b) GAS CHROMATOGRAPHY/HIGH RESOLUTION MASS SPECTROMETRY

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

c) AUTOMATED GC/MS DATA REDUCTION

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

d) CLOSED-LOOP INSTRUMENT CONTROL

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

REFERENCE - PART B(i)

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

Joshua Lederberg
School of Medicine
Stanford University
Stanford, California 94305

Rapid Calculation of Molecular Formulas from Mass Values

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

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

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

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

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

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

Step 3. We therefore remove 12 H's (12.0939) to give $m'' = m' - 12H = 252.035 \pm 0.001$. The table now has entries at 0.034 ($H_8N_4O_8$), 0.035 ($H_{10}NO_9$) and 0.036 ($H_8N_5O_5$). These will be completed in Step 4. 12 H's are again removed until m_f falls below − 0.0498, the bottom of the table. In our example, this occurs at the next cycle.

Step 4. The table entries are now completed as follows

			Add C's to make up m''	Adjust borrowed H's	Check mass (compare 259.0900 ± 0.0010)
34	0.034216	$H_8N_4O_8$ m_i = C_{16}	C_5	$C_5H_{15}N_4O_8$	259.089
35	0.035559	$H_{10}NO_9$ m_i = C_{14}	C_7	$C_7H_{17}NO_9$	259.090
36	0.036895	$H_8N_5O_5$ m_i = C_{14}	C_8	$C_8H_{13}N_5O_5$	259.092

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

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

In fact $m'' = m - 6N - 7H = 168.017 \pm 0.001$ leads to $C_6H_{11}N_8O_4$, $m = 259.090$. Further, $m - 12N - 7H = 83.999 \pm 0.001$. We read this as $m_i = 84$; $m_f = -0.001$ and find two entries in the table: −0.000826 (H_6NO_{10}) and 0.000510 ($H_2N_2O_6$), whose m_i however >84.

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

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

This work has been supported in part by the Advanced Research Projects Agency (contract SD-183), the National Aeronautics and Space Administration (grant NGR-05-020-004), and the National Institutes of Health (grant GM-00612-01).

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

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

Table of Mass Fractions for all Combinations[a] of H, N, O ⟨H ≦ 10 N ≦ 6 O ≦ 11⟩

Index	$m_f \times 10^6$	H	N	O	=C	Index	$m_f \times 10^6$	H	N	O	=C	Index	$m_f \times 10^6$	H	N	O	=C
−49	−49787	0	2	11	17	0	0	0	0	0	0	31	31537	10	3	11	9
−45	−45765	0	0	9	12	1	510	2	5	6	14	32	32363	4	2	1	14
−38	−38554	0	4	10	18	2	1853	4	2	7	12	34	34216	8	4	8	16
−37	−37211	2	1	11	16	4	4532	2	3	4	9	35	35559	10	1	9	14
−34	−34532	0	2	8	13	5	5875	4	0	5	7	36	36895	6	5	5	13
−30	−30510	0	0	6	8	6	6385	6	5	11	21	38	38238	8	2	6	11
−25	−25978	2	3	10	17	7	7211	0	4	1	6	40	40917	6	3	3	8
−24	−24635	4	0	11	15	8	8554	2	1	2	4	41	42260	8	0	4	6
−23	−23299	0	4	7	14	10	10407	6	3	9	16	42	42770	10	5	10	20
−21	−21956	2	1	8	12	11	11750	8	0	10	14	43	43596	4	4	0	5
−19	−19277	0	2	5	9	13	13086	4	4	6	13	44	44939	6	1	1	3
−15	−15255	0	0	3	4	14	14429	6	1	7	11	46	46792	10	3	8	15
−14	−14745	2	5	9	16	15	15765	2	5	3	10	49	49471	8	4	5	12
−13	−13402	4	2	10	18	17	17108	4	2	4	8	50	50814	10	1	6	10
−10	−10723	2	3	7	13	18	18961	8	4	11	20	52	52150	6	5	2	9
−9	−9380	4	0	8	11	19	19787	2	3	1	5	53	53493	8	2	3	7
−8	−8044	0	4	4	10	20	21130	4	0	2	3	56	56172	6	3	0	4
−6	−6701	2	1	5	8	21	21640	6	5	8	17	57	57515	8	0	1	2
−4	−4022	0	2	2	5	22	22983	8	2	9	15	58	58025	10	5	7	16
−2	−2169	4	4	9	17	25	25662	6	3	6	12	62	62047	10	3	5	11
−1	−826	6	1	10	15	27	27005	8	0	7	10	64	64726	8	4	2	8
						28	28341	4	4	3	9	66	66069	10	1	3	6
						29	29684	6	1	4	7	68	68748	8	2	0	3
						30	31020	2	5	0	6	73	73280	10	5	4	12
												77	77302	10	3	2	7
												81	81324	10	1	0	2
												88	88535	10	5	1	8
(−0.049 to −0.0008)						(0 to 0.03)						(0.03 to 0.088)					

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

[Reprinted from Journal of Chemical Education, Vol. 49, Page 613, September, 1972.]

PART B(ii):

ANALYSIS OF THE

CHEMICAL CONSTITUENTS OF BODY FLUIDS

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

OBJECTIVES:

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

BACKGROUND AND PROGRESS:

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

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

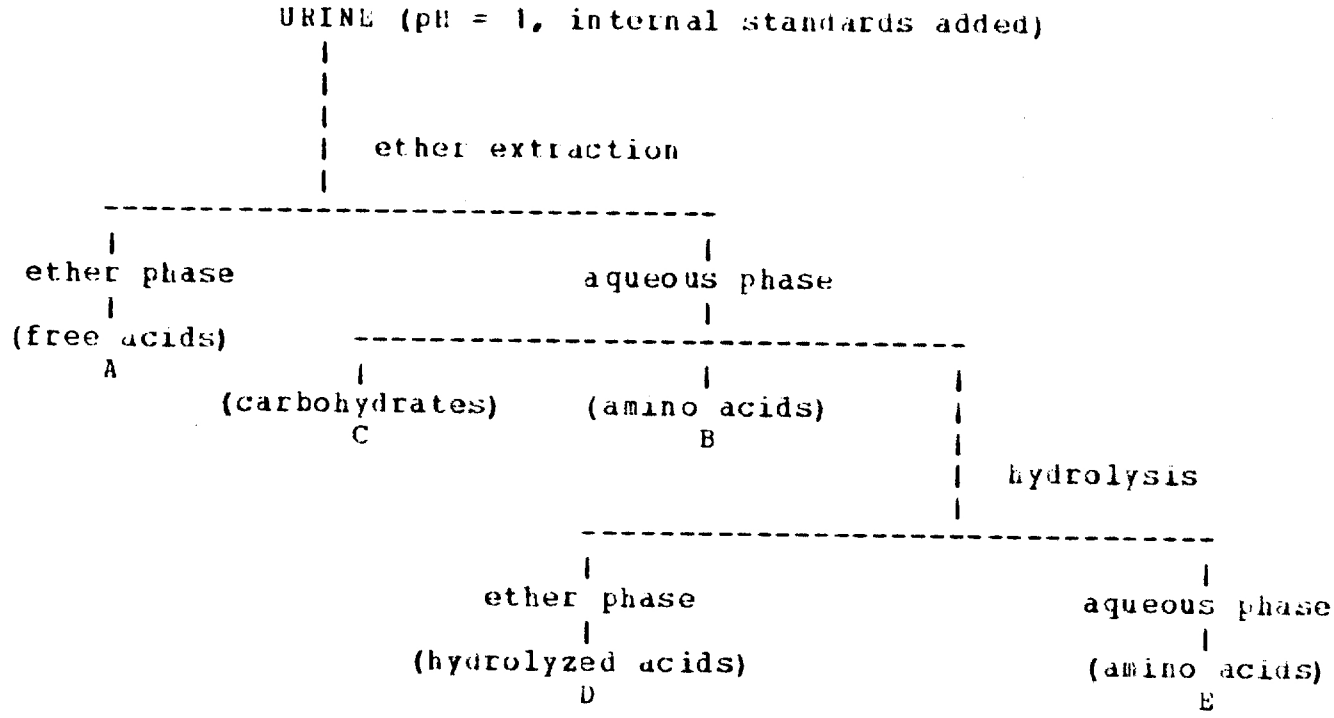
specific and provide a powerful means of positive identification of metabolites in human body fluids (ref 11-13). Of these techniques, gas chromatography is the most convenient to interface to the mass spectrometer because the carrier gas can easily be removed as the analysis proceeds on a continuous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

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

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

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

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



The experimental procedure used for working with a urine sample is as follows. To an aliquot (2.5 ml.) of a 24 hour urine sample is added 6N hydrochloric acid until the ph is 1. Two internal standards, n-tetracosane and 2-amino octanoic acid are then added. Ether extraction isolates the free acids (fraction A) which are then methylated and analyzed by gas chromatography-mass spectrometry. An aliquot of the aqueous phase (0.5 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). Another aliquot (0.5 ml) of the aqueous phase can be derivatized for the detection of carbohydrates (Fraction C).

Concentrated hydrochloric acid (0.15 ml) is added to the urine (1.5 ml) after ether extraction and the mixture hydrolyzed for 4 hours under reflux. Ether extraction separates the hydrolyzed acid fraction (D) which is then methylated and analyzed by GC/MS. A portion of the aqueous phase (0.5 ml) from hydrolysis of the urine is concentrated to dryness and derivatized and analyzed for amino acids (Fraction E).

As an example of the application of these methods to biomedical problems, we can use some recent studies we have undertaken on the urine of a patient suffering from acute lymphoblastic leukemia. The gas chromatographic profile (Figure 1) of the amino acid fraction of his urine showed the presence of an abnormal peak (A). The mass spectra (Figure 2) recorded during the lifetime of this chromatographic peak identified this

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

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

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

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

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

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

FUTURE PLANS

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

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

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

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

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

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

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

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

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

REFERENCES

1) Schwartz, M.K., "Biochemical Analysis," Anal. Chem., 44, p. 9R, (1972).

2) Heathcote, J.G., Davies, D.M., and Haworth, C., "The Effect of Desalting on the Determination of Amino Acids in Urine by Thin Layer Chromatography." Clin. Chim. Acta, 32, p. 457 (1971).

3) Davidow, B., Petri, N.L., and Quame, B., "A Thin Layer Chromatographic Screening Procedure for Detecting Drug Abuse," Amer. J. Clin. Pathol., 50, p 714, (1968).

4) Efron, K. and Wolf, P.L., "Accelerated Single-column Procedure for Automated Measurement of Amino Acids in Physiological Fluids," Clin. Chem., 18, p 621, (1972).

5) Burtis, C.A., "The Separation of the Ultraviolet-absorbing Constituents of Urine by High Pressure Liquid Chromatography," J. Chromatog., 52, p 97, (1970).

6) Wilson-Pitt, W., Scott, C.D., Johnson, W.P., and Jones, G., "A Bench-top, Automated, High-resolution Analyzer for Ultraviolet Absorbing Constituents of Body Fluids," Clin. Chem., 16, p. 657 (1970).

7) Dalgliesh, C.E., Horning, E.C., Horning, M.G., Knose, K.L., and Yarger, K., "A Gas-Liquid Chromatographic Procedure for separating a Wide Range of Metabolites Occurring in Urine or Tissue Extracts," Biochem. J., 101, p. 792 (1966).

8) Teranishi, R., Mon, T.R., Robinson, A.B., Cary, P., and Pauling, L., "Gas Chromatography of Volatiles from Breath and Urine," Anal. Chem., 44, p. 18, (1972).

9) Pauling, L., Robinson, A.B., Teranishi, R., and Cary, P., "Quantitative Analysis of Urine Vapor and Breath by Gas-liquid Partition Chromatography," Proc. Nat. Acad. Sci. USA, 68, p. 2374, (1971).

10) Zlatkis, A. and Liebich, H.M., "Profile of Volatile Metabolites in Human Urine," Clin. Chem., 17, 592 (1971).

11) Mrochek, J.E., Putts, W.C., Rainey, W.T., and Burtis, C.A., "Separation and Identification of Urinary Constituents by Use of Multiple-analytical Techniques," Clin. Chem., 17, p.72 (1971).

12) Horning, E.C. and Horning, M.G., "Human Metabolic Profiles Obtained by GC and GC/MS," J. Chromatog. Sci., 9, p. 129, (1971)

13) Jellum, E., Stokke, O., and Eldjarn, L., "Combined Use of Gas Chromatography, Mass Spectrometry, and Computer in Diagnosis

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

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

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

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

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

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

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

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

P-56

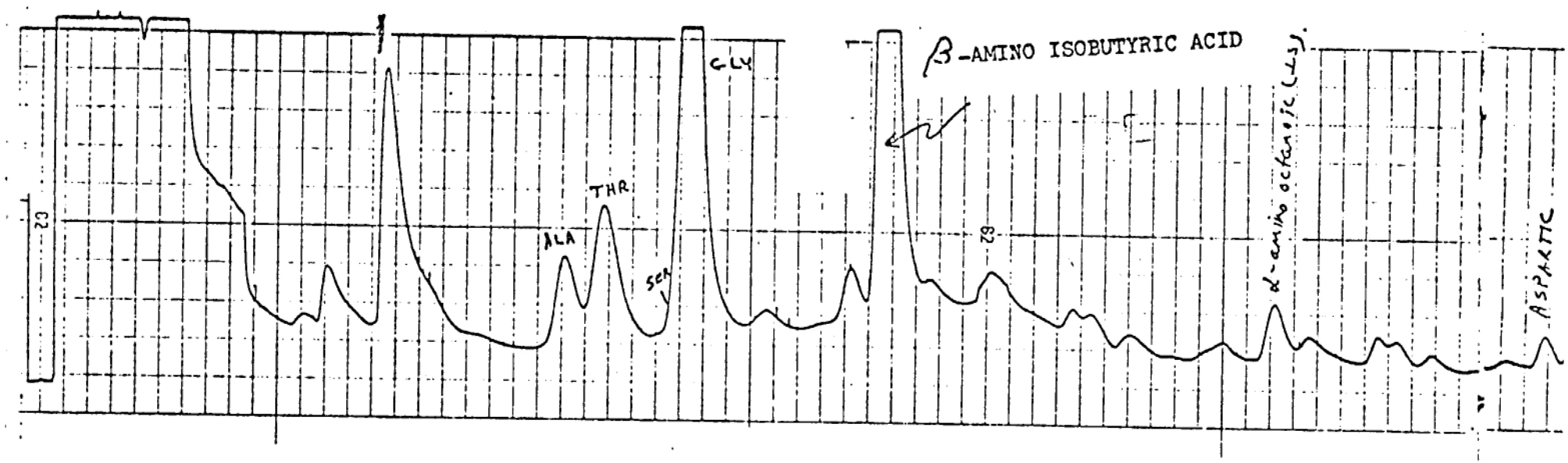


FIGURE 1

Gas Chromatogram of the Amino Acid Fraction of Urine

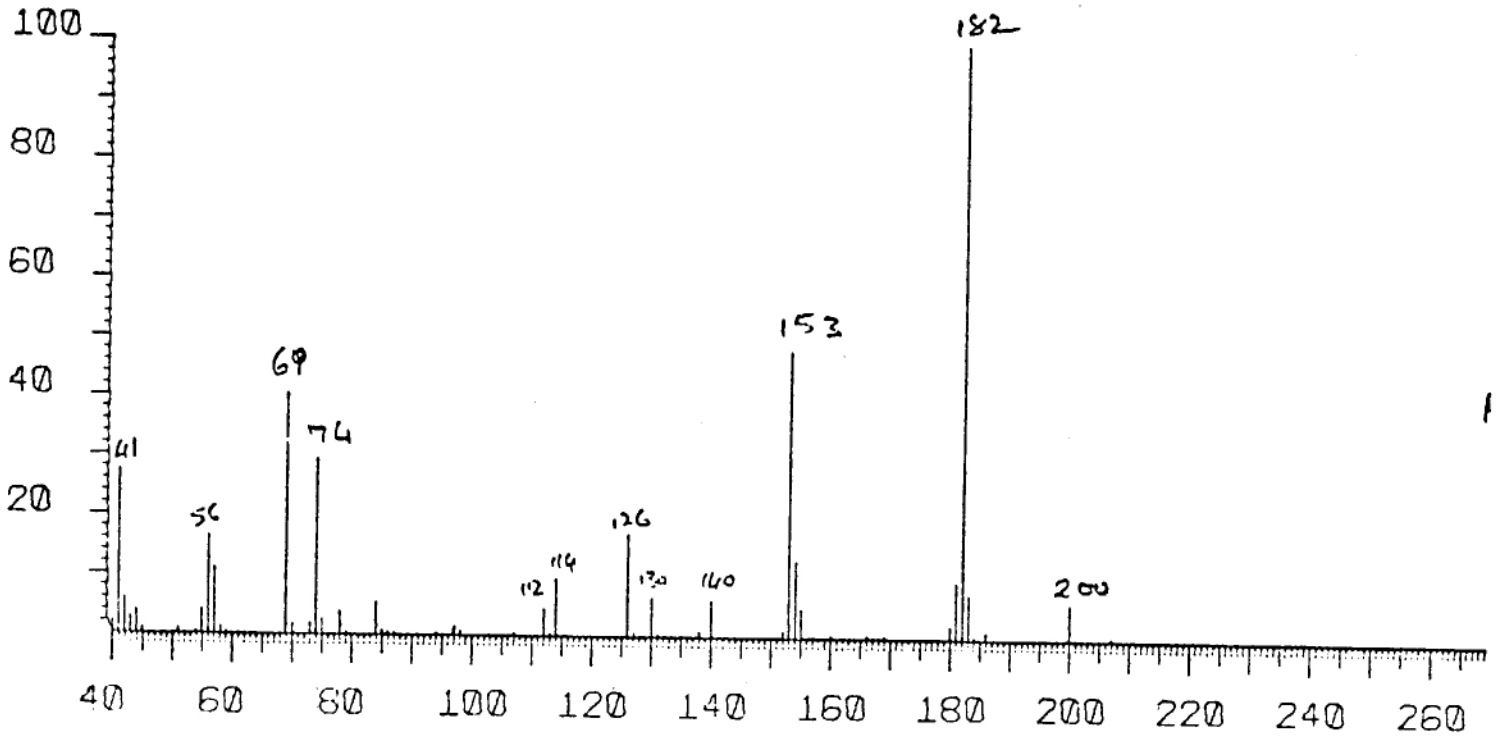


FIGURE 2

Mass Spectrum of Beta-Amino Isobutyric Acid

PART C:

EXTENSION OF THE

THEORY OF MASS SPECTROMETRY BY COMPUTER

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

OBJECTIVES:

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

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

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

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

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

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

PROGRESS:

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

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

Summary of Segment 1

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

Summary of Segment 2

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

Summary of Segment 3

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

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

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

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

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

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

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

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

The output of this program, even though crude in many

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

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

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

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

Plans:

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

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

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

The program is now limited to forming rules which are more

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

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

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

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

PUBLICATIONS -- PART C

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

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

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

PART D:

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

PART D. CARBON-13 NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

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

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

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

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

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

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

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

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

PUBLICATIONS -- PART D

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

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

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

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

Raymond E. Carhart[2] and Carl Djerassi,* Departments of Computer Science and Chemistry, Stanford University, Stanford, California, 94305, U. S. A.

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

I. INTRODUCTION

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

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

However, recent studies indicate that C-13 nuclear magnetic resonance (CMR) spectroscopy[8] is an exception. For several classes of compounds,[9] rules have been obtained which allow one to predict the CMR spectrum of a substance from its molecular structure, and in all cases, the rules indicate that the chemical shift of any Carbon, even one in a large alkyl chain-end, depends heavily upon branching at nearby centers. Thus, it appears that CMR spectroscopy, either alone or in combination with other methods, could be a powerful tool in the computerized analysis of molecular structure. This paper outlines the methods by which such an analysis may be carried out for the acyclic amines, and describes a FORTRAN IV computer program,[10] entitled AMINE, in which these methods are implemented.