRAL, ACMF and the PDP-11 so that ACME generated (via DENDRAL) requests can be serviced at the mass spectrometer and resulting data returned promptly.

Part B-ii. CHEMICAL CONSTITUENTS OF URINE.

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

```
             URINE (pH = 1, internal standards added)
               |
               |
               |
               | ether extraction
               |
  --------------------------
  |                        | ether phase             aqueous phase
(free acids)    --------------------------------
     A          |                   |          |
          (carbohydrates)     (amino acids)    |
                C                   B          |
                                               | hydrolysis
                                               |
                            --------------------------------
                            |                              |
                       ether phase                      aqueous
                   (hydrolysed acids)               (amino acids)
                            D                              F
```

The experimental procedure used for working with a urine sample is as follows. To an aliquot (25 ml.) of a 24 hour urine sample is added 6N hydrochloric acid until the pH is 1. Two internal standards, n-eicosane and 2-amino octanoic acid are then added. Ether extraction isolates the free acids (fraction A) which are then methylated and analysed by gas chromatography-mass spectrometry. An aliquot of the aqueous phase (2 ml) is concentrated to dryness, reacted with n-butanol/hydrochloric acid followed by methylene chloride containing trifluoroacetic anhydride. This procedure derivatizes any amino acids (or water soluble amines) which are then subjected to GC/MS analysis (fraction B).

If desired another 2 ml aliquot of the aqueous phase can be derivatized for the detection of carbohydrates (Fraction C). Our experience has been that this fraction generally contains few components and it can be eliminated without detriment to the overall urine analysis.

Concentrated hydrochloric acid (1.25 ml) is added to the urine (12.5 ml) after ether extraction and the mixture hydrolysed for 4 hours under reflux. Ether extraction affords the hydrolysed acid fraction (D) which is then methylated and analysed by GC/MS. A portion of the aqueous phase (2 ml) from hydrolysis of the urine is concentrated to dryness and derivatized and analysed for amino acids (Fraction E) as described under step B.

Urinary output from any individual will vary to some extent with diet. In order to suppress the problem of dietary variation it was decided to monitor the urine of premature infants in the Stanford Nursery of the Pediatrics Department. These infants are sustained on a carefully regulated diet and their hospital confinement is usually of the order of one month such that their urinary excretion could be investigated as a function of time.

Preliminary studies on approximately 20 urine samples from premature infants provided the experience necessary for a selection of the best operational techniques for chromatographic separation. This work has been carried out in the Department of Genetics where a suitable gas chromatograph and mass spectometer were available. The mass spectrometer (Finnigan Quadrupole, model 1015) used to date in this investigation is interfaced for data acquisition to the ACME computer system. During the gas chromatography-mass spectrometric analysis of a urine fraction over six hundred mass spectra are recorded in 45 minutes. A data system is mandatory to handle this avalanche of data and until one is functioning on the Varian-MAT 711 mass spectrometer we anticipate using the quadrupole instrument for the routine analysis of urine.

In the preliminary study of 20 urine samples from premature babies the only abnormal metabolite observed was p-hydroxyphenyl lactic acid which occurred in three of the samples. This compound's presence reflects the known ability of some premature infants to metabolize p-hydroxyphenyl pyruvic acid to the corresponding lactic acid. In all cases we observed the excretion of p-hydoxyphenyl lactic acid to drop to normal levels after several days presumably as particular enzyme functions became operative in the child.

Following these preliminary studies a joint program was formalized between the Departments of Genetics and Pediatrics to investigate late metabolic acidosis of the premature. A copy of the protocol to be used in this investigation is attached to this report.

At this time several urine samples from premature infants have been investigated but only one child was acidotic when the urine sample was collected. This urine sample was definitey abnormal and it appears to contain large quantities of p-hydroxy mandelic acid and p-hydroxyphenyl lactic acid. These abnormal metabolites were present in each of three daily samples of urine submitted to GC/MS analysis. It is interesting that the occurrence of p-hydroxyphenyl lactic and p-hydroxy mandelic acids in urine has been associated with abnormally high tyrosine levels while in our case tyrosine is present in normal concentrations.

The investigation of acidotic premature infants, although just commencing, shows promise that any organic acids causing acidosis will be identified by our analytical techniques.

In addition to these clinical aspects described above, work is continuing on the computer analysis of the mass spectra generated from urine specimens. Work has progressed on the construction of library lookup routines operating on data tapes obtained from Dr. Egil Jellum, Oslo, Norway, a former collaborator in our laboratory.

Part C. EXTENDING THE THEORY OF MASS SPECTROMETRY BY A COMPUTER

Objectives:

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 hope to provide computer programs which are of some practical help to the theory-forming scientist.

Mass spectrometry is the task domain for the theory formation program, called Meta-DENDRAL, 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 formalized to any great degree, and there remain gaps in the theory, but discovering new explanatory rules and systematizing them is taking place throughout the country, albeit slowly.

We have described the design and partial implementation of the Meta-DENDRAL program in a paper presented at the 7th Machine Intelligence Workshop (Edinburgh, Scotland, June, 1972). A copy of that paper is attached and should be consulted for details. It will be published in the proceedings of the conference (Machine Intelligence 7, B. Meltzer & D. Michie, eds., in press).

Our objective is to explore the theory formation problem for mass spectrometry within the context of AI research. As mentioned earlier we hope to produce intermediate programs which will aid chemists in formulating new pieces of theory as well. The following subgoals have guided our research along one dimension, although we have often been forced to consider other dimensions of the problem. The discussions of progress and future work are structured around these subgoals.

(1) Collect a suitable set of known mass spectra together with representations of the molecular structures from which the spectra were derived.

(2) Summarize and interpret the data with respect to possible explanations of the individual data points. This re-representation of the data is a critical step in extracting explanatory rules, for the data points are, for the first time, associated with possible mechanistic origins ("causes").

(3) Peruse the summary to make plans for intelligent rule formation. Any of the possible mechanisms described in the summary-interpretation phase could be incorporated in a rule of mass spectrometry. But planning will allow the rule formation program to start with explanatory rules which are likely to make good reference points for the whole rule formation process.

(4) Incorporate the possible mechanisms into general rules (rule formation). By bringing more and more of the descriptive mechanisms under rules, the rule formation program explains more and more of the original data points. This is difficult for many reasons, however. For instance, the rules must be general enough to avoid writing a new rule

for each data point. Yet there are numerous ways of generalizing rules, with few prospective guidelines to focus attention on the elegant generalizations which explain many data points simply. Various alternatives for rule formation, which we are exploring, are described in the progress section.

(5) Evaluate the rules to decide retrospectively whether each proposed rule is worth keeping or not. If so, it may be further modified in light of more data. If not, it will be discarded in favor of rules which are simpler, explain more data, or are otherwise better suited for incorporation into the emerging theory.

(6) Codify the rules into a theory. Although a set of phenomenological rules can predict the mass spectral behavior cf the class of molecules, further codification is needed to increase the explanatory power of the rules. This may mean something as "simple" as collapsing rules or subsuming rules under one another. Or, at a deeper level, it may mean finding relationships and principles which explain why the phenomenological rules are good predictors.

(7) Finally, it will be necessary to compare alternative theories (at whatever level) that come out of the program in order to choose the best one. Part of this research means experimenting with different criteria of "best" theory. Although the philosophical literature is full of suggested criteria, no one has ever tried to make them precise enough for use in a program.

Progress:

Meta-DENDRAL has progressed in the last year within several of the problem areas mentioned above. The attached paper (MI 7) describes much of our progress in mapping out a detailed strategy for attacking the problem. In addition, we have explored many issues related to alternative design or implication strategies. The unedited notes of our frequent group meetings are attached to show the issues discussed and some of the direction of our experimentation.

(1) Collection of mass spectrometry data was no problem because of the files kept for the Heuristic DENDRAL program and the availability of the mass spectrometer. Deciding which set of data to explore, however, was more difficult.

We had initially hoped to do theory formation for a large heterogeneous class of molecules in order to test the ability of the program to separate classes of molecules with dissimilar mass spectrometric behavior and group the similar classes of molecules. We had initially started working with the collection of saturated aliphatic monofunctional compounds and their mass spectra, already collected for previous Heuristic DENDRAL work. Later it was decided that we could make a more direct assault on the theory formation problem by choosing a set of homogeneous compounds whose mass spectrometry was already well characterized. It was hoped that we could formulate rules which corresponded closely with the known characterizations after examining only a small number of compounds and their spectra (tens of compounds, not thousands). The class of molecules chosen was the class of estrogenic steroids. This was an especially good choice because (a) the estrogens have been studied extensively - and thus there are known rules with which to compare the program's "discovered" rules - and (b) the estrogens, partly because of their biological interest, are not well enough characterized - thus the intermediate results of the program's analysis of estrogen mass spectra are interesting and immediately useful to science.

(2) 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 mechanisms of interest. 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 a manuscript to be submitted. 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. 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 alkaloids.

As always, we have modified the program many times after it produced its initial results in order to add new items of information to the summary or to reformat the summary - both aimed at making the program a more useful tool for chemists instead of just a computer science research tool. In a sense this is a diversion. But we feel it is important in interdisciplinary research to satisfy many goals (within the project) to maintain the high motivation and cooperative spirit which have characterized this project from the start.

(3) 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 just waiting to be found.

In a planning program currently being 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 disambiguated).

The program will be made much more sophisticated as we gain more experience with it. Even the output of this crude program, however, is useful to humans 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 disambiguation of other sets of mechanisms and also serve as a "core" of rules to be extended and modified in the course of detailed rule formation.

(4) The process of rule formation is the most difficult to define precisely. We have explored various strategies which are described briefly below and discussed in the attached notes of meetings. Although

we have in hand programs which formulate rules from the summary data, we are not completely satisfied with any of them. Thus, much work remains to be done on rule formation.

The following outline, written by Dr. Sridharan and taken from our internal working notes, encapsulates the dimensions of the rule formation problem we have considered and some of our explorations within those dimensions. Not all of the items presented there have been explored by writing computer programs, although we intend to do much of this in the future. Part I of this encapsulation presents two ways of characterizing theories. The formal representation mentioned in I-A was developed in the Machine Intelligence 7 paper. The less formal characterization of I-B is the subject of much of the philosophy of science literature which we are researching.

Rule Formation Work in Meta-DENDRAL

- I. Theory Representation and Formalization of Theory Formation Task
  - A. Formal Representation
    - i) Kinds of theory classes
Action based, Partial, 0-1 theories
    - ii) Set theoretic framework and theory definition using Generalized Cover Theory
    - iii) Definition of spaces: of theories, of rules, of situations, of actions
  - B. Characterization of Theories
    - i) How much prior chemistry assumed.
    - ii) How much ms theory assumed/Consistency
    - iii) Internal consistency
    - iv) Simplicity/complexity
    - v) Testability/falsifiability
    - vi) Performance with respect to data, predictive performance
    - vii) Predictive scope, Generality
    - viii) Explanatory power
    - ix) Projectability
    - x) Degree of instantiation
    - xi) Ambiguity
    - xii) Efficiency
- II. Exploration of Methodology and Paradigms
  - A. Model Building
    - i) Statistical analyses
    - ii) Discrete, charge localized model
    - iii) Fluid flow class of models
    - iv) Quantum Mechanical model
  - B. Deriving S-A Rules
    - i) Derive S-A rules from model and data
    - ii) Derive S-A rules from summarization of data
      - a) Constructive method
Generalization, Specialization, Validation, Evaluation and Codification
      - b) Generative method
Generation, Validation and Heuristic guidance
- III. Confrontation with the Realities of Data
  - A. Large volumes of data
  - B. Richness or high information density in data
  - C. Ambiguity
  - D. Limitation to the significance of data

a) Recording resolutions
b) Reproducibility limits

E. Need to watch for errors and mistakes in data, besides the need to manage data in the presence of such errors

Part II of the outline of Meta-DENDRAL work points to numerous places in the discussion notes concerning questions of the level of theory to be built and the program strategies to be used. We have concentrated on level II-A-ii - a more or less descriptive model of mass spectrometry written in terms of discrete atoms, bonds, and electronic charge. The programs already written, with one exception, use this model. The exception is the statistical programming work by Professor Ed Blaisdell, a visitor to Stanford last summer from the chemistry department of Juniata College (Huntingdon, Pennsylvania). The programs he developed attempted to derive a regression model from statistical analysis of the data in order to predict the strength of processes as a function of properties of the molecule. Items iii and iv of II-A are models of mass spectrometry which computer programs could conceivably work in. But our discussions, as yet, have not led to actual programs which will allow us to try out our ideas with some precision.

The strategies mentioned in Part II-B all fit within Artificial Intelligence paradigms, but so far we have little guidance on how to choose a good strategy. Part II-B-i refers to a Gelernter-like strategy of problem solving in which, in our case, a rough model of mass spectrometry in the program serves as a reference for checking the plausibility of proposed additions to the theory being built, say by statistical analysis. The so-called constructive model (II-B-ii-a) of the rule formation process is the one the programs have been working with mostly. It is the one described at the beginning of this section as the method we are following. While this is true, we do not wish to exclude the other methods from consideration until some detailed experiments have been performed. The generative method (II-B-ii-b) is the closest to the well-known heuristic search paradigm of Artificial Intelligence programs. Mr. Carl Farrell is pursuing this approach in his Ph.D. dissertation (directed by E. A. Feigenbaum and B. G. Buchanan). Outlines of his dissertation and computational procedure are attached to this report for reference.

The last section of the outline (III) covers a large part of the discussions in our meetings this year. Because we are working with real, and not ideal, experimental data, our rule formation problem is much more complex than, say, grammatical inference problems as currentlly formulated. Working in an idealized task domain could remove these difficulties, but we feel we would thereby lose much of the fascinating complexity of this problem.

(5-7) Many discussions have taken place on the topics of rule evaluation, codification of rules into theories, and theory evaluation. However, we have considered it premature at this point to begin writing computer programs for thse tasks until the rule formation problem itself was on firmer ground.

Plans:

Our plans for the coming year are to focus on specific gaps and problems in the design and implementation of the theory formation research now in progress. In particular, we will continue working with the mass spectra of estrogens, concentrating especially on the rule formation subtask described above.

We expect the programs to contribute to the formulation of new theory by humans for specific classes of molecules. At the same time, we expect to capture in the program more of the judgmental elements of rule formation.

B. SUMMARY OF RESOURCE USAGE

This section called for in the annual report instructions does not apply since the project is resource-related research and has no resources, per se.

II.C. RESOURCE EQUIPMENT LIST

EQUIPMENT LOCATED IN MAIN RESOURCE AREA

Equipment						Cost		
Description/ Identification	Manufacturer	Type	Model No.	Date Installed	Date Accepted	Purchase Price	Annual Rent	Source of Funds

In addition to equipment listed in the prior year report, the following items costing over $1,000. were purchased with funds from this grant:

Oscilloscope	Tektronix	Storage	5103	2-2-72	2-2-72	1,567.97

The following parts for the Varian MAT 711 mass spectrometer were capitalized as system additions:

Sample Rod and Inlet Filament	$1,717.80
Spare Source	1,827.50
Exit Slit	1,020.00

A large number of component parts were purchased (some capitalized) for the construction of interfacing hardware used in data-to-computer information processing. Major items were:

Digital Equipment Co.	Unibus Repeater	$1,953.02
Digital Equipment Co.	Digital Interface	1,250.00

D. SUMMARY OF PUBLICATIONS

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", Journal of the American Chemical Society, 94, 5962-5971 (1972).

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).

H. Brown, L. Masinter, L. Hjelemeland, "Constructive Graph Labeling Using Double Cosets". Stanford Computer Science Memo 318 (1972).

Grant No. 5 R25 [illegible]12-02
Section III-A

A. Summary of Resource Expenditures
By budget Periods

		5/1/71- 12/31/71 Actual Previous	1/1/72- 12/31/72 Current	1/1/73 12/31/73 Estimated Next
1.	Personnel			
	a. Salaries & Wages	$53,571.	101,851.	111,885.
	b. Fringe Benefits	7,878.	16,014.	18,195.
	Subtotal	61,449.	117,865.	130,080.
2.	Consultant Services	-0-	-0-	-0-
3.	Equipment			
	a. Purchased	99,483.	14,395.	-0-
	b. Maintenance	-0-	-0-	8,000.
	Subtotal	99,483.	14,395.	8,000.
4.	Supplies	5,139.	10,983.	5,000.
5.	Travel	88.	2,466.	1,000.
6.	Alterations & Renovations	-0-	-0-	-0-
7.	Publication Costs	-0-	-0-	-0-
8.	Other			
	a. Computer Services	33,286.	66,250.	48,228.
	b. Other	14,667.	3,034.	1,200.
	Subtotal	47,953.	69,284.	49,428.
9.	Subtotal - Direct Costs	214,122.	214.993.	193,508.
10.	Indirect Costs (Rate by Budget period)			
	a. Previous @ 59% of S, W & B (5/1/71 - 8/31/71)	12,008.		
	and @ 46% of NTDC (9/1/71-12/31/71)	25,071.		
	b. Current @ 46% of NTDC		81,032.	
	c. Next @ 46% of NTDC			77,858.
11.	Total Costs	$251,201.	295,125.	271,366.

Grant No. 5 R24 RR00612-02
Section III-A

A. Summary of Resource Expenditures for Period January 1, 1972 through December 31, 1972

Budget Categories	Current Budget (as awarded)	Actual Expenditures thru 8/31/72	Est. Additional Expenditures & Obligations for Remainder of Current Budget Period	Total Estimated Expenditures and Obligations
Personnel (Salaries	101,851.	81,837.	20,014.	101,851.
Fringe Benefits	16,014.	12,318.	3,696.	16,014.
Consultant Services	-0-	-0-	-0-	-0-
Equipment	14,395.	10,224.	4,171.	14,395.
Supplies	10,083.	10,000.	83.	10,083.
Travel (Domestic)	2,466.	700.	1,766.	2,466.
Other:				
Computer Service	66,250.	61,322.	4,928.	66,250.
Other	3,034.	2,239.	795.	3,034.
Total Direct Costs	214,093.	177,875.	36,218.	214,093.
Indirect Costs	81,032.	61,287.	19,745.	81,032.
Totals	295,125.	239,162.	55,963.	295,125.

At present we estimate that the expenditures in the period January 1, 1972, through December 31, 1972, will equal the budget set forth in the Grant Award Statement.

Grant No. 5 R24 RR00612-02
Section III-A

A. Summary of Resource Expenditures
Estimate Next Budget Period Jan. 1, 1973 through December 31, 1973

	Part A	Part B	Part C	Total
Personnel (Salaries)	38,931.	48,636.	24,318.	111,885.
Fringe Benefits	6,359.	7,864.	3,972.	18,195.
Consultant Services	-0-	-0-	-0-	-0-
Equipment				
a. Purchase	-0-	-0-	-0-	-0-
b. Maintenance	-0-	8,000.	-0-	8,000.
Subtotal	-0-	8,000.	-0-	8,000.
Supplies	200.	5,500.	200.	5,900.
Travel (Domestic)	500.	-0-	500.	1,000.
Other:				
a. Computer Services	12,388.	10,000.	25,840.	48,228.
b. Publications	600.	-0-	600.	1,200.
Total Direct Costs	58,978.	80,000.	55,430.	194,408.
Less: Exemptions	<9,688.>(1)	-0-	<17,420.>(1)	<27,108.>(1)
Net Total Direct Costs	49,290.	80,000.	38,010.	167,300.
Indirect Costs @ 46% NTDC	22,673.	36,800.	17,485.	76,958.
TOTAL	81,651.	116,800.	72,915.	271,366.

(1) Computer Service from Campus Facility.

B. Expenditure Details - Direct Costs Only

	Part A	Part B	Part C	Total
Personnel (1)	38,931.	48,636.	24,318.	111,885.
Fringe Benefits	6,359.	7,864.	3,972.	18,195.
Consultant Services	-0-	-0-	-0-	-0-
Equipment: Maintenance	-0-	8,000.	-0-	8,000.
Supplies:				
Misc. office, etc.	200.	-0-	200.	400.
Helium	-0-	500.	-0-	500.
Gas Chromatograph columns	-0-	1,000.	-0-	1,000.
Organic chemicals	-0-	1,000.	-0-	1,000.
Electronic supplies & replacement parts	-0-	3,000.	-0-	3,000.
Subtotal	200.	5,500.	200.	5,900.
Travel: Domestic	500.	-0-	500.	1,000.
Other:				
Computer Services:				
Terminal Service	2,300.	-0-	2,300.	4,600.
Computer Time & Storage-SCC	9,688.	-0-	17,420.	27,108.
Computer Time & Storage-ACME	400.	10,000.	6,120.	16,520.
Subtotal	12,388.	10,000.	25,840.	48,228.
Publications:	600.	-0-	600	1,200.
Grand Total Direct	58,978.	80,000.	55,430.	194,408.

(1) Details of the salary budget by person are being submitted under separate cover.

C. Budget Explanation - Justification

The total budget for the twelve month period, January 1, 1973 - December 31, 1973, is equal to the level of support recommended for the applicable 03 budget period, namely $194,408. The two largest items in this budget request are salaries and computer services. Comments on each budget item are as follows:

1. Personnel (salaries): The salary budget was figured on the basis of current salary rates for each individual on the project plus a 5.5% allowance for merit increases and promotions. Additional details of the salary budget by person are being submitted under separate cover.

2. Fringe Benefits: These have been budgeted at standard university rates which are 16% for the fiscal year ending August 31, 1973, and 17% for the fiscal year ending August 31, 1974.

3. Equipment - Maintenance: Budgeted funds are necessary to maintain the major existing hardware components, including the PDP-11/20 computer as well as the MAT 711 mass spectrometer which ceases to be covered by the manufacturers warranty next year.

4. Supplies: Outside of normal office supplies, the budget includes helium, gas chromatograph columns, organic chemicals and electronic supplies and replacement parts. Budgeted funds are for supplies needed for gas chromatograph operation and chemistry procedures necessary to isolate fractions of urines and other body fluids. Also included is an allocation for needed electronic supplies and replacement parts for in-house equipment maintenance and modifications.

5. Travel: A minimal level of domestic travel funds have been requested for visits by project personnel to other laboratories where work of mutual interest is under way, and for trips to relevant scientific meetings.

6. Computer Services: The current budget for ACME computer services is carried forward prorated for half of the next year, after which time the existing ACME grant expires. It is assumed that DENDRAL computing requirements will be met as one application of the proposed, separately funded SUMEX facility (proposal currently under review by NIH) after July, 1973.