

Technical Report No. IRL 1056

CYTOCHEMICAL STUDIES OF PLANETARY MICROORGANISMS EXPLORATIONS IN EXOBIOLGY

Status Report Covering Period October 1, 1966 to April 1, 1967
For
National Aeronautics and Space Administration
Grant NsG 81-60



GPO PRICE \$ _____

CESTI PRICE(S) \$ _____

Hard copy (HC) 3.00

Microfiche (MF) 1.65

FACILITY FORM 602

N67-38527

(ACCESSION NUMBER)

(THRU)

98

(PAGES)

(CODE)

NP-88326

(NASA CR OR TMX OR AD NUMBER)

04

(CATEGORY)

853 July 65

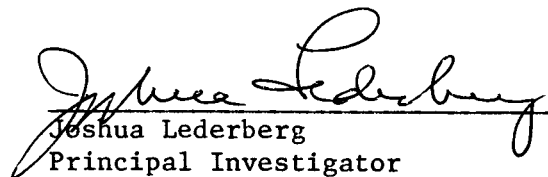
Instrumentation Research Laboratory, Department of Genetics
Stanford University School of Medicine
Palo Alto, California


Report to the National Aeronautics and Space Administration
"Cytochemical Studies of Planetary Microorganisms - Explorations in Exobiology"

NsG 81-60

Status Report Covering Period October 1, 1966 to April 1, 1967

Instrumentation Research Laboratory, Department of Genetics
Stanford University School of Medicine
Palo Alto, California


Joshua Lederberg
Principal Investigator


Elliott C. Levinthal, Director
Instrumentation Research Laboratory

A. INTRODUCTION

This Status Report covers the activities of the Instrumentation Research Laboratory from October 1, 1966 to April 1, 1967. Major technical efforts are described in separate technical reports and papers. The status report refers to these and summarizes continuing projects.

We are now located in the new laboratory facilities provided by NASA Grant NsG-(F)-2. Work under grant NsG 81-60 includes areas of research that are closely related to efforts being carried out in the Department of Genetics under other grants or contracts. This includes Air Force Contract AF 49(638)1599 for "Molecular Biology Applications of Mass Spectroscopy," National Institute of Neurological Diseases and Blindness Grant NB-04270 entitled "Molecular Neurobiology" and for work carried out by the Advanced Computer for Medical Research (ACME) program supported by the National Institutes of Health, Division of Research Facilities and Resources under Grant FR00311-01. There is collaboration with the work in the Computer Science Department on artificial intelligence carried out under support of the Advanced Research Projects Agency SD 183. In addition, work is being done on "Genetic Studies of Mammalian Cells," National Institutes of Health under Grant CA04681-08. The relationship of the work carried out under this NASA grant to these other activities continues to prove of great mutual benefit in all cases.

The general project areas of the resume are:

- I. Fluorometry
- II. Gas Chromatography and Optical Resolution
- III. Mass Spectrometry
- IV. Computer Managed Instrumentation
- V. Atmospheric Effects on Photographic Resolution

These projects contribute to technical mastery of problems in exobiology by furnishing specific analytical techniques of high sensitivity and discrimination for the detection of exotic life. In addition, the management of instrumentation in many laboratories via a time-shared computer (the ACME project) is a system prototype for the automated biological laboratory.

A status report prepared by ACME for NIH covering this same report period is included as Appendix A of this report.

In connection with the work on Computer Manipulation of Chemical Hypotheses, a recent report on DENDRAL - A Computer Program for Generating and Filtering Chemical Structures, is included as Appendix B.

During the six month period described above, four papers were submitted to journals for publication, in addition to those of Professor Djerassi's laboratory and the ACME group. A listing of these papers is included in this status report. Information covering personnel changes is presented.

B. PROGRAM RESUME

I. Fluorometry

A number of Bacillus bacteria have been isolated from Chilean desert soil and provisionally identified according to their aminopeptidase profile. These isolates are now being identified by classical taxonomic techniques at Ames Research Laboratory in order to test the usefulness of the aminopeptidase method of identification. In addition, other types of microorganisms including fungi and yeasts will be tested by Ames to see how widespread the aminopeptidases are in nature.

II. Gas Chromatography and Optical Resolution

a. Gas Chromatography

In order to try and understand what factors affect the GLC resolution of diastereoisomers a number of widely differing amides and esters have been synthesized and tested on the gas chromatograph. The results are summarized in Table 1.

In the previous report we discussed the GLC behavior of diastereoisomeric menthyl esters and showed that it was necessary to have an additional amide function in the resolving agent to achieve separation of the diastereoisomers. In the diastereoisomeric amides this condition is not necessary, probably because of the greater rigidity of the amide bond. The importance of conformational immobility in the resolution of diastereoisomeric amides is further confirmed by the high α -values obtained for the cyclic amines 2-ethylpiperidine and decahydroquinoline. Finally we have used GLC of diastereoisomers to correlate the absolute configuration of organic compounds. This is based on the observation that diastereoisomers derived from a homologous series have similar chromatographic behavior (i.e., the L-L diastereoisomer has a longer retention volume than the L-D). Using this technique, we have already

Table 1. Resolution of diastereoisomeric esters and amides by GLC*

Resolving agent	$ \begin{array}{c} R_1 \diagdown \\ R_2 - C - \\ R_3 \diagup \end{array} \begin{array}{c} O \\ \\ -C- \end{array} $			Ratio of retention times of diastereoisomers (α)**				
	R_1	R_2	R_3	Menthol	Valine Methylester	Amphetamine	2-ethyl- piperidine	Decahydroquinoline
α Methylbutyric	H	CH ₃	C ₂ H ₅	1.0	1.0	1.0	1.08	1.09
α chloropropionic	H	CH ₃	Cl	1.0	1.12	1.0	1.0	1.0
2-chloro-3-Me butanoyl	H	CH ₃ \diagup CH CH ₃ \diagdown	Cl	1.0	1.22	1.04	1.16	1.19
2-chloro-4-Me pentanoyl	H	CH ₃ \diagup CH-CH ₂ CH ₃ \diagdown	Cl	1.0	1.18	1.10	1.0	1.11
α phenylbutyric	H	C ₂ H ₅	C ₆ H ₅	1.0	1.0	1.08	1.0	1.07
α phenoxy propionic	H	CH ₃	C ₆ H ₅ O	1.0	1.0	1.0	1.15	1.08
N-TFA-alanine	H	CH ₃	CH ₃ CONH	1.13	1.06			
N-acetyl-alanine	H	CH ₃	CH ₃ CONH	1.15	1.0	1.0		
N-formyl-alanine	H	CH ₃	HCONH	1.11				
N-benzoyl-alanine	H	CH ₃	C ₆ H ₅ CONH	1.12				
N-TFA-proline	H	(CH ₂) ₃	CH ₃ CON	1.09	1.28	1.25	1.25	1.34

* GLC resolutions were carried out on 5' x 1/8" stainless steel columns packed with either 0.5% EGA or 1% cyclohexanedimethanol succinate on gas-chrom Q.

** α values are quoted for the best conditions of resolution using the above columns. Attempts will be made using capillary columns to resolve the diastereoisomers which were not resolved on packed columns ($\alpha = 1.0$).

correlated the absolute configuration of a series of α -phenylacetic acids and are now using GLC to correlate the configuration of some naturally occurring alkaloid bases with cyclic amines of known configuration.

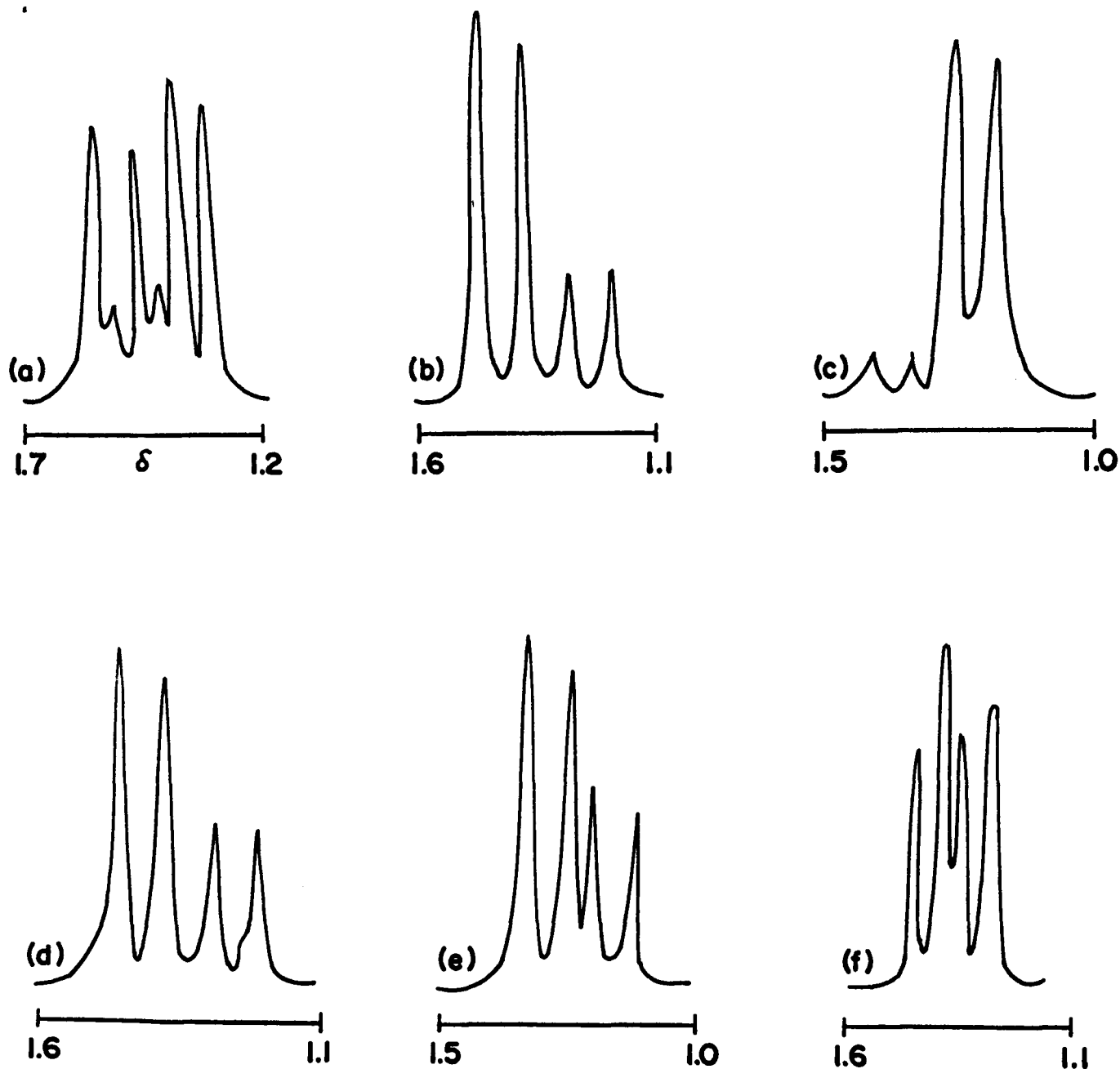
b. Determination of Steric Purity of Peptides by NMR Spectroscopy

Since chemical modification of a sample is not required, NMR spectroscopy may be a convenient technique for the determination of "optical" purity of peptides. Magnetic non-equivalence in diastereoisomeric dipeptides has been demonstrated earlier¹, but a second study using strict pH control concluded that diastereoisomeric alanyl peptides have identical NMR spectra.² We have now reexamined the NMR data of diastereoisomeric di and tri peptides and our results show that NMR can be used to detect contamination of one diastereoisomer with the other. The presence of an aromatic amino acid in the peptide greatly enhances the observable change in chemical shift and in these cases the steric purity and the absolute configuration of the peptide can be determined by NMR (Fig. 1).

Besides the determination of optical purity of free peptides, much effort has been devoted to the study of racemization during the formation of the peptide linkage. Present techniques rely on the separation of "model" diastereoisomers by fractional crystallization, counter current distribution, vapor phase chromatography, and paper chromatography. Because a cumbersome physical separation is not necessary when NMR is used to analyze diastereoisomeric mixtures, we have examined the methyl resonances of several model n-acylated peptide derivatives (Fig. 2). Our results suggest that a NMR measurement is a convenient technique to study racemization in peptide synthesis. A detailed investigation on the influence of coupling agents, solvent and N-acyl groups on the steric purity of peptide bond formation is now in progress.

¹T. Wieland and H. Bende, Chem. Ber., 98, 504 (1965).

²M. VanGorkom, Tetrahedron Letters, 5433 (1966).



FIGURES 1 and 2

N.m.r. spectra in the region of the methyl resonance for alanyl peptides. (a) L-alanyl-D-alanine and D-alanyl-D-alanine (1:9; D_2O ; pH 5; SDSS); (b) L-alanyl-L-phenylalanine and L-alanyl-D-phenylalanine (3:1; D_2O ; pH 3; SDSS); (c) L-phenylalanyl-L-alanine and L-phenylalanine-D-alanine (1:9; D_2O ; pH <1; SDSS); (d) DL-alanyl-DL-phenylalanine (commercial sample; D_2O ; pH 5; SDSS); (e) glycyl-DL-phenylalanyl-D-alanine (D_2O ; pH 5; SDSS); (f) glycyl-DL-alanyl-L-phenylalanine (D_2O ; pH >10; SDSS).

III. Mass Spectrometry

a. Analysis of Natural Products

The Atlas CH-4 Mass Spectrometer in Professor Djerassi's laboratory in the Department of Chemistry has yielded the results reported in the following papers:

- Gutzwiller, J.; Djerassi, C.: Mass Spectrometry in Structural and Stereochemical Problems CXXI. (Massenspektrometrie und ihre Anwendung auf strukturelle und stereochemische Probleme CXXI. Fragmentierungen und Wasserstoffverschiebungen in 4-Keto-androstanen). Helv. Chim. Acta, 49, 2108 (1966).
- Duffield, A. M.; Carpenter, W.; Djerassi, C.: Mass Spectrometry in Structural and Stereochemical Problems CXXIX. The Course of the Electron Impact Induced Elimination of Hydrogen Sulphide from Aliphatic Thiols. Chem. Comm., 109 (1967).
- Diekman, J.; Djerassi, C.: Mass Spectrometry in Structural and Stereochemical Problems CXXV. Mass Spectrometry of Some Steroid Trimethylsilyl Ethers. J. Org. Chem., 32, 1005 (1967).
- Brown, Peter; Djerassi, C.: Mass Spectrometry in Structural and Stereochemical Problems CXXX. A Study of Electron Impact Induced Migratory Aptitudes. (In press)
- Harris, R. L. N.; Komitsky, F., Jr.; Djerassi, C.: Mass Spectrometry in Structural and Stereochemical Problems CXXXII. Electron Impact Induced Alkyl and Aryl Rearrangements in α,β -Unsaturated Cyclic Ketones. (In press)
- Harris, R. L. N.; Komitsky, F., Jr.; Djerassi, C.: Mass Spectrometry in Structural and Stereochemical Problems CXXXIV. Electron Impact Induced Alkyl and Aryl Rearrangements in α -Arylidene Cyclic Ketones. (In press)
- Duffield, A. M.: Mass Spectrometric Fragmentation of Some Lignans. J. Heterocyclic Chem., 4, 16 (1957).
- Shapiro, R. H.; Serum, J. W.; Duffield, A. M.: Mass Spectrometric and Thermal Fragmentation of 1-Substituted 3-phenyl-2-thioureas. J. Org. Chem. (In press)

b. Sequence Analysis of Peptides by Mass Spectrometry

The determination of the amino acid sequence of a particular protein is often necessary for an increased understanding of a biological process. Present sequencing techniques rely on a stepwise chemical degradation procedure which is monitored by fluorescence labelling, total hydrolysis of the peptide and a chromatographic separation. Recently, several mass spectrometric methods for sequencing have been proposed which are based on the fact that the structure of a linear molecule is determined unequivocally by using only the possible fragments which contain one end of the chain. Thus in a molecule A-B-C-D-E-F, the six fragments of the molecule containing part A are sufficient for the determination of the sequence. Several methods for marking the end of the peptide chain are possible, but so far only high resolution mass spectrometry combined with a sophisticated computer program appear to provide a general method for sequence analysis.

In view of the large capital investment in such an installation, we are investigating the use of low resolution mass spectrometry for this purpose. Our approach involves the incorporation of chlorine into the N-terminal amino acid of the peptide chain. The isotope ratio of the heteroatom ($^{35}\text{Cl} : ^{37}\text{Cl} = 75.5:24.5$) assists in the identification of all chlorine containing fragments in the mass spectrum and the m/e of these fragments give the peptide sequence. The chemical operation involved in the chlorine labelling of the peptide is the dissolving of the sample in formic acid and treatment with nitrosyl chloride at 0°C . After removal of the excess reagent and the solvent, the residue is treated with diazomethane. The resulting chloropeptide esters are then used for mass spectrometry. Mass spectra of chlorine labeled tri, tetra and penta peptides containing neutral and acidic amino acids have already yielded the correct amino acid sequence. The behavior of the basic amino acid and tyrosine in the NOCl reaction and the mass spectrometry of peptides containing these amino acids are still under investigation.

c. Mass Spectral Microanalysis of Organic Solids

We are continuing our study into the efficacy of laser induced vaporization as a mechanism for enabling spatial resolution in the mass spectral analysis of organic solids of biological interest.

Experiments have been conducted with an Optics Technology Model 100 pulsed ruby laser coupled to a Bendix Time-of-Flight (TOF) mass spectrometer (MS). The manufacturer's specifications cite a peak power of 1×10^3 watt, pulse width of 0.5×10^{-3} sec, and beam divergence of 0.5×10^{-3} rad at threshold. None of these quantities are given precise definition by the manufacturer.

Our measurements have indicated a lasing threshold of 190 joules input with a laser output of 3.5×10^{-2} joules at the maximum available input of 440 joules. The wavelength of the laser radiation is 6943 Å. Beam divergence at threshold appears to be in approximate accord with the specified figure. We have some evidence that the full angle beam divergence at maximum output is approximately 5×10^{-3} rad; we have not, however, performed accurate determinations of the directional dependence of the beam brightness of the radiation issuing from the laser.

Figure 3 illustrates schematically the optical configuration utilized for the mass spectral studies undertaken to date. A portion of the radiation issuing horizontally from the pulsed ruby laser 1 is reflected downward by the beam splitter 2 toward the simple biconvex converging lens 3 of focal length 100 mm, transmitted through the planar glass window 4 to focus at the target 5 supported by probe 6 inside the source chamber in the vacuum environment of the MS.

Alignment and aiming are accomplished by arranging that the quasi-CW radiation from a 1 milliwatt Optics Technology Model 170 He-Ne gas laser 7, reflected off mirror 8, partially transmitted through splitter 2, and reflected off the ruby and mutually parallel mirrors in the laser

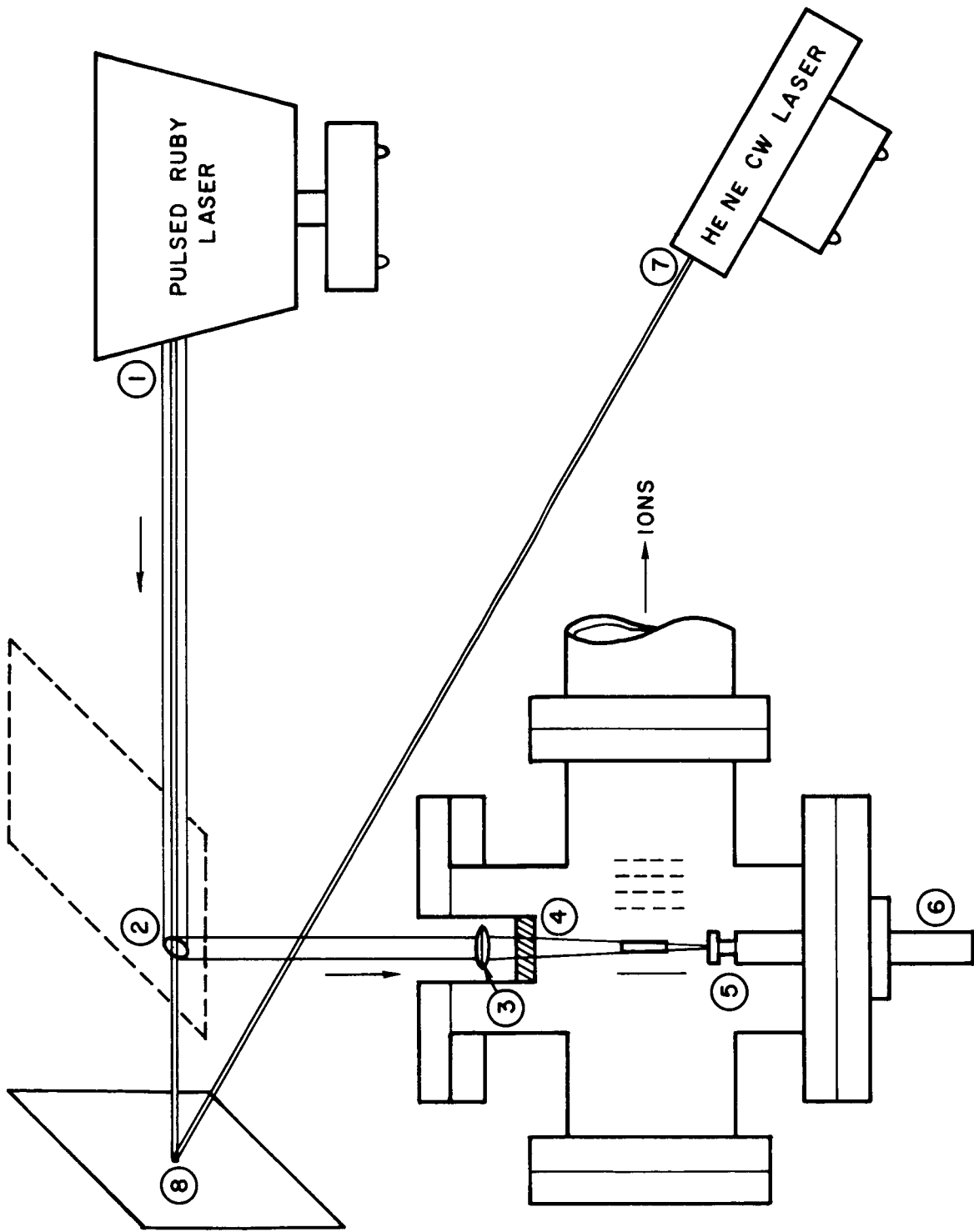


FIGURE 3

Schematic of Laser Optical Configuration
Used for Preliminary Studies

head 1 returns to the point of origin at the gas laser 7. Since this alignment ensures normal reflection of CW radiation off the laser optics, it follows that the pulsed ruby radiation will be incident at the same point on the target as that illuminated by the portion of the CW radiation following the path 1-2-3-4-5. The point of impact of the laser radiation is adjusted by horizontal translation of the converging lens 3.

Each 100 μ sec the MS fires a 0.25 μ sec duration burst of ionizing electrons transversely across the source chamber immediately above that target. That portion of the material that is vaporized by the laser and is positively ionized by each electron burst is impelled, by application of a 2700 volt kick, into a linear field-free drift space at the terminal end of which is a high speed ion detector.

The voltage kick V imparts to each ion a velocity v given approximately by

$$v \approx \sqrt{2neV/m} , \quad (1)$$

where n is the number of electrons removed from the molecule, e is the magnitude of the electronic charge, and m is the mass of the ion. The elapsed time between application of the kick and detection of arrival of an ion is given approximately by

$$t \approx \ell/v , \quad (2a)$$

where t is the transit time down the drift tube of length ℓ .

Substitution of (1) into (2) yields

$$t \approx \ell \sqrt{m/(2neV)} . \quad (2b)$$

Conversion of the mass-to-charge dependent velocity dispersion into ion time-of-arrival dispersion at the detector enables masses to be identified from the cathode ray oscilloscope (CRO) display of ion current versus time. The capture of maximum spectral information over one or a succession of MS repetition cycles (several of the 100 μ sec intervals) has involved photographing the CRO display of output. Utilization of a

procedure based upon the approximate quadratic relationship between m and t expressed in (2b) has enabled us to associate mass numbers with recorded peaks out to the neighborhood of 400 atomic mass units (amu).

Figure 4 is a schematic representation of the initial version of the pulsing circuitry utilized external to the MS to enable the recording of either one or a series of successive spectra. The system allows for the establishment of a controllable delay between firing of the laser and the initiation of the CRO display. The number of successive spectra displayed can be freely varied. Successive spectra are vertically displaced from one another on the CRO. The joint resolution of the CRO and the 10,000 ASA Polaroid film is insufficient to enable unambiguous mass identification over the range 0-400 on a single trace. Recording of this mass range has generally required 5 to 10 laser shots, the mass range window for each shot being successively displaced to higher masses.

The mass spectral analysis of the vapor produced by laser irradiation of a powdered crystalline target of N-dinitrophenyl (DNP)-L-isoleucine is presented in Table 2. There is also presented, for comparative purposes in this table, the spectrum obtained by conventional crucible warming of the same sample to 40°C. The structural and graphic formulae for this sample are presented in Figure 5.

We have found a high degree of variability in the amplitude of the spectrum on repeated runs with different preparations of the same target material. We at first suspected the laser-optical chain (laser output, optical alignment, sample placement). It is our present belief, however, that the degree of powdering of the crystalline material has been responsible for most of the variability.

The photographic recording process and subsequent identification and evaluation of mass peaks is at present a time consuming process that does not readily lend itself to rapid scanning of a heterogeneous target.

Table 2. Relative Intensities of Laser Spectrum and Crucible (40 Deg. C) Spectrum of (DNP)/L/Isoleucine.

Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.	Mass Laser Cruc.					
1	0	1	51	14	1	101	10	4	151	0	0	201	0	16	251	0	0
2	0	1	52	25	1	102	14	1	152	0	0	202	20	3	252	72	5
3	0	0	53	23	1	103	27	1	153	0	0	203	0	0	253	0	0
4	0	0	54	22	0	104	17	0	154	0	0	204	0	0	254	0	0
5	0	0	55	40	2	105	7	0	155	0	0	205	0	0	255	0	0
6	0	0	56	8	0	106	5	0	156	0	0	206	24	0	256	0	0
7	0	0	57	100	3	107	8	0	157	0	0	207	0	0	257	0	0
8	0	0	58	7	0	108	0	0	158	0	0	208	0	0	258	0	0
9	0	0	59	11	0	109	0	0	159	0	0	209	0	0	259	0	0
10	0	0	60	0	0	110	0	0	160	0	0	210	0	0	260	0	0
11	0	0	61	3	0	111	0	0	161	0	0	211	0	0	261	0	0
12	0	0	62	10	0	112	0	0	162	0	0	212	0	0	262	0	0
13	0	0	63	38	1	113	0	0	163	0	0	213	0	0	263	0	0
14	3	5	64	28	0	114	8	0	164	10	0	214	0	0	264	0	0
15	4	0	65	14	0	115	0	0	165	0	0	215	0	0	265	0	0
16	0	5	66	8	0	116	0	0	166	11	1	216	0	0	266	0	0
17	3	20	67	14	0	117	12	0	167	0	0	217	0	0	267	0	0
18	14	100	68	8	0	118	3	0	168	0	0	218	0	0	268	0	0
19	0	0	69	45	1	119	12	0	169	0	0	219	0	0	269	0	0
20	0	0	70	7	0	120	2	0	170	0	0	220	0	0	270	0	0
21	0	0	71	0	0	121	0	0	171	0	0	221	0	0	271	0	0
22	0	0	72	0	0	122	5	0	172	0	0	222	0	0	272	0	0
23	0	0	73	5	0	123	0	0	173	0	0	223	0	0	273	0	0
24	0	0	74	12	0	124	0	0	174	0	0	224	13	0	274	0	0
25	0	0	75	48	2	125	0	0	175	0	0	225	0	0	275	0	0
26	0	0	76	19	1	126	0	0	176	0	0	226	0	0	276	0	0
27	46	2	77	20	1	127	0	0	177	0	0	227	0	0	277	0	0
28	67	100	78	20	1	128	0	0	178	0	0	228	0	0	278	0	0
29	100	7	79	0	0	129	0	0	179	0	0	229	0	0	279	0	0
30	62	1	80	0	0	130	7	0	180	0	0	230	0	0	280	0	0
31	12	1	81	0	0	131	7	0	181	0	0	231	0	0	281	0	0
32	28	37	82	11	0	132	3	0	182	0	0	232	0	0	282	0	0
33	0	0	83	0	0	133	0	1	183	0	0	233	0	0	283	0	0
34	0	0	84	7	0	134	14	1	184	0	1	234	0	0	284	0	0
35	0	0	85	0	0	135	0	0	185	0	0	235	0	0	285	0	0
36	0	0	86	0	0	136	0	0	186	0	0	236	0	0	286	0	0
37	0	0	87	5	0	137	0	0	187	0	0	237	0	0	287	0	0
38	0	0	88	0	0	138	0	0	188	0	0	238	0	0	288	0	0
39	55	2	89	0	0	139	0	0	189	0	0	239	0	0	289	0	0
40	22	2	90	11	0	140	0	0	190	0	0	240	8	1	290	0	0
41	100	7	91	0	0	141	0	0	191	0	0	241	0	0	291	0	0
42	18	1	92	0	0	142	0	0	192	0	0	242	0	0	292	0	0
43	24	2	93	0	0	143	0	0	193	0	0	243	0	0	293	0	0
44	0	1	94	0	0	144	0	0	194	0	0	244	0	0	294	0	0
45	37	1	95	0	0	145	0	0	195	0	0	245	0	0	295	0	0
46	0	0	96	0	0	146	0	0	196	22	1	246	0	0	296	0	0
47	0	0	97	0	0	147	7	0	197	0	0	247	0	0	297	0	1
48	0	0	98	0	0	148	7	0	198	0	5	248	0	0	298	0	0
49	0	0	99	10	3	149	0	0	199	0	8	249	0	0	299	0	0
50	22	0	100	0	3	150	0	0	200	0	12	250	0	0	300	0	0

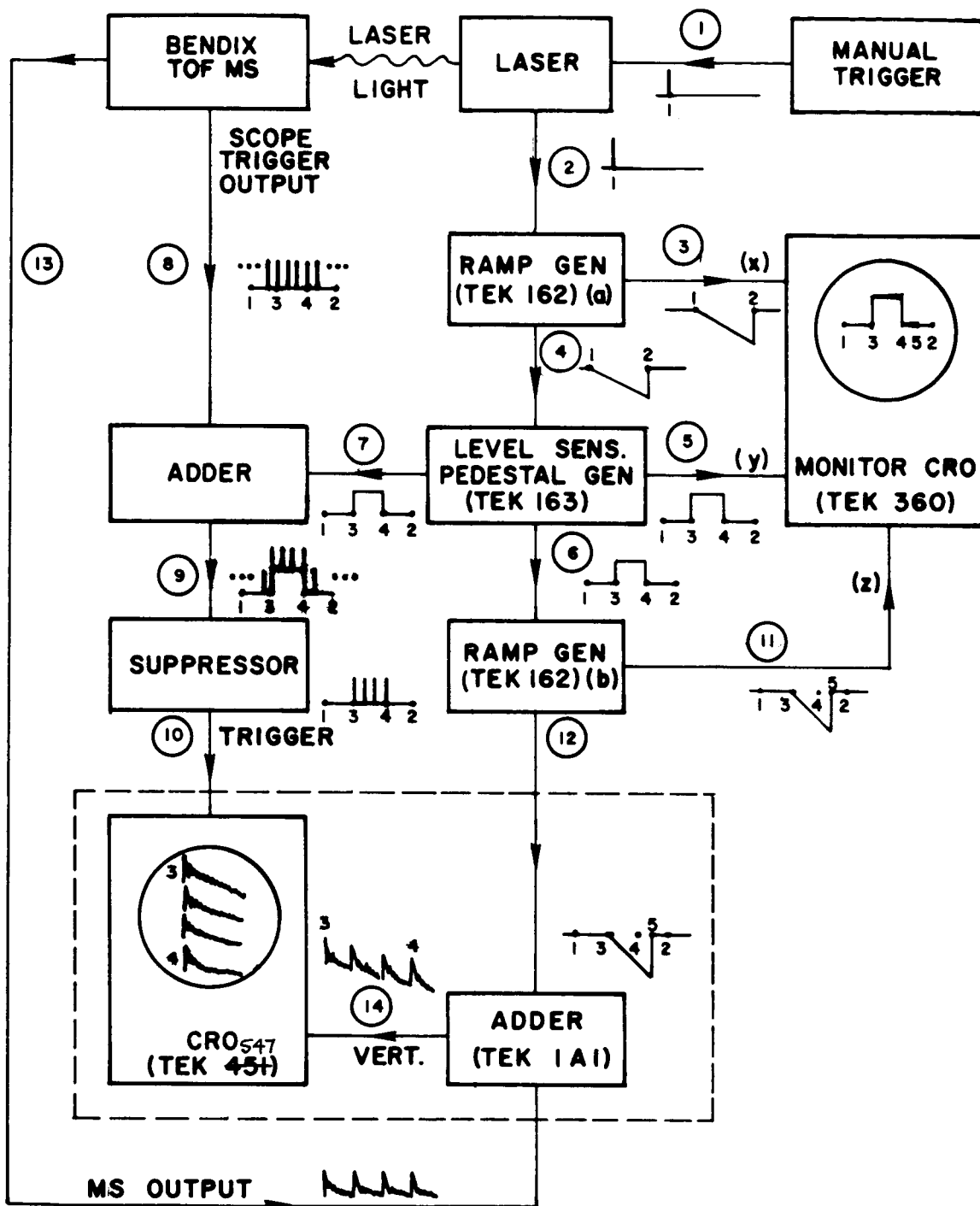
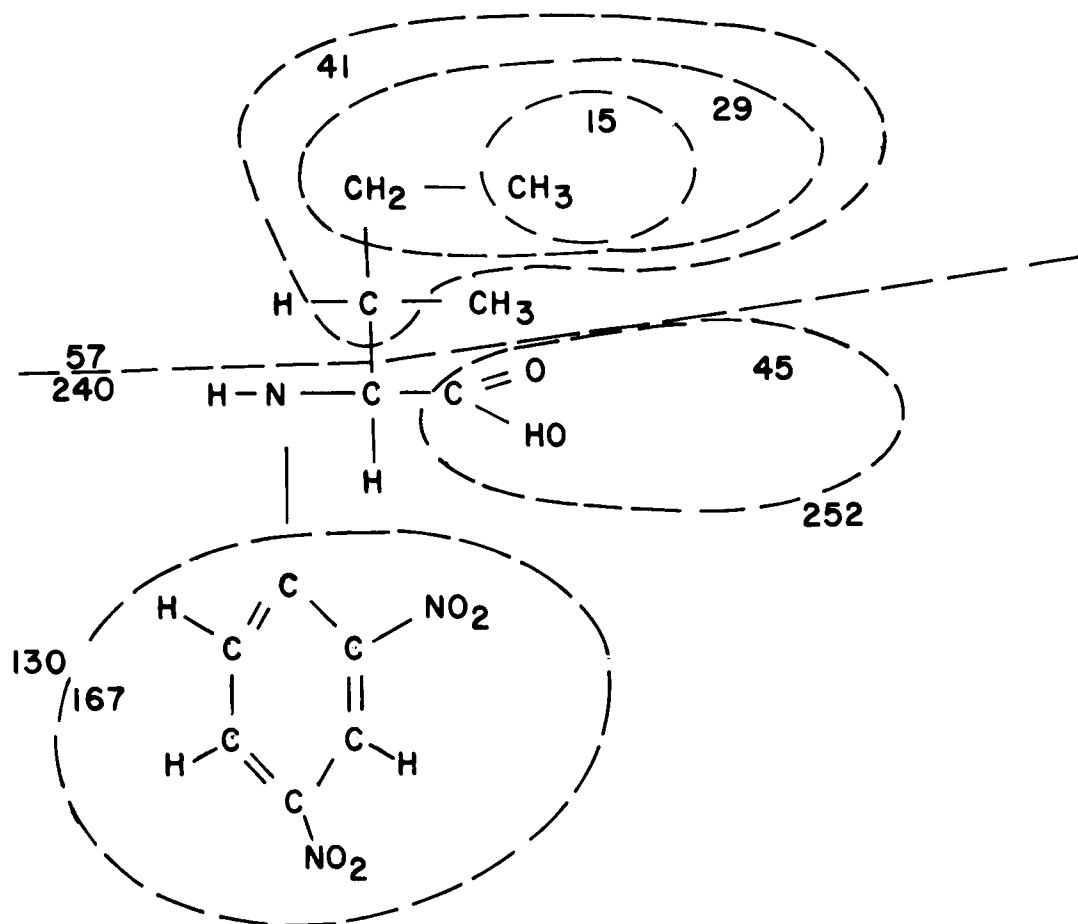


FIGURE 4

Block Diagram of Pulse Circuitry Utilized to Enable Recording of Either One or a Series of Successive Mass Spectra

EMPIRICAL FORMULA OF (DNP)-L-ISOLEUCINE: $C_{12}H_{15}O_6N_3$



GRAPHIC FORMULA OF (DNP)-L-ISOLEUCINE

FIGURE 5

Empirical and Graphic Formulae for (DNP)-L-Isoleucine.
The dashed lines suggest non-rearrangement subgroupings of atoms that might give some of observed mass numbers.

This limitation is for the moment, however, academic. For we have found that the present experimental laser configuration appears to deliver insufficient focused pulsed areal energy density to vaporize many potentially interesting materials. We are now undertaking a re-instrumentation designed to rectify this deficiency.

Our reinstrumentation is intended to accomplish three purposes:

- (1) Reduction of the focused spot size;
- (2) More precise aiming of the laser;
- (3) An increase of the areal energy density delivered at the focus of the laser beam.

It is convenient, for the purpose of discussing our approach, to introduce the concept of brightness defined by the relation

$$B(\vec{r}, \hat{n}, t) = \frac{d^2 P(\vec{r}, \hat{n}, t)}{d\omega_{\hat{n}} dA_{\hat{n}}}, \quad (3)$$

where \vec{r} denotes position in space, \hat{n} a direction at that position, t time, $d\omega_{\hat{n}}$ an element of solid angle about \hat{n} , and $dA_{\hat{n}}$ an element of area orthogonal to \hat{n} at the position \vec{r} .

It is possible to demonstrate, though we shall not do so here, that in the geometrical optics limit the ray brightness in a lossless, homogeneous, isotropic medium is a ray invariant; i.e., the brightness propagating along a ray with the speed of light in the medium is a constant of the motion.

Except for accountable losses arising from reflection, absorption, etc., it follows that the brightness of the laser radiation propagating from the laser to the focus of an introduced converging lens is invariant.

The computation of the power delivered per unit area in the direction of the optical axis of the system and at the vicinity of crossover of

the rays, the "point" of focus, may be deduced from knowledge of the brightness throughout the radiation field on a transverse section directly in front of the rod, specification of the optics, and an appropriate integration utilizing Eq. (3) at the focus. We shall for this discussion make the following simplifying assumptions: 1) the brightness is uniform throughout the cone of directions into which the principle portion of the laser radiation issues from the rod, and 2) correction factors related to the obliquity of rays to the optical axis can be neglected. We thus can conclude from Eq. (1) that the power delivered per unit area at the focus is, neglecting losses,

$$\frac{dP}{dA} \approx B\omega , \quad (4)$$

where B is the laser brightness and ω is the solid angle of convergence of the radiation to the focus.

Expression (4) tells us that in order to optimize delivered areal power density one must secure a bright laser and condense the radiation into the focus through the largest possible solid angle. It is anticipated that what one must optimize is some blend of areal power density and areal energy density delivered in the pulse. The delivered areal energy density can be expressed in terms of a representative pulse brightness B and pulse width τ by

$$\frac{dU}{dA} \approx B\tau\omega . \quad (5)$$

For a given laser brightness, the delivered power density is optimized then by opening up the solid angle into which the radiation is condensed. If it happens that collimated light enters the lens element immediately preceding the focus then in terms of the F-number of the lens, defined by

$$F = f/D, \quad (6)$$

where f is the lens focal length and D is the lens diameter, (4) may be written

$$\frac{dP}{dA} \approx \frac{\pi B}{(2f)^2} . \quad (7)$$

There is a fundamental theoretical limitation on the brightness of a laser, which may be deduced as follows. Simplifying (3) in a manner consistent with the approximation, the laser brightness may be expressed in the form

$$B = \frac{P}{A\omega} , \quad (8)$$

where P is the laser power, ω is the solid angle into which the laser emits its radiation, and A is the cross-sectional area of the rod. In terms of the full angle beam spread θ_f of the laser, the solid angle may be expressed as

$$\omega = \pi(\theta_f/2)^2 . \quad (9)$$

Expressing A in terms of the rod diameter D ,

$$A = \pi(D/2)^2 , \quad (10)$$

we obtain, upon substitution of (9) and (10) into (8),

$$B = \left(\frac{2}{\pi \theta_f D} \right)^2 P . \quad (11)$$

The minimum theoretically attainable value for the full angle θ_f is determined by diffraction limitations to be

$$\theta_f \approx \lambda/D , \quad (12)$$

where λ is the wavelength of the laser radiation. Thus we obtain

$$B < \left(\frac{2}{\pi \lambda} \right)^2 P . \quad (13)$$

The maximum power density in the vicinity of the "point" of focus of the condensed radiation will be limited, referring to (7), therefore to

$$\left. \frac{dP}{dA} \right|_{\text{focus}} \approx \frac{1}{\pi} \frac{P}{(F\lambda)^2} \quad (14)$$

And since in practice

$$F \gg 1, \quad (15)$$

(14) becomes

$$\left. \frac{dP}{dA} \right|_{\text{focus}} \approx \frac{P}{\pi\lambda^2} \quad (16)$$

Expression (16) is, of course, consistent with our recognition of the fact that with light of wavelength λ it is impossible to form a focal spot of size smaller than λ .

In our reinstrumentation we are, with reference to (4), optimizing ω while planning for the moment to use the same laser we have employed for our earlier work. A schematic of the optical system now under construction is given in Figure 6.

The radiation issuing horizontally from the 1/4" diameter ruby rod in the laser 1, is diverged by the simple plano-concave-73 mm focal length lens 2, reflected downward by mirror 3 condensed by an 35 mm focal length F:1.4 Komura "35 mm camera" objective placed ~ 650 mm from lens 2, and directed through a planar glass window to a focus at the target.

The ratio of delivered energy density at the focus achieved with the diverging element 2 to that obtained without the element is ideally

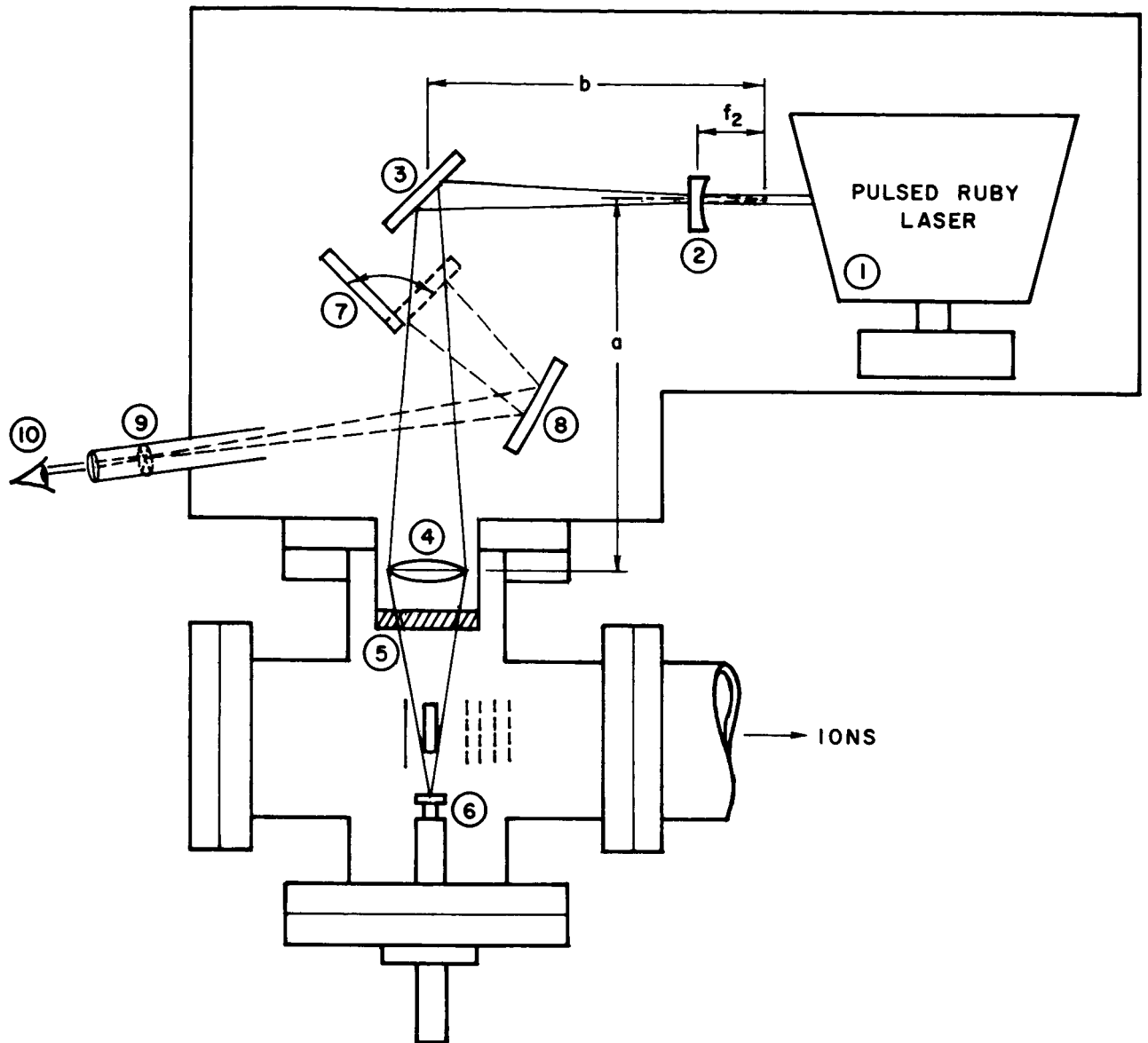


FIGURE 6

Schematic of Laser Optical Configuration
Presently Under Construction

$$\begin{aligned}
\frac{(dU/dA)_{\text{focus with } \underline{2}}}{(dU/dA)_{\text{focus without } \underline{2}}} &\approx \frac{\omega_{\text{focus with } \underline{2}}}{\omega_{\text{focus without } \underline{2}}} & (17) \\
&= (a + b)/f_2 \\
&= \left(\frac{657}{85}\right)^2 \\
&= 60
\end{aligned}$$

It is clear that the ~ 60 -fold increase in energy density has been accompanied by an ~ 8 -fold reduction in the spot size at the focus. Lens losses and instrumental aperture constraints in the MS source will limit us to somewhat less than the 60-fold theoretical enhancement. We have in a laboratory mockup of the system been able to produce 35μ diameter craters in steel with the ~ 10 millijoules transmitted to the target.

We plan with this new system to irradiate various portions of the sample by moving the sample about beneath the focused laser radiation, rather than by moving the point of focus as with the earlier system.

d. Computer Manipulation of Chemical Hypotheses

This work is mainly supported by ARPA contract under the direction of Professor E. Feigenbaum, but Professor Lederberg's part in it is pertinent to the present NASA contract.

A polished form of the DENDRAL program has now been translated and is running smoothly on the PDP-6 LISP system. A recent report giving detailed documentation is included as an appendix to this report, together with an example of well-filtered output of the ISOMERS function.

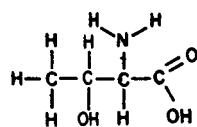
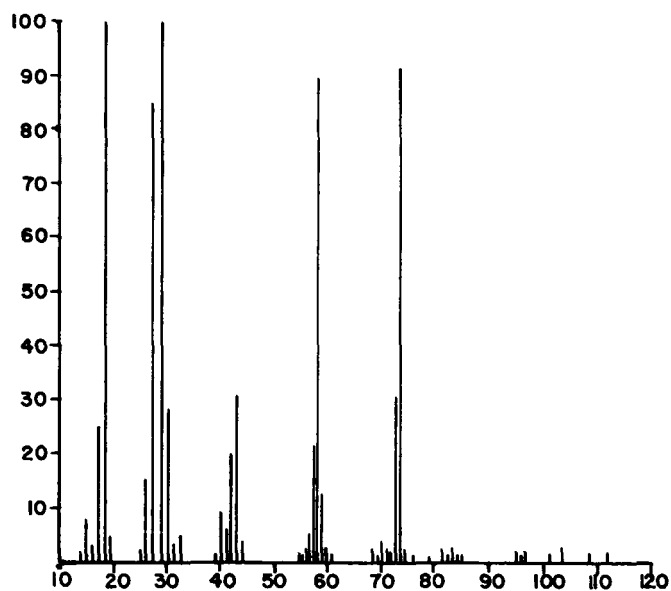


FIGURE 7

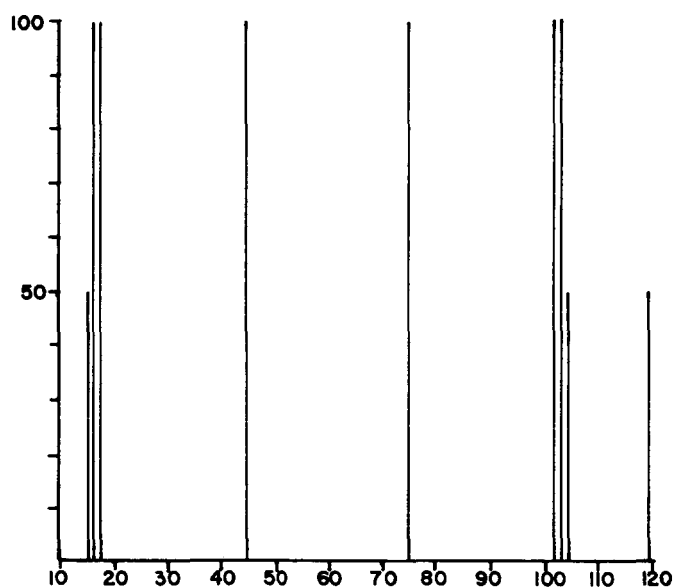


FIGURE 8

Threonine: Idealized Spectrum (Fig. 7), Actual Spectrum* (Fig. 8)

* N. Martin, "An Investigation of the Mass Spectra of Twenty-Two Free Amino Acids," NASA Technical Report No. IRL-1035, Instrumentation Research Laboratory, Genetics Department, Stanford University, Palo Alto, California, p. 26 (1965).

The program will now solve mass spectral problems based on a primitive, approximate theory of molecular fragmentation, taking only a few seconds for the example of threonine. The input data are an idealized spectrum such as would be obtained if each chemical bond had an equal probability of rupturing to generate two fragment ions (Fig.7). A more realistic theory of mass spectrometry is now being programmed to allow real data (Fig. 8) to be interpreted.

IV. Computer Managed Instrumentation

a. ACME Program

As discussed in the last status report, the Instrumentation Laboratory is engaged in a cooperative program with the Advanced Computer for Medical Research Group (ACME). A status report prepared by ACME for the NIH covering this same report period is included as an appendix to this report. This report describes some of the collaborative efforts of both activities.

One such collaborative effort has been the design and construction of a computer driven CRT. This utilized a 4K core memory supplied by IRL which was suitable for a recirculating, computer-driven, display. The techniques are applicable to using PDP-8's or LINC computers as a basis for CRT displays. This is particularly pertinent since there are quite a number of these small computers existing within the Medical School.

Certain requirements of IRL, including mass spectrometry instrumentation, will require a reasonably high speed input to the ACME 360/50. Such a unit has been built by IBM and termed a "270X." One such unit has been delivered and installed as part of the ACME 360/50 system. It has undergone preliminary testing and it appears that it will be quite suitable for the needs of this laboratory.

Some proposed uses of the 270X will require digital plotting at the experiment site. This laboratory has connected their Calcomp digital plotter to the 270X at the ACME site. The ACME programmers have written plot programs to demonstrate the technical use of the 270X to drive this plotter. This will be quite beneficial as it insures that plotting capability will be available at any remote 270X laboratory installation without additional extensive costs for plotter interfaces. The ACME programmers are cooperating in adapting the Calcomp routines to this load of operation. If successful, this will make it much easier for the researcher in his laboratory to program high quality and informative graphical outputs.

b. Mass Spectrometry

The activity in computer instrumentation for mass spectrometers have related to the Bendix TOF, the AEI MS-9 mass spectrometer in the Chemistry Department, and the EAI 300 Quadrupole mass spectrometer.

1. Mass Peak Identification for the Bendix TOF Mass Spectrometer

As reported in part IV, section b. of the previous status report, mass spectra from the Bendix TOF are being sent through a logarithmic amplifier then digitized and stored on magnetic tape by the LINC computer. Also included in the report was the procedure of applying a transformation to the time axis of the spectra in an attempt to place the mass positions at equal intervals on this time axis. Because of certain drifts in the TOF, the mass values of this altered spectrum cannot be accurately determined simply by their position on the new scale. The method described previously for determining their position was quite time consuming with the limited computational abilities of the LINC and will have to wait for the implementation of the ACME system for practical usage.

A method has been developed for identifying integer mass peaks by utilizing the display scope of the LINC computer to display a portion of a spectrum and a raster of the peak positions.

The present method displays a portion of the spectrum (after the rough linearization) on the screen with a generated raster indicating the approximate separation of the mass positions (see Figure 9). The raster can be stretched and translated using the potentiometers on the LINC console to fit the peak positions of the displayed spectrum. This approach combines the user's ability to distinguish meaningful peaks with the computer's capacity to calculate the raster, store the digitized spectrum, and store the information about the mass peaks once the user has adjusted the raster to fit the particular part of the spectrum being displayed.

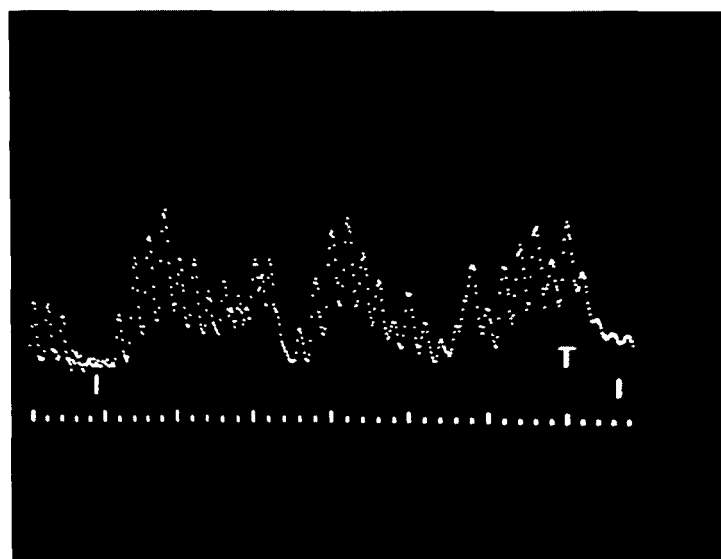


FIGURE 9

A Portion of a Mass Spectrum
as Displayed on the LINC Oscilloscope

The program is operated under the locally written LOSS monitor system (see R. K. Moore, An Operating System for the LINC Computer, NASA Technical Report No. IRL-1038) and uses its conventions for handling the data tape (the spectrum). Upon loading the program, the first portion of the spectrum (actually the transformed spectrum as described above) is read into the memory and displayed on the oscilloscope (8 cm by 8 cm display area). About 40 mass positions are within view of the

user. Initially the user positions an illuminated "pointer" over a known mass peak and enters the associated mass value using potentiometers on the console. Lifting a console switch identifies this position in the spectrum with the particular mass position. At this time a raster is also displayed on the screen with mass positions equal to zero modulo five being enhanced. Thus a raster is developed for the set of peaks based on the particular position and mass number entered through the potentiometers. The distance between the elements of the raster may now be expanded until the elements and the mass peaks coincide. This expansion takes place about the reference mass previously mentioned so as not to disturb its position. The user can, however, translate the entire spectrum to improve the matching between the raster and the observed mass peaks by adjusting another potentiometer.

With the matching accomplished the researcher can now either have the amplitudes at the mass positions typed out (still in logarithm form) or, by setting a switch recorded on magnetic tape for further processing (see below). The mass positions considered in this operation are those between a set of vertical bars on the screen which act as parentheses around the masses under consideration. The user has complete flexibility in determining the portion of the displayed peaks to type or record or may investigate an area more than once to observe changes due to altering the raster. In general, the latter will yield little change since the program searches half a peak on either side of the raster position to find the maximum in the signal.

Having investigated the displayed portion of the spectrum the user moves the reference mass up to a position in the rightmost quarter of the screen by moving the pointer to a mass position on the raster and lifting a toggle switch. Lifting another switch will move the data on the rightmost quarter of the screen to the extreme left and read in the next portion of the spectrum. The raster can again be adjusted and another set of mass values typed or recorded.

In actual practice it has been found that only a slight amount of raster adjustment is needed (less than a peak width) in any one portion of the spectrum and thus the user can move along quite rapidly. This latter condition prompts one to think about automatic adjustment of the scale using an algorithm which examines the spectrum to find the spacing for each successive portion. While this would work well in the lower part of the spectrum (say up to mass 75), the large voids in the higher mass regions make these methods undependable. Once familiar with the system, a user can cover the mass positions up to 300 in three or four minutes if the output is being stored on tape rather than typed.

Once the data is on magnetic tape in the form of mass number and amplitude number pairs, it is readily usable in a number of ways. It can be put on standard seven channel digital tape or possibly sent directly to another computer as intended under the ACME concept. Presently the data is used as input to another LINC program which takes the antilog of the mass amplitudes and produces a bar graph (see Fig.10) in which the largest peak is considered to have value 100 and all other peaks are scaled accordingly.

The output, peak identification and bar graph is much the same as that reported previously, however, from a considerably different approach. The method described here is basically simpler and requires less computer capacity but requires much more machine-operator interaction. Since there is much to be said for both approaches, depending on the computer facility available, it is felt that pursuing both methods is well worthwhile.

2. MS-9 Fast Scan Data Acquisition

This laboratory is cooperating with the Chemistry Department of Stanford to instrument the AEI MS-9 mass spectrometer in a fast scan, on-line system.

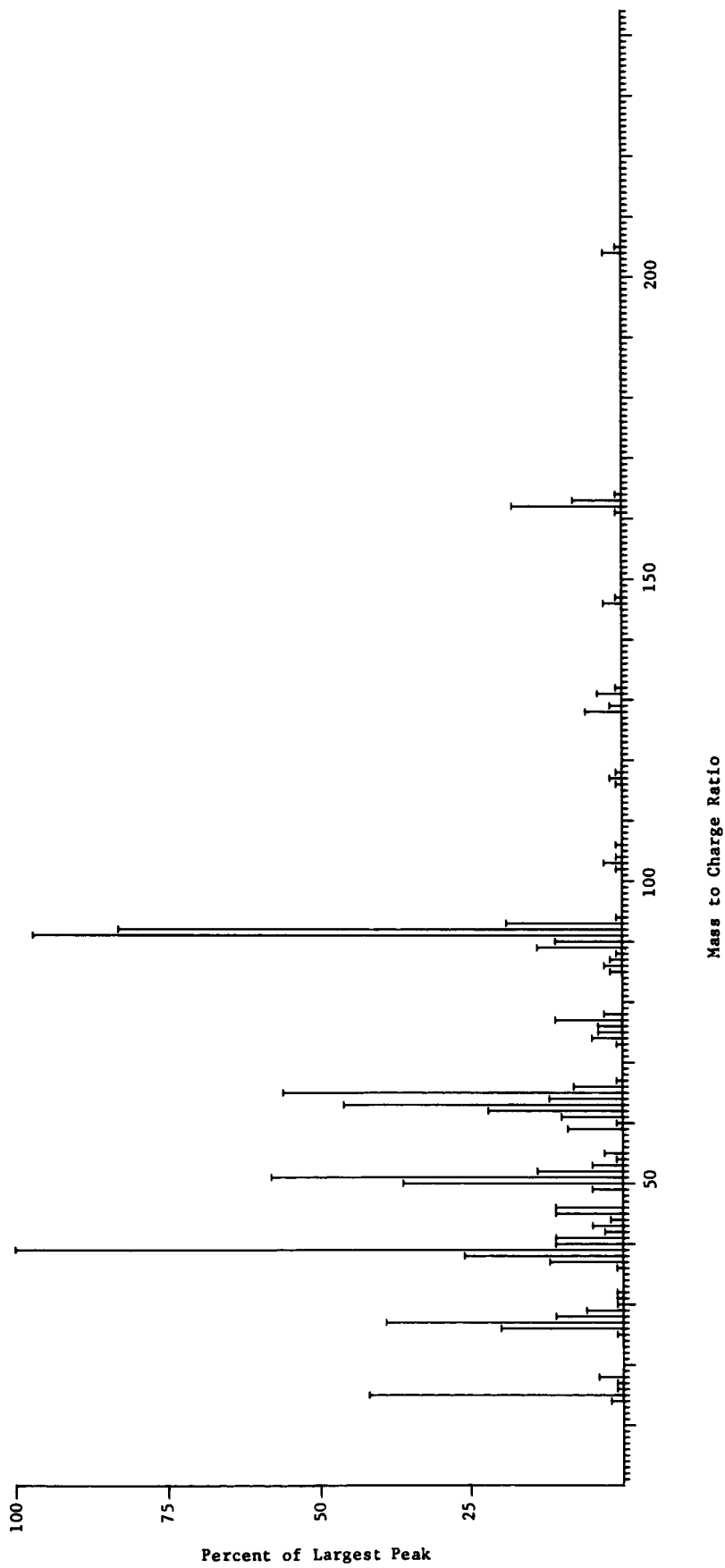


Figure 10
Mass Spectrum of Phenyl Alanine Methyl Ester HCL

The MS-9 is a high resolution mass spectrometer. Though in its initial configuration was a rather slow instrument with regard to gathering data, the high resolution ability of this instrument has made it of great interest to our department's work. As early as two years ago, the manufacturer of this instrument announced plans to provide a high speed capability utilizing magnetic tape. Actual ability of the manufacturer to deliver matured very slowly and only in the last few months has this ability been evident. In the meantime, assessments by our group indicated that the technique employed, although useful, was limited by the state of the art in analog magnetic recordings. It was concluded that much was to be gained, at very little extra cost, by direct computer connection. This would put the researcher in much better control of his experiment and enable him to more effectively direct the course of the experiment or data analysis.

The availability of ACME, the previously mentioned 270X remote digital connection to the 360/50, and the digital plotting capability are important to this project.

General technical planning for this MS-9 project was completed early in this reporting period. An occasion arose in February for one of our engineering staff to visit the manufacturer of the MS-9. At that time information was gathered enabling the final decision on which of the manufacturer's modification would be used and which would be developed by this laboratory. The time schedule for this MS-9 project calls for completion to the point of hardware and program debugging by the end of July 1967 and an operational system by November 1967. Upon completion of this project, full details will be reported.

3. Instrumentation of the EAI Quadrupole Mass Spectrometer

During this reporting period a decision was made to develop a computer interface for the EAI (Electronic Associates Inc.) QUAD 300 Quadrupole mass spectrometer. This mass spectrometer is connected to analyze the

effluent from a Varian aerograph gas chromatograph in connection with Pasteur Probe experiments.

The EAI Quadrupole mass spectrometer and the Bendix TOF mass spectrometer, though completely different in principles of operation, do have similarities in control of mass sampling position. With fairly minor changes to the interface, it would be suitable for either mass spectrometer. With suitable minor alterations it can operate off of either the LINC computer or the ACME system.

It was decided to initially connect the EAI mass spectrometer, via this interface to the LINC, because of availability of remote teletype control. This mass spectrometer interface will eventually be transferred to the ACME system. ACME will provide an additional increase in flexibility and usefulness since the ACME system features increased ease of programming and data management.

Hardware for the interface was 90 percent complete by the end of the reporting period. Connection had been made to the LINC computer and the EAI Quadrupole. Testing of the hardware and software has begun and it is expected to be operational by the end of April.

Previous computer instrumentation efforts in mass spectrometry has been mainly to develop data logging systems. The system, referred to here, includes computer control of many of the mass spectrometer functions. For example, the spectra will not be scanned in the conventional mode. In this system, the computer will direct the mass spectrometer to sample the spectrum at a specific known peak. This can be done in any directed order. For example, if desired, the computer program can make a decision which peak or peaks should be measured at a given phase of the experiment. This decision could be based upon previous acquired data and the logic incorporated into the program.

The total system also includes several self-calibration procedures. These have been designed to automate, or at least to simplify, the numerous adjustments that are normally necessary to insure acquisition of high quality spectral information.

c. Particle or Cell Separator

A cell separator based on measurement of electrical resistance changes during the flow of the particles through a small chamber was recently described by Fulwyler³. The basic difference between this type of apparatus and commercially available particle counters using the resistance principle is that the liquid passed through the detection orifice is further forced into a stream. As shown by Rayleigh in 1879⁴, such a cylinder of liquid is dynamically unstable under the action of surface tension, and it can be broken into uniform and equally spaced droplets by launching into it an ultrasonic wave. The size of the droplets is determined by the inside diameter of the orifice and the wavelength of the wave introduced in the column.

By sensing resistance changes at the cavity formed by the orifice the volume of each non-conducting particle passing through is detected. Each droplet is charged to a potential related directly to the volume of the particle it contains (if any). After deflection while dropping through a uniform electric field, the droplets are collected into various vessels.

This type of apparatus should be valuable in many biological investigations and a modification of it has been built here. We have also started to investigate the possibility of replacing the resistance

³M. A. Fulwyler, Science, 150, 910 (1965).

⁴Lord Rayleigh, Proc. London Math. Soc., 10, 4 (1879).

detector with an optical one so as to be able to separate particles according to their optical characteristics under various types of illumination (see Section 7).

The particle separator consists of the following basic parts:

1. Reservoir, orifice and detector
2. Size discriminator
3. Droplet forming apparatus
4. Charging cylinder and deflection plates
5. Collection cups

Figure 11 shows the entire apparatus broken down into the basic parts mentioned above.

1. Orifice and Detector

The particles, suspended in a conductive media, are stored in a reservoir. This solution under pressure is forced through the orifice to form a liquid column. The orifice structure constructed here is made by laminating a glass or plastic disc about 250 micron thick having a central hole of about 100 microns with a platinum disc of about 400 microns having a hole, concentric to that of the insulating disc, of 75 microns in diameter. This metal disc is used as the ground electrode. This structure forms one side of the detector resistance bridge. The 250 microns long by 100 micron diameter cavity of the insulating disc determines the steady-state resistance of the orifice structure. Any non-conducting particle entering this cavity will alter the cavity resistance according to the volume of the particle. This volumetric modulation of the orifice resistance by particles entering the cavity is sensed and amplified by the detector amplifier. Since two or more particles in the cavity simultaneously would give a signal comparable to that of a single large particle the solution should be dilute enough to minimize such coincidence.

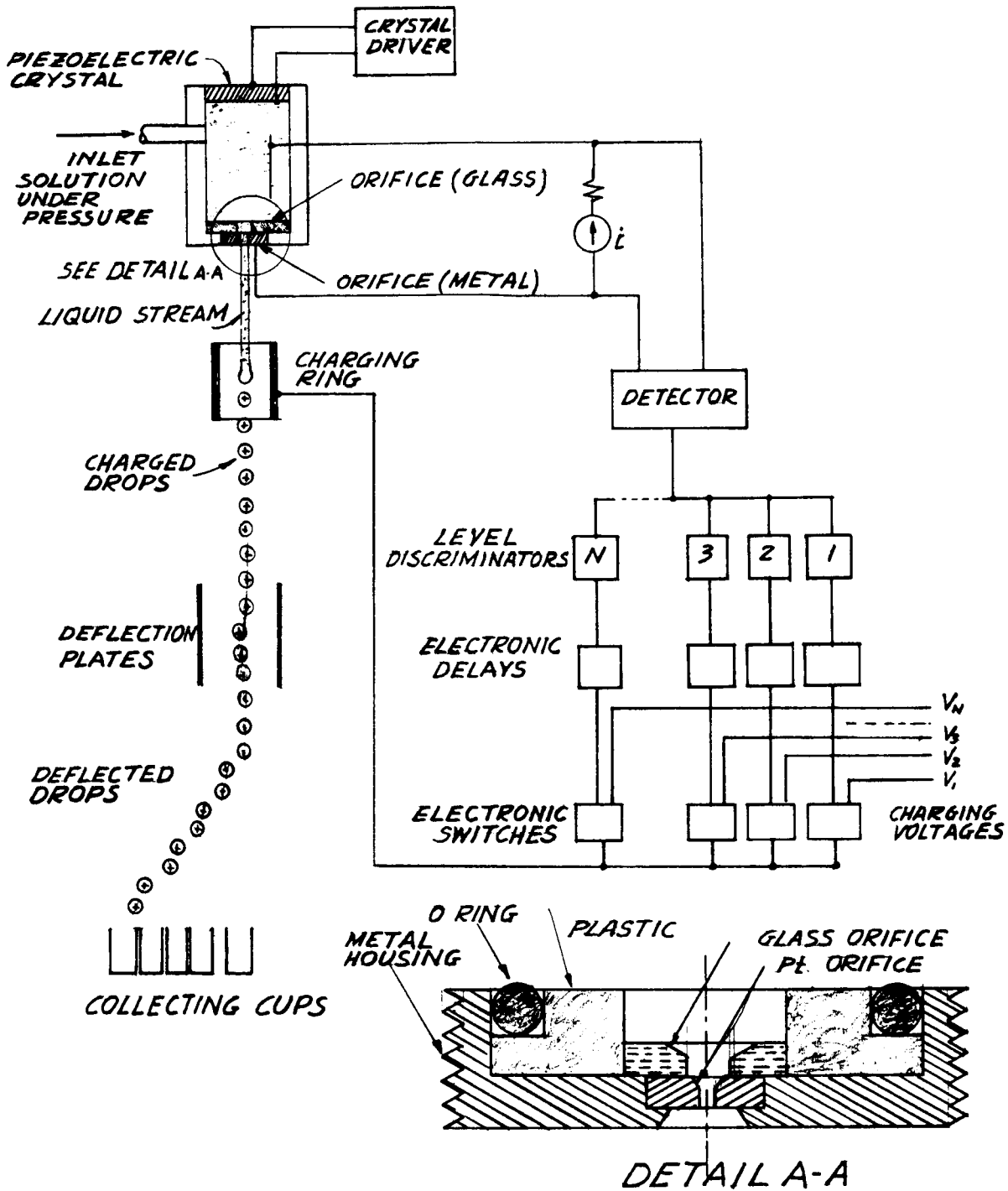


FIGURE 11

If a non-conducting particle of length ℓ_1 and diameter d_1 enters the sensing cavity the resistance change is given by

$$\Delta R = \frac{4\rho\ell_1}{\pi} \left[\frac{d_1^2}{(d^2 - d_1^2)d^2} \right] \quad (18)$$

where

- ρ - resistivity of solution
- d - diameter of insulating cavity.

The voltage change due to the presence of a particle in the cavity is

$$\Delta V = i\Delta R$$

where i is the current through the cavity.

The amplified voltage is thus a function of the size of the particle in the orifice cavity at any given time. This amplified signal is fed into the discriminator.

2. Size Discriminator

The size discriminator consists of an array of level detectors. The number of discriminators used depends on the number of size separations desired with an upper useful limit depending on the signal-to-noise ratio.

Each discriminator is a level detector demultiplexer that separates pulses of a given size range. It is hoped that with the present signal-to-noise ratio this device will be able to discriminate particles of 5 micron diameter and larger with a one micron resolution.

3. Droplet Forming Apparatus

The emerging liquid stream is dynamically unstable and will break into droplets. However, for any useful utilization of the droplets these must be of equal size and uniformly spaced. They should also break away from the liquid column at a precise distance away from the orifice.

Uniform size droplets can be produced by ultrasonically exciting the emerging column. A piezoelectric crystal of a resonant frequency of 80 KC is excited and its acoustical wave excites the emerging liquid column by means of a waveguide, tuned to the exciting frequency.

Rayleigh⁵ showed that the "wavelength" for which the droplet formation occurs at a minimum time from the stream formation is given by

$$v/f = \lambda \approx 9a \quad (19)$$

where a is the radius of the jet, v is the velocity of the jet stream, and f is the exciting frequency.

In our case the jet diameter is about .8 of 75μ or 60μ ⁶. The exciting frequency is 80,000 cycles per second and the velocity is between 16 and 35 meters per second at reservoir pressures in the range between 10 and 50 psi. This velocity range gives a λ range of $7a$ to $14a$. This λ falls within the empirical limits of a required for uniform droplet formation. From the principle of conservation of volume per unit length of the jet the size of the droplet will be $\pi a^2 \lambda$. The number of droplets per unit time is determined by the exciting frequency. Thus $\frac{4}{3} \pi r^3 = \pi a^2 \lambda$ where r is the radius of the droplet and $\lambda = \frac{v}{n} = v/f$ and n the number of droplets per second.

⁵Lord Rayleigh, Proc. Roy. Soc., (London), 29, 71 (1879).

⁶N. R. Lundblad and J. M. Schneider, Rev. Sci. Inst., 42, 635 (1965).

Thus

$$r = \frac{3}{4} a^2 \lambda$$

In order to meet Rayleigh's criteria

$$v = \lambda f = 9 a f = 9 \times 30 \times 10^{-6} \times 80 \times 10^3 = 22 \text{ m/sec.}$$

This velocity is in the range observed in the apparatus. Thus with the present orifice and exciting ultrasonic frequency the conditions can be made optimum for uniform droplet formation. In fact the drops observed experimentally seem to be essentially uniform. We have ordered lower frequency transducers in order to measure the effect of varying the frequency. This lower frequency will result in larger size droplets. However, the droplets will still be of smaller volume than the orifice chamber, so no reduction in particle concentration should be required.

4. Charging Cylinder and Deflection Plates

The charging cylinder is placed so that it contains the point where the emerging liquid jet is broken into droplets. The voltage applied to the cylinder is varied according to information extracted at the orifice with an appropriate delay. The jet passes through the uniform electric field of the cylinder and acquires a charge proportional to the voltage applied. The drops separate from the jet and pass through the uniform field of the deflecting plates and are deflected by an amount proportional to their charge. The equation correlating the various parameters of the charging and deflecting system is:

$$k = \frac{4\epsilon_o \ell V_{12} V_d}{s r_1^2 \ln \frac{r_2}{r_1} 7 \times 10^3 p} \quad (20)$$

where

- k = ratio of horizontal to vertical velocities of the droplets
- ϵ_0 = permittivity
- l = length of deflection plates (m)
- s = separation of deflection plates (m)
- p = upstream pressure (lb/in²)
- r_2 = radius of charging ring (m)
- r_1 = radius of droplet (m)
- V_{12} = charging potential (volts)
- V_d = deflection potential (volts)

Since V_{12} and V_d enter as a product in the above equation, it is obvious that either one could be changed for separation purposes. However, if the droplets are charged uniformly, a much longer delay is required between particle detection and application of the appropriate deflection voltage. It is advisable to keep this delay to a minimum for high speed operation.

Certain physical limitations are imposed in selecting values for the various parameters in the above equation. The distance s is limited due to electrical discharge from V_d . The length l is also limited to avoid the deflected droplets hitting the plates. The maximum charging potential is limited by electronic component limitations while r_2 is limited by alignment requirements. Deflections of the order of 0.4 in. at 4 in. below the deflection plates are observed with droplets charged to 300 volts at the charging ring and deflected by 10,000 volts applied to the deflection plates. This appears to be adequate.

5. Collecting Cups

The collecting cups are a series of vessels spaced approximately 1/8-inch apart and connected to bigger vessels by means of tubing.

6. Status of Cell Separator

All the mechanical parts of the system have been completed. The electronic equipment required has also been designed and built.

Some preliminary experimentation with the apparatus has indicated the possibility of separating particles down to 5 microns in diameter using the parameters listed above. The signal-to-noise ratio of the system is such as to permit a 1 micron resolution. The most critical portion of the entire apparatus seems to be the orifice. Clogging of the orifice has presented some problems. Due to the fact that electrolysis occurs at the orifice and the intermediate reservoir, bubbles generated in this area tend to affect the direction of the jetstream. This is objectionable for continuous operation. We are in the process at this time of fabricating a new orifice structure incorporating details forwarded to us by Mr. Fulwyler. These details concern the mechanical configuration of the platinum orifice and polishing of the platinum electrode so as to minimize the generation of gas bubbles.

7. Detection of Fluorescent Single Cells in Motion

Fluorescence techniques are very valuable for detecting the presence of small quantities of material. Experiments are under way to determine the feasibility of detecting fluorescent single cells, and separating such cells from non-fluorescent, or weakly fluorescent material, using the particle separation techniques discussed above. In such an instrument the cells would flow past an optical assembly containing a light source, appropriate filters, and a photodetector. Non-uniformity of cell size, variations in flow rate, inhomogeneous distribution, etc., would lead to varying pulse lengths and amplitudes at the detector. Pulse amplitude would be used to trigger the separation circuitry. The experimental equipment for evaluation of this technique takes the form of a microscope with the various components fitted to an optical support rail. The light source is imaged after primary filtration on the surface of a rotating chopper disc. Two diametrically opposed holes of 2.87 mm dia.

drilled on a 5.08 cm radius of the disc provide two light pulses per revolution. The disc is rotated at 30 revolutions per second by a synchronous motor. The modulated output of the chopper system consists of pulses 300 μ seconds in length occurring at a rate of 60 cycles per second. A pair of microscope objectives are used to image the source and collect the modulated light. The modulated beam is then directed to a microscope stage via a dark field condenser. Comparison of fluorescent and non-fluorescent cells is made by examining suitable areas of a test slide on the stage. The magnified microscopic image falls upon an aperture plate after secondary filtration. The purpose of the aperture plate is to limit the size of the field. Aperture size may easily be changed. Detection is finally carried out by a 1P21 multiplier phototube whose amplified output will be used for subsequent operations. The overall system simulates a microscope observing cells flowing in a capillary system. After optimization of parameters a suitable flow system will replace the conventional cell slide, and the detector output will be used to control cell sorting procedures. The initial design of the experimental apparatus is complete and parts are under construction.

d. Investigation into Separation of Specific Fractions of DNA

As discussed in the last progress report, it appeared worthwhile to investigate whether specific DNA fractions could be separated by elution from hydroxyapatite columns. Such columns were reported by Miyazama and Thomas⁷ to retain native DNA when eluted with 0.08 M phosphate buffer, while permitting thermally denatured DNA to pass through. They found that fractions containing DNA of differing G-C content could be separated by eluting with 0.08 M phosphate at successively higher temperatures and termed such experiments thermal chromatography. It seemed probable that short stranded DNA containing particular genetic markers would elute over the narrow temperature range associated with complete strand separation⁸.

⁷Y. Miyazama and C. A. Thomas, J. Mol. Biol., 11, 223 (1965).

⁸W. R. Guild, J. Mol. Biol., 6, 214 (1962).

Preliminary experiments using commercially available salmon sperm DNA showed that the hydroxyapatite columns were able to distinguish between fully native and partially denatured DNA. From 75 to 90% of this DNA was eluted at relatively low (<0.1M) phosphate concentrations from columns maintained at 60°C. The thermal melting curve of the original sample was quite broad with a tail extending almost to room temperature, indicating relatively high denaturation. On the other hand, the thermal melting curves of the fraction eluting at high phosphate concentration (>.10M) were quite sharp with half of the DNA melting over a range of $\pm 2^\circ\text{C}$ around the midpoint of the thermal transition at about 85°C (Fig. 12, Curve A). The thermal chromatograms showed much of the material eluting at lower temperatures than the melting curve, with the actual elution temperature varying somewhat depending on the condition of the column. As shown in Figure 12, on freshly prepared columns the elution temperature (Curve B) was only slightly lower than the melting temperature. However, on one column in use for several weeks the elution temperatures were much lower (Curve C). A thermal melting curve on an early fraction from this column is shown in Figure 12, Curve D. The eluted fraction was essentially undenatured. Similar fractions rerun through the column eluted almost completely in 0.1 M buffer at 60°C indicating that they were permanently changed. Thus the column appeared to provide a very sensitive indication of minor changes in native DNA.

Similar performance was exhibited on this column by phage λ DNA although the elution temperatures were only a few degrees below the corresponding melting temperatures. Only about 15% of the absorbance in the original sample eluted at 60°C in the 0.1 M buffer, indicating that this DNA was relatively undenatured. When freshly prepared columns were used with this DNA the elution temperatures in the thermal chromatograms (Figure 13, Curve A) were considerably higher than the melting temperatures (Curve B).

FIGURE 12
 Thermal Chromatograms and Thermal Melting Curves
 of Salmon Sperm DNA

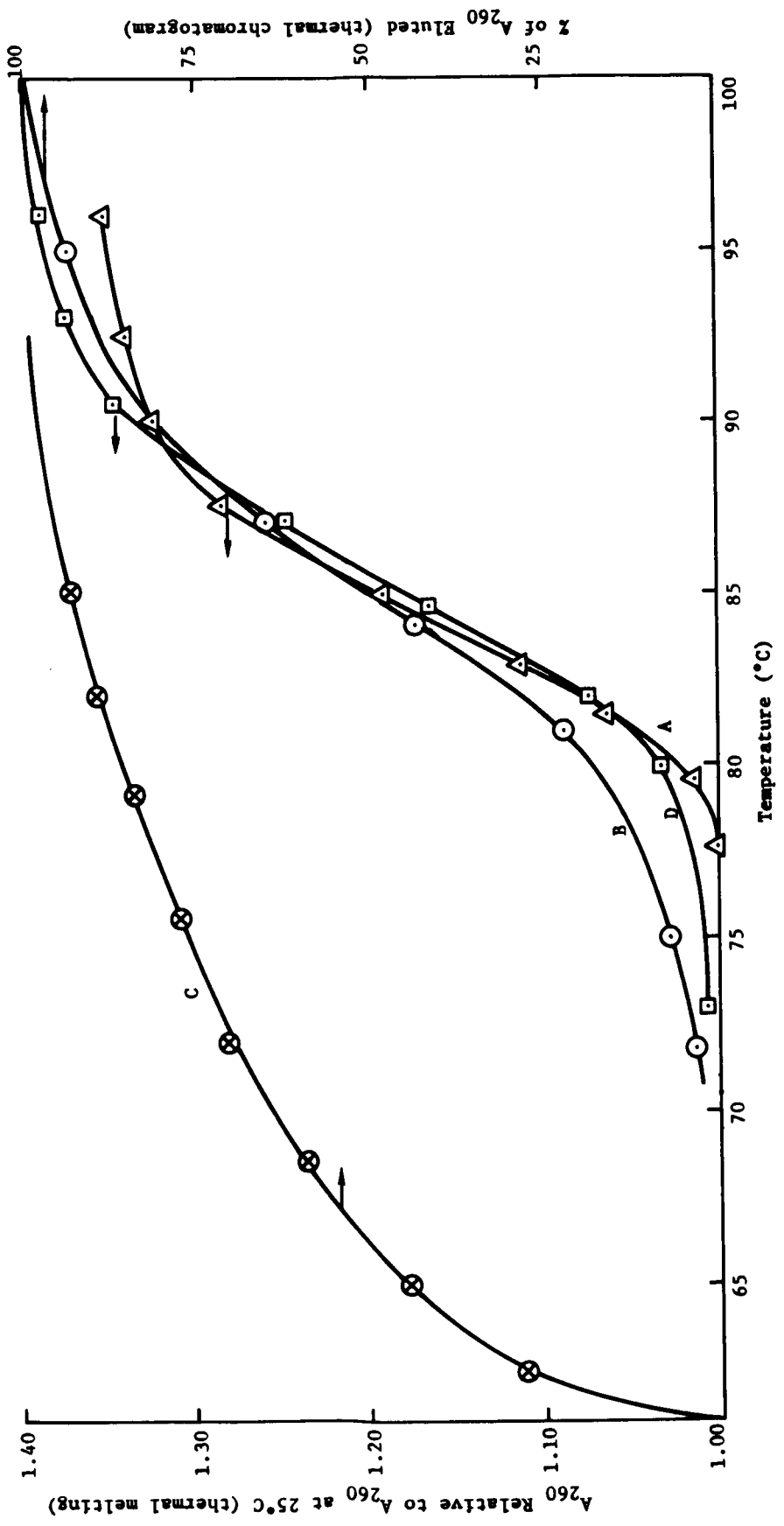


FIGURE 13
 Thermal Chromatography and Thermal Melting Curve
 of Phage λ DNA

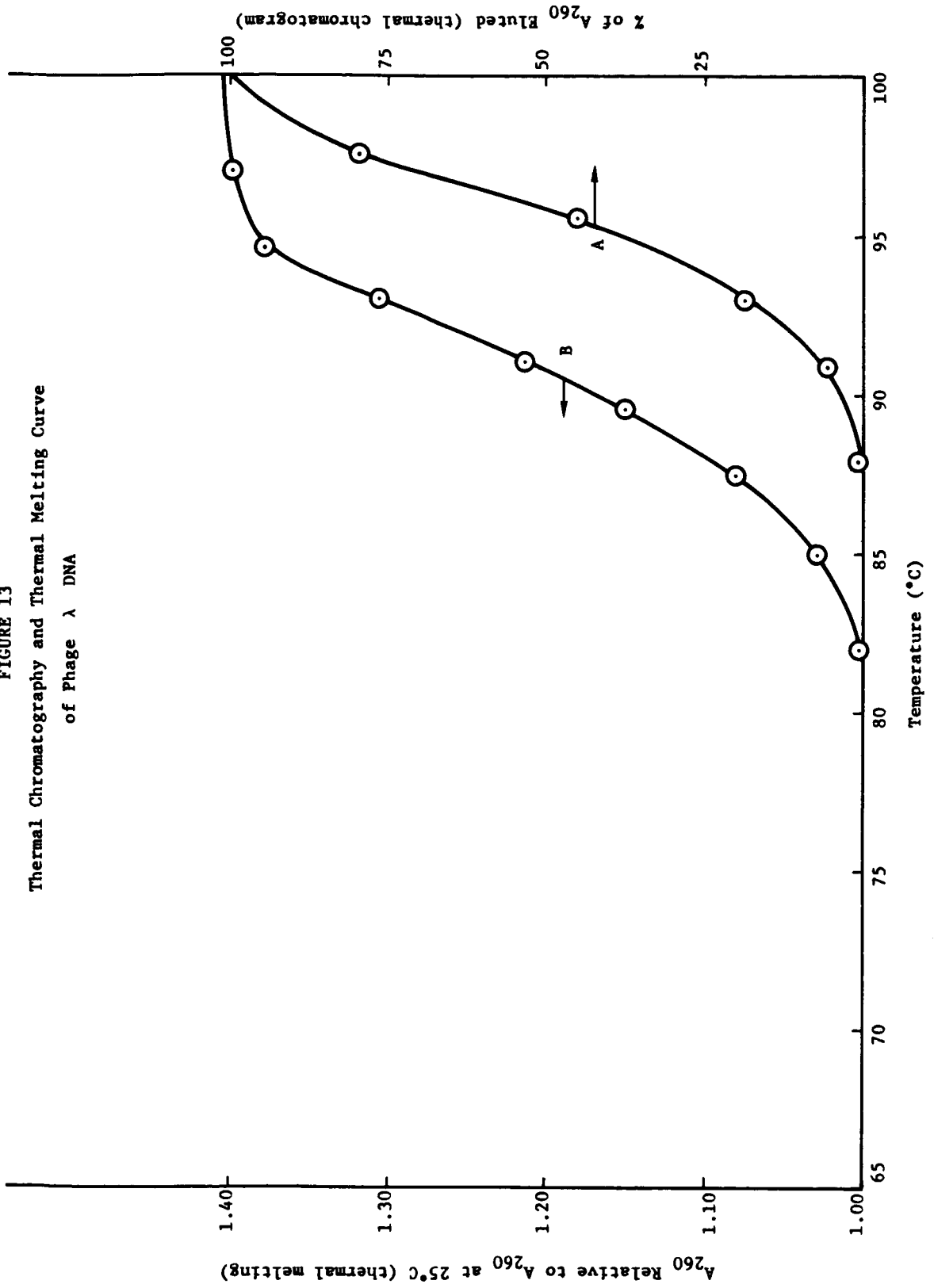
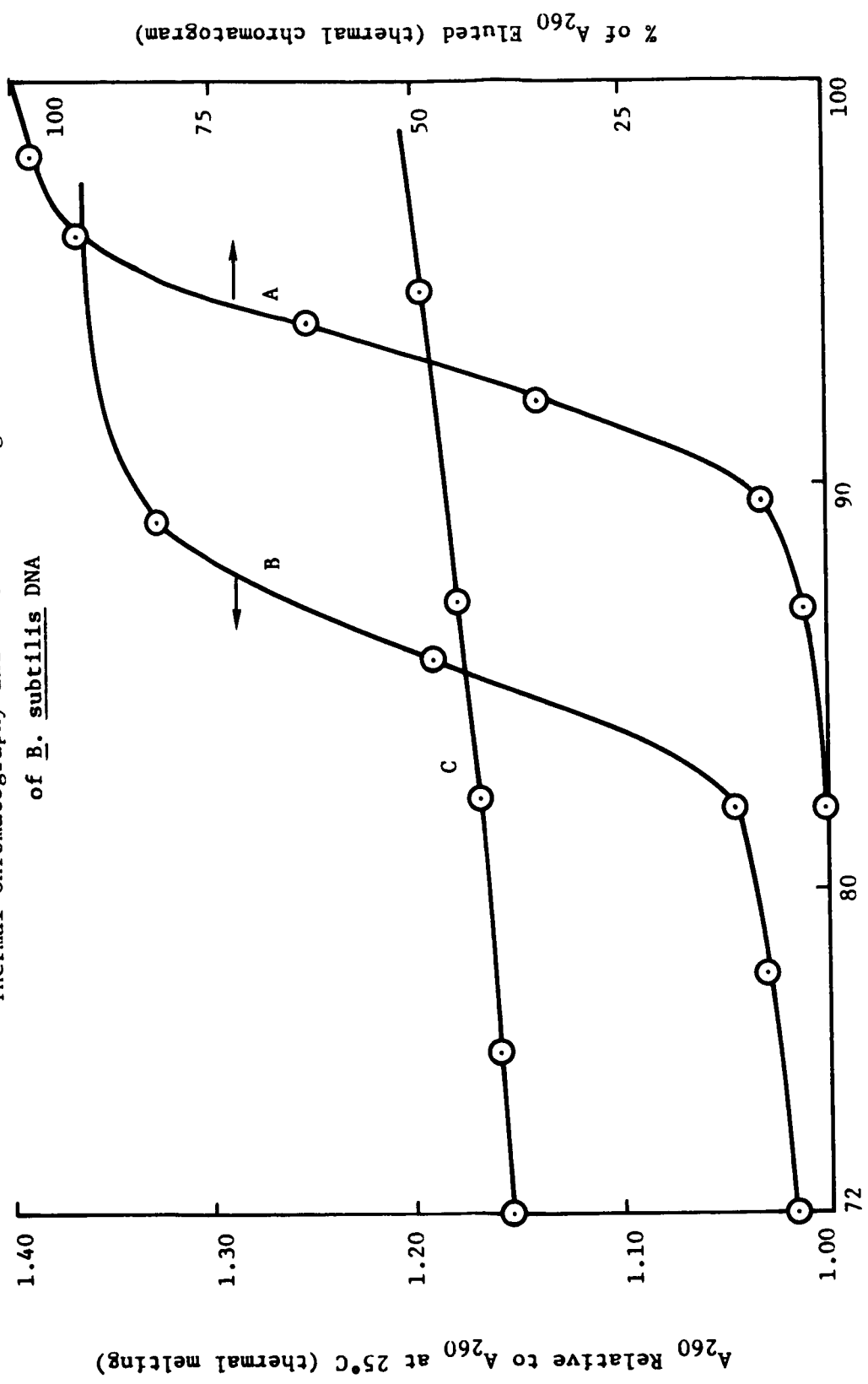


FIGURE 14
 Thermal Chromatography and Thermal Melting Curves
 of B. subtilis DNA



About 25% of the absorbance of a sample of B. subtilis DNA was eluted from freshly prepared columns at 60°C in 0.1 M phosphate buffer. Melting curves on the eluted fractions showed little or no hyperchromicity indicating that some impurities were present in the original material. The thermal chromatogram temperatures (Figure 14, Curve A) were again much higher than the normal melting temperatures shown in Curve B. The thermal chromatogram fractions were essentially completely denatured as shown by the absence of a distinct melting temperature in the melting curve of one of these fractions (Curve C).

Apparently neither phage λ or B. subtilis DNA are eluted on freshly prepared columns until the strands are completely separated as postulated. However, the shorter stranded salmon sperm DNA elutes at lower temperatures. The behavior of short strands of B. subtilis DNA will be determined in future work.

e. Temperature Gradient Control

In order to provide accurate temperature control for experiments such as thermal melting of DNA, work is currently proceeding on instrumentation to detect temperature variations of 0.001°C occurring in a small volume of liquid, and to effect a programmed control of the temperature rise of the liquid with very high accuracy.

A Fenwal Electronics bead thermistor type GB 38 P12 is used as the sensing element which forms one arm of a balanced AC bridge. In order to reduce self heating of the thermistor to an acceptable level, the bridge is driven by a 600 millivolt 400 cycle signal. After amplification of 1000X by a Fairchild A006 operational amplifier the signal is synchronously demodulated in a full wave bridge circuit. The filtered DC level which appears as a result of bridge imbalance will be processed by the ACME computer and used to control the rotation of a stepping motor. This motor in turn restores the input bridge circuit to balance by driving a precision potentiometer in the appropriate arm of the bridge.

The required temperature gradient program will be stored in the computer; when the magnitude and phase of bridge imbalance indicates a departure from this gradient correctional temperature increments will be applied to the liquid.

The characteristics of the thermistor, input bridge, amplifier and output circuit have been evaluated and recorded in several test periods ranging from four to eight hours. Results indicate that a temperature change of 0.001°C can be reliably detected and a usable signal derived for operational purposes. Approximately one second is required for the measurement period but this could readily be shortened under computer control permitting a rapid temperature step or increased gradient at reduced accuracy. No firm developmental approach has been decided upon for the temperature control elements pending a determination of the experimental operability of the detector. Two controlling procedures will be necessary; one applied at the main oil bath, the other in the immediate area of interest. It appears that programmed pulsed control of the oil bath with proportional control at the liquid container should permit control of the required order. Either a Peltier junction device, a small IR filtered lamp or a combination of both could be used to effect final temperature control liquid.

V. Atmospheric Effects on Photographic Resolution

One of the limitations on photographic resolution of planet surfaces is set by atmospheric refractive index variations caused by turbulence. This topic has been of previous interest to one of our staff members. The work has been expanded here to include analysis of the effects of the atmospheres of both the earth and Mars. It appears that ultimate resolution limits on observations of the earth from very high altitudes are in the range of a few centimeters, depending on the exact definition of resolution. The limits would be much smaller on Mars, and thus atmospheric refractive index variation effects will be negligible on any currently considered photographic reconnaissance missions of the Martian surface. A paper describing this work is in preparation.

C. PERSONNEL AND ORGANIZATION

There is one addition to the professional staff of our organization and her curriculum vitae is listed below.

VIRGINIA A. CLOSE

PERSONAL DATA: Born Danbury, Connecticut, April 6, 1942. Single, American citizen.

EDUCATION:

1965 A.B., Northeastern University, Boston, Massachusetts
Major: Biology

1966 Bridgewater State College, Bridgewater, Mass. and
Boston State College, Boston, Massachusetts
15 graduate credits in Education

APPOINTMENTS:

1961-65 Undergraduate Cooperative Work Experience as
Research Assistant

Cancer Research Laboratory
New England Center Hospital
Boston, Massachusetts

Pediatric Clinical Center
Boston City Hospital
Boston, Massachusetts

Walker Biochemistry Laboratory
Massachusetts Eye and Ear Infirmary
Boston, Massachusetts

1966 Teacher of 9th Grade General Science and
10th Grade Biology
Walpole High School
Walpole, Massachusetts

VIRGINIA A. CLOSE

PUBLICATIONS:

"Identification by L-phenylalanine Inhibition of Intestinal Alkaline Phosphatase Components Separated by Starch Gel Electrophoresis of Serum," J. H. Kreisher, V. A. Close and W. H. Fishman, Clinical Chemica Acta, 11, 122-7 (1965).

"Aminopeptidase Profiles of Various Bacteria," J. W. Westley, P. J. Anderson, V. A. Close, B. Halpern and E. M. Lederberg, J. Appl. Microbiol., 1967 (in press).

D. PAPERS AND REPORTS

October 1, 1966 to April 1, 1967

Publications

1. B. Halpern, L. Chew, J. W. Westley, "Investigation of Racemization during Peptide Bond Formation by GLC of Diastereoisomeric t-BOC-Amino Acid Amides," J. Anal. Chem., 39, 399 (1967), IRL-1053.
2. B. Halpern, J. W. Westley, B. Weinstein, "Chemical Shift of a Magnetic Non-Equivalent Isopropyl Group Due to Steric Hindrance," Chem. Comm., 160 (1967), IRL-1057.
3. B. Halpern and J. W. Westley, "Correlation of Absolute Configuration of α -Alkylphenylacetic Acids by GLC," Chem. Comm., 237 (1967), IRL-1058.
4. P. J. Anderson, V. A. Close, B. Halpern, E. J. Lederberg, J. Westley, "The Identification of Bacteria via Their Aminopeptidase Specificity," J. Appl. Microbiol., (1967), in press, IRL-1059.

APPENDIX A

ACME Progress Report to NIH

October 1, 1966 - April 1, 1967

DESCRIPTION OF RESEARCH PROJECT

During the report period nearly all of the available computer facilities were devoted to the development, programming, and check-out of the ACME computer system.

The staff of the project was occupied with a number of tasks which will be reported in separate paragraphs. These will cover:

- 1) System design and implementation
- 2) Program translator
- 3) Data acquisition and distribution programming
- 4) Improvements in terminal man/machine interaction
- 5) Data transmission tested design
- 6) Production and installation of basic transmission units
- 7) Design of a CRT display unit for medical data
- 8) Development of a basic statistical and text processing library for interactive use
- 9) Consultation with medical staff and faculty
- 10) Education for medical staff and faculty
- 11) Installation of the computer equipment
- 12) Operation of the computer facility.

The progress in most areas has been extremely satisfactory. In the short time between receipt and check-out of the basic IBM equipment (December 15, 1966) and this date (April 15, 1967), we have been able to transfer our systems work from the variety of computers used for development and to begin offering a very limited but true timesharing service to the medical school. The education program has led to a great deal of enthusiasm on the part of the medical staff and faculty.

The greatest lack as of this date is the delay in file capability. We can expect to have this capability available by June 1, 1967.

During this project much assistance and exchange has been effected between ACME and the Central Campus Facility at Stanford. With the current delays in software delivery of IBM's timesharing system the ACME System will be used at the Central Facility.

Continued exchange is going on between ACME and the Stanford Linear Accelerator Center in the areas of high speed data transmission and graphic support and between ACME and the Allen-Babcock Company regarding terminal control and systems development. In addition, there has been a joint educational and informational effort in the 1800 computer area with the Syntex Laboratories of Palo Alto.

Throughout the ensuing paragraphs references are made to the working papers of the ACME Project, or ACME Notes, which are appended to this report as Appendix I.

Interest in the ACME System has been wide-spread.

Presentations have been invited and given at:

- | | | | |
|-------|-------------------------------|---|----------|
| I. | UCLA | Health Sciences Computing Facility | 3/18/66 |
| II. | ONR | Workshop on Psycho-biology & Computers
(Naval Post-Grad School, Monterey,
California
The paper has been published in the
proceedings of the conference. | 5/17/66 |
| III. | IBM | Research in Cambridge, Mass. | 6/10/66 |
| IV. | IEEE | Workshop on Progress in Time Sharing
Lancaster, Pa. | 10/12/66 |
| V. | University | | |
| V. | University of Toronto, | Toronto, Canada | 8/4/66 |
| VI. | Argonne National Laboratories | Conference on Time-Sharing Model 50's
The paper is due to be published as part of the
proceedings. | 11/1/66 |
| VII. | COMMON Meeting, | New Orleans, La.
A report has been published through COMMON prior
to this meeting which is the result of a joint
effort of IBM and four 1800 users, including ACME:
Report of the 1800 Time Sharing Executive System
Review Committee. | 11/29/66 |
| VIII. | ACM, Los Angeles Section | The presentation has been reviewed in the March
issue of DATAMATION. | 2/1/67 |
| IX. | CALTECH, | Pasadena | 3/29/67 |

In addition, a system description has appeared in IBM's internal documentation.

Even though our system is barely operational, IBM has made arrangements under which the University of Witwatersand, Johannesburg, South Africa and the University of Paris, France have visited us, and are currently communicating extensively on the applicability of the ACME System for their installations.

The progress experienced would not have been possible without the enthusiastic support of the staff.

SENIOR ACME STAFF

<u>NAME</u>	<u>HIGHEST DEGREE</u>	<u>YEARS OF EXPERIENCE</u>	<u>PREVIOUS POSITION OR EXPERIENCE</u>
Gio Wiederhold	BS 1957	9	Visiting Prof., IIT, Kanpur, India Head of Programming, UC Berkeley
* Gary Y. Breitbard	MA 1963	8	Student Compiler UC Berkeley
Charles Class	AA 1963	4	Operator, Stanford
Linda Crouse (part-time)	129 units	4	Desert Research Institute U.C., Davis
* David E. Cummins	BA 1963	5	Programmer, IBM
* Robert Flexer	MS 1965	1	Post-graduate work, UC, Berkeley Electrical Engineering
* Ann Hintz	BA 1961	5	Associate Systems Engineer Associate Programmer, IBM
Klaus Holtz	BS 1964	4	Engineer at Linear Accelerator, Stanford
Zeva LaHorgue (part-time)	BA 1958	2	State Dept. of Public Health, Berkeley Bureau of Chronic Diseases
Gerald Miller (leave of absence)	BA 1964	5	Systems Programmer III, U.C., Berkeley
Mabel Moore	BS 1965	2	Genetics Dept., Stanford
* Arun Patel	MA 1965		Electrical Engineer
William S. Sanders	MS 1964	7	Time sharing projects, General Electric
Voy Wiederhold (part-time)	MA 1960	7	Programmer, UCLA, Health Sciences

In addition to the secretary there are three technician/operators on the full-time staff and a number of part-time student assistants.

Charged to funds provided by the Josiah Macy, Jr. Foundation.

1. SYSTEM DESIGN AND DEVELOPMENT

Thanks to the additional funding provided by the Josiah Macy Jr. Foundation we had been able to do design, simulate, and test hypotheses for the ACME computer system prior to the grant period.

A simulation, written in Burroughs Algol and run on the Stanford Computation Center B5500 computer (ACME Note QL-1) indicated that the scheduling algorithm using "yielding" logic rather than "time slice cutoff" logic was valid. As data for the simulation figures from the MIT timesharing system (MAC) (ACME Note OP-1) and the large computer experience from the Stanford and U.C. Berkeley 7090/7094 systems were used to insure a system responsive to qualitatively large demands. This yielding logic which is a radical departure from large timesharing systems (ACME Note AY-1) currently in existence promises to give the facilities and computing power required by serious researchers. One system using this logic currently in a less general environment is the MEDLAB system at the LDS Hospital, Salt Lake City, Utah, set up by Dr. Homer Warner.

Another aspect of a system in a life-science area is that the data tends to be voluminous. To aid the researcher in the problem of keeping track of the amount of data that can be handled by a large computer system a formal method for handling data is being implemented. All data stored by the system can be automatically identified with the user's name, project, the date and time, file sequence numbers and the actual data name. (ACME Notes FC-1, FI-3, FD-1, FF-1, FP-1, FE-1) Indexes to the data are kept separately to facilitate updating, safe-keeping, and search for data previously collected (ACME Notes FA-1, FLU-1).

Storage modes are limited to textual data and real type numeric data. IBM support for the mechanical storage device (the IBM 2321 pie file) did not arrive until February 15, 1967, and a considerable percentage of the ACME effort is going into the provision of the file capability.

The remainder of the system is operational and is able to support the hardware maximum of 14 users at typewriter terminals, and to respond to real time data requests. (ACME Notes IOA-2, IO-2, MA-1)

The majority of the system is written in FORTRAN, including the scheduling and file supervisory mechanisms so that this system can be transformed to operate on other equipment with a minimum of effort. It also enables us to adjust the system easily on the basis of data gathered from experience, to accommodate equipment changes, and to introduce new ideas and procedures.

2. THE PROGRAM TRANSLATOR

The ACME/PL translator is a true incremental compiler. As far as we can determine it is the only one currently in existence, even though much discussion on the usefulness of such a tool in the timesharing environment has taken place.

It generates absolute binary machine language code from the PL/1 statements typed in by the user.

Through careful subsetting of the PL/1 language (ACME Note PL-2) we can deliver to the user the computational power which is otherwise limited to batch systems. The compiler is operational, although language extensions continue to be added. This rapid development was again made possible largely due to the fact that the predecessor to this compiler, the STUDENT compiler at the Berkeley Computation Center IBM 7040-7094 system had been written in FORTRAN; that planning funds were made available by the Josiah Macy foundation; that the Stanford 7090 system and the IBM 360-50 computers at the Stanford Linear Accelerator and at the Allen-Babcock Company were made available for testing and development (ACME Notes HL-2, HDC-1 and YA-1)

Initial measurements of the performance of the generated code indicate a slightly higher execution speed than IBM's PL/1 compiler currently achieves, although that compiler runs in a batch mode only, and considerably higher speeds than interpretive systems as are generally in use now in interactive time-sharing systems. (ACME Notes KO-6, MT-2, ND-1, NT-4, NS-1)

The compiler takes care also of command handling, (ACME Notes RC-1, LA-2, PC-2) user debugging facilities (ACME NOTE RU-2, LM-3) and is designed to present to the user a consistent one-level interface. The user input/output facilities are not yet completed, but available are the facilities we expect to be used generally: free format multi-value input in response to system prompts and system formatted multi-value output. (ACME Notes FS-2, OF-1, OP-1, Ty-2)

3. DATA ACQUISITION AND DISTRIBUTION PROGRAMMING

The programming to control real time data acquisition and distribution has not yet been integrated into the system. This work could not commence until the computer facility was operating smoothly. Considerable difficulty was initially apparent in our efforts to keep the transfer of information compatible with IBM's Operating System Standards.

This goal required intensive and extensive study of the system. (ACME Notes CP-2, OS-2, CL-1, DN-1, DG-2) We have overcome this hurdle now and we are very satisfied with IBM design in this area. A change of IBM support staff recently has helped us also in the communication required with the manufacturer. We are now able to communicate properly with non-IBM supported devices as PDP-8's, linc-computers, displays, and data transmission apparatus. The integration of these programs into the time-sharing system is scheduled to begin May 1, 1967, and limited user availability is expected in June. In the meantime, check-out of hardware and some production data acquisition is taking place on a scheduled non-time-sharing basis using the software in its current status.

4. MAN MACHINE INTERACTION

In order to improve the man-machine interaction ACME is equipping its terminals with an indicator panel (ACME Note LI-2) so that the user is always aware of the status of the computer and his problem. In addition an ATTENTION Key has been programmed (ACME Note PA-2) to give the user ability to immediately interrupt whatever action the system or his program is carrying out. Another aspect of satisfying the responsiveness of the system is the data acquisition implementation, which assures that the user controls when his equipment is to be sampled. (ACME Note HA-1)

5. DATA TRANSMISSION TESTING AND DESIGN

Considerable effort has been made in testing and developing means for economic data transmission throughout a modern building such as the Stanford Medical School. (ACME Note HDT-1)

This has led us to connect our IBM terminals to the IBM computer directly. The use of telephone communication equipment is hereby avoided with an attendant cost reduction. The cost reduction is largely due to the fact that data transmission use of voice telephone facilities, which are engineered with another set of problems in mind is less than ideal. An ACME built switching panel is located at the computer operator's console for connecting users into the system. (ACME Notes KA-1, KB-2)

In cases where we have to leave the building data-phone and voice frequency FM coded data transmission are used.

ACME has to thank the Instrumentation Research Laboratory of the Department of Genetics who are sharing both their facilities and experience to make the current level of work possible.

6. HARDWARE PRODUCTION AND INSTALLATION

As a result of the field work detailed above, standard circuits have been designed and built. The majority of these used integrated circuit logic, and are mounted and checked out within the ACME Facility. (ACME Note H7-1) These are now in stock as off-the-shelf items and can be combined with standard power supplies in a rack mounted units containing the required number and types of data input and output devices. Two labs (Respiration and Pediatric Cardiology) are currently routinely transmitting data via these links to ACME and the ACME Engineering staff is working at the installation of additional links. (ACME Note HT-1) An input/output typewriter has been connected to one input line to give lower case alphabetic keypunching capability. (ACME Note KP-1)

A link to a small computer has recently been checked out and others are being assembled.

Much of this effort is being done on a cost sharing basis with the laboratories involved, both to conserve ACME funds, and to insure joint real interest in the projects.

7. CATHODE RAY TUBE DISPLAY

In order to present data quickly in high data rate interactive experiments, cathode ray tube displays are of great value. The data to be displayed in much life-science work has the form of time series graphs, annotated with the results of the computer analysis.

With this in mind, as a joint project of ACME and the Instrumentation Research Laboratory a CRT display has been designed with the following features:

- 1) Tube driven by digital logic to insure stability of display
- 2) The digital logic controlled by an independent memory to give an economic source for the required regeneration cycles.
- 3) A memory organization oriented toward vector display to allow effective use of the unit for time-series graphs.

The proto-type of this unit is currently under test, being driven by FORTRAN programs in the 360 and has demonstrated the feasibility of the approach.

The cost of the parts has been about \$6,500 of which the majority is accounted by the CRT tube itself and the core memory unit. The connection to the computer follows ACME small computer conventions and the programming logic is similar to the driving of the CALCOMP digital plotters so that the same programs may be used.

8. STATISTICAL AND TEXT PROCESSING LIBRARY

A beginning has been made with the development of a library to process data on this interactive system.

A number of statistical highly interactive routines are currently available on ACME's Babcock terminal, (ACME Notes EB-1, EBA-1, EBB-1, EBD-1, EBL-1) while a survey has been made of existing statistical routines (ACME NOTE ES-1) which are candidates for inclusion in the system. Testing of various of these is currently in progress, whereas about a dozen are currently available to the users on the Babcock terminal. Much experience is being gathered to organize these routines so that they may be used directly by the medical researcher without having to consult professional or semi-professional programming staff. It is hoped that by the end of the summer a fairly complete statistical library will be available and that the efforts of the group can then be diverted more to the problems of analyzing continuously arriving data.

As a byproduct and extension of the compiler development some text processing routines have become available. These are not yet integrated into the timesharing system, but are available on a stand-alone basis. The current capabilities include text sorting (ACME Notes KC-1, KH-5) concordance preparation, word list with frequency count generation, key word in context type indexing and capability for specifying uninteresting words for deletion (stop words). In the process of check-out are options for searching-for-sentences-containing-specific-words, text comparisons, and generation of data for further statistical processing. The routines are oriented toward the processing of large files and economic usage of core memory. (ACME Note WTXT-1) They have been used at Stanford for analyzing Rorschach test responses, psychiatric diagnosis and setting up clinic appointments.

9. CONSULTATION

A fair amount of staff time has been spent in discussing with staff and faculty of the medical school the feasibility and approach to a large number of projects. ACME has gathered a good impression of the range of problems that the system will have to respond to, and also found a few that cannot be solved with current technology and facilities. As a result of those discussions we feel that the medical school will be waiting to use every resource that becomes available through the system.

10. EDUCATION

To educate the medical faculty and staff in a manner that is directly related to their problems is one of the tasks of a specialized medical facility. To enable an early start for this area a terminal to the Allen-Babcock time-sharing system is being rented through the Stanford Computation Center.

A monthly seminar is being conducted to inform the medical school of the progress with the Project, and to give us the opportunity to hear speakers from other institutions discuss their work in relevant areas.

During November through February a series of 15 four-and-a-half hour courses were conducted which were successfully completed by 167 members of the medical school faculty and staff. (ACME Notes ABC-1 thru 10)

The current demand for these terminals, which have only limited computing capability and no data acquisition facilities, exceeds their availability to the extent that weekly sign-up is required.

An initial draft of a user's manual (ACME Note AM-1) has just been completed which again is oriented towards solving medical research problems. A new series of courses is due to start in May using ACME's own facilities.

11. EQUIPMENT INSTALLATION

The installation of the computer equipment in the specially built structure was finally completed on April 8, 1967.

Due to delays in approval of the various funds used to construct a special structure and adapt it to the computer's requirements, primitive and novel means were used to enclose the computer and keep it operating while the construction progressed. The dust has finally settled down and the facility is now not only functional, but also extremely attractive and much commented on by visitors to the STANFORD-PALO ALTO HOSPITAL and the Medical School.

Thanks are due here to the Stanford Business School who made a computer floor available, the medical school architect's office who believed that the impossible was possible, and IBM who were willing to risk their equipment to the elements and the construction crews.

12. COMPUTER OPERATIONS

The operation of the computer is handled through arrangements with the Stanford Central Facility which enables us to secure reliable 24-hour-7-day staffing without having to employ redundant back-up staff. The technical aspect of operations is supervised by a member of the ACME staff. Full staffing has resulted in very reliable operations under unfavorable conditions and our machine has shown an availability during the period from 12/12/66 to 4/09/67 of 97.5 percent after scheduled maintenance (4.0%). Much of the computer's time is being used for ACME hardware check-out. The ACME System development takes priority currently, but users problems are routinely run at least overnight.

The system has a very poor batch performance since the requirements for time-sharing have taken precedence in both hardware and software selection. (ACME Notes CN-2, CQ-1, DL-1, OD-1) It will be interesting to compare cost to productivity ratios when the time-sharing service is in full swing.

Gio Wiederhold

LIST OF TERMINALS

<u>NAME</u>	<u>DEPARTMENT</u>	<u>DATE OF INSTALLATION</u>
Ed Brown	Dept. of Medicine	
Dr. J. W. Bellville	Anesthesia	
Dr. K. M. Colby	Computer Science Dept.	12/20/66
Dr. K. L. Chow	Neurology	
Dr. E. Dong	Surgery	1/16/67
Dr. Djerrassi	Genetics/Chemistry	
Dr. W. Forrest	Anesthesia	12/20/66
Dr. Fred Fox	Clinical Lab.	
Dr. Allen Gates	Gynecology	1/16/67
Dr. A. Goldstein	Pharmacology	
Dr. D. Harrison	Cardiology	
Dr. L. Herzenberg	Genetics	
Dr. A. M. Iannone	Neurology	
Dr. K. Killam	Pharmacology	1/16/67
Dr. Kopell	Psychiatry, VAH	12/20/66
Dr. J. Lederberg	Genetics	
Dr. S. Liebes	Genetics	
Dr. Mesel/Conn	Pediatrics	2/6/67
Dr. Mesel/Northway	Pediatrics	
Dr. Mesel/Radiology	Pediatrics	
Dr. T. Nelson	Surgery	
W. Reynolds	Genetics	
Dr. L. Rosenberg	Med. Micro-Biology	
Dr. Stewart (3)	Fleischman Lab.	
Dr. L. Stryer (2)	Biochemistry	
Dr. Luetcher/Van Kessel & Dr. Wasserman	Dept. of Medicine	1/16/67
Dr. Bagshaw	Radiology	
Dr. T. Merigan	Infectious Diseases	
Dr. George Wertheim	Psychiatry	
Dr. S. Kountz	Surgery	1/16/67
J. Wong	Genetics	

APPENDIX B

DENDRAL - A COMPUTER PROGRAM FOR GENERATING
AND FILTERING CHEMICAL STRUCTURES

February 15, 1967

DENDRAL - A COMPUTER PROGRAM FOR GENERATING
AND FILTERING CHEMICAL STRUCTURES

by Georgia Sutherland

Abstract: A computer program has been written which can generate all the structural isomers of a chemical composition. The generated structures are inspected for forbidden substructures in order to eliminate structures which are chemically impossible from the output. In addition, the program contains heuristics for determining the most plausible structures, for utilizing supplementary data, and for interrogating the on-line user as to desired options and procedures. The program incorporates a memory so that past experiences are utilized in later work.

The research reported here was supported in part by the Advanced Research Projects Agency of the Office of the Secretary of Defense (SD-183)

Table of Contents

1. Dendral Representation of Chemical Structures
2. Dendral Implementation
3. The Dendral Program
4. Modifications to the Dendral Program
5. Graph Matching in Dendral
6. Summary and Discussion

1. DENDRAL REPRESENTATION OF CHEMICAL STRUCTURES

Dendral is a system of topological ordering of organic molecules as tree structures.* The essential features are detailed in the first of the references and are summarized here.

Proper Dendral includes precise rules to maintain the uniqueness and the non-ambiguity of its representations of chemical structures. Each structure has an ordered place, regardless of its notation; the emphasis is upon topological uniqueness and efficient representation of molecular structures. The principal distinction of Dendral is its algorithmic character. Dendral aims (1) to establish a unique (i.e., canonical) description of a given structure; (2) to arrive at the canonical form through mechanistic rules, minimizing repetitive searches and geometric intuition; and (3) to facilitate the ordering of the isomers at any point in the scan, thus also facilitating the enumeration of all of the isomers.

The Dendral representation of a structure is made up of operators and operands. The operators are valence bonds issuing from an atom. Each bond looks for a single complete operand. An operand is (recursively) defined as an unbonded atom, or an atom whose following bonds are all satisfied in turn by operands. Hydrogen atoms are usually omitted, but are assumed to complete the valence requirements of each atom in the structure. If the structure has unsaturations (one unsaturation for each pair of hydrogen atoms by which the structure falls short

*References: J. Lederberg, DENDRAL-64, A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs, Parts I-V, Interim Report to the National Aeronautics and Space Administration, December 1964.

of saturation), these are indicated by locations of double and triple bond operators. Single, double, and triple bonds are represented by . , : , and : respectively. The operator : may be represented by = and the operator : by \$ or < depending on the available character set.

As an example, the molecule $\text{NH}_2 - \text{C} \begin{array}{l} \diagup \text{O} - \text{CH}_3 \\ \diagdown \text{S} \end{array}$

has one unsaturation and may be written in many ways, including:

- 1) C.O.C.:NS
- 2) C.O.C:..SN
- 3) O..CC.:NS
- 4) O..C.:NSC
- 5) C..:O.CNS
- 6) C..:O.CSN
- 7) C:..NSO.C (canonical)
- 8) C:..SNO.C
- 9) S:C..O.CN
- 10) N.C.:O.CS

Each of these ten notations is a non-ambiguous representation of the molecule. However, proper Dendral also specifies that the representation be unique. The key to obtaining the unique or "canonical" representation is the recognition of the unique center of any tree structure and the subsequent ordering of successive branches of the tree.

The centroid of a tree-type chemical structure is the bond or atom that most evenly divides the tree. A molecule will fall into just one of the following categories, tested in sequence. Let V be the count of non-hydrogen atoms in the molecule. Then either

A. Two central radicals of equal count are either (1) united by a leading bond (V is even) or (2) sister branches from an apical node (V is odd); or

B. Three or more central radicals, each counting less than V/2, stem from a single apical node.

In the first case, the centroid is a bond, and the canonical representation is an operator followed by two operands. In the other two cases the centroid is an atom, and the canonical representation is an operand in the form of an atom followed by two or more bonds and operands. In every case where two or more bonds follow an atom, the operands must be listed in ascending Dendral order.

Dendral order (or simply "weight") is a function of the composition and arrangement of a structure and finds its primary use when comparing two operands (radicals). The weight of a radical is evaluated by the following criteria (in descending significance): count, composition, unsaturation, next node, attached substructures.

Count is the number of skeletal (non-hydrogen) atoms. Of two radicals, the one with the higher count is of higher weight.

Composition refers to the atoms contained in the radical. An arbitrary ordering of the atoms makes carbon less than nitrogen less than oxygen less than phosphorus less than sulfur, $C < N < O < P < S$. (This ordering is alphabetical as well as by atomic number.) When comparing two radicals of the same count, the one with the fewer number of carbons has lesser weight. If carbons are equal, the one with the fewer nitrogens is of lesser weight. And so forth.

Unsaturation counts the number of extra bonds (1 for a double bond, 2 for a triple bond) in the radical, including those (if any) in the afferent link (the bond leading into the radical). Of two radicals, the one with the greater number of unsaturations has the greater weight.

The next node or apical node refers to the first atom in the radical (the one connected to the afferent link). When comparing two apical nodes, the following three criteria are evaluated (in order of decreasing significance):

Degree is the number of efferent (attached) radicals. The apical node with the most radicals attached to it has the greater weight.

Composition refers to the type of atom. A carbon atom is the lowest apical node, while a sulfur atom is the highest.

Afferent link refers to the bond leading to the apical node. A single bond afferent link is the lowest, a triple bond is the highest.

If the above criteria fail to determine which of two radicals has the greater weight, then the radicals appendant on the two apical nodes must be arranged in increasing order and compared in pairs. The first inequality in weight of appendant radicals determines the relative weight of the original radicals.

The canonical representation for the molecule in the example given earlier is notation #7. It must be a central atom molecule since its count (ignoring hydrogen atoms) is 5; and the non-terminal carbon atom is the only atom which has all its appendant radicals with counts less than $5/2$. Of the three appendant radicals, the one containing two atoms has the highest count and thus is the heaviest. Of the two radicals containing a single atom each, the one with the double bond is the heavier because it has more unsaturations.

Even-count molecules may have a bond for center, if the count of the molecule is evenly divided by cutting that bond. Thus, the canonical form for $\begin{array}{c} \text{NH}_2 \\ \diagdown \\ \text{CH}_2 \end{array} - \begin{array}{c} \text{OH} \\ \diagup \\ \text{CH}_2 \end{array}$ is .C.NC.O, a leading bond, the first dot, calling for two operands.

The collection of rules and conventions described above provides a unique and non-ambiguous representation for any non-ringed chemical structure. (Ringed structures have been dealt with by more complex rules.) In addition, the rules also allow us to construct the "lowest" structure which can be made from a composition (collection of atoms). Once this lowest structure has been made, it may be transformed by a process of rearranging its atoms and saturations into the "next to lowest" structure. This "incrementing" process may be continued until the "highest" structure has been made.

The computer program which is described in later sections of this report is designed to do these operations and therefore to construct all of the (topologically possible) isomers of a composition.

2. DENDRAL IMPLEMENTATION

The task set forth in the Dendral Report is the manipulation of chemical graphs to produce all the isomers of a given chemical formula. The list processing language, LISP, was used to write a computer program implementing the proposed scheme. The choice of this language led to the representation of the chemical graphs as tree type lists.

The list representation is a straightforward translation of Dendral dot notation. The bonds are represented by the integers 1,

2, and 3. Each radical is a sublist and is enclosed in parentheses. The list notation for a structure is a three part list. The first part specifies the afferent link; the second part specifies the apical node; and the third part specifies the efferent radicals in list notation. If the structure is a molecule, its afferent link will be NIL . For a central-bond molecule the apical node is NIL also. Central-bond molecules have two efferent radicals; central atom molecules have two or more efferent radicals, and radicals may have any number (including zero) of efferent radicals.

As an example, the dot notation C.:OOS.N is translated into (NIL C (1 O)(2 O)(1 S(1 N))); and the molecule represented in dot notation as .C.:SS C.O.P..CC:C becomes (NIL NIL (1 C(1 S)(2 S))(1 C(1 O(1 P(1 C)(1C(3C)))))).

The list notation is easily manipulated by the computer program, so all operations are performed using this representation. Output, however, is given in Dendral dot notation. Certain functions are available within the Dendral program for converting back and forth between the two types of notation. The function INDOT reads dot notation and converts the dots to integers. (This is the so-called Dendral Polish notation.) The function UNSTRING converts Polish notation to list notation. The function DOTORD also reads dot notation, but converts it to canonical dot notation. The functions MOLORD (for molecules) and RADORD (for radicals) convert list notation into canonical list notation. The functions TOPMOL (for molecules) and TOPRAD (for radicals) convert list notation to dot notation and print the dot notation.

The Dendral program operates on a chemical composition in order to produce the structural formulas of all isomers with that composition. A composition is a list of numbers of atoms, such as: C_4H_{10} , or C_2H_5NO . The internal form of a composition is a list of dotted pairs in which the number of hydrogen atoms is replaced by the number of unsaturations (extra bonds) in the composition. The formula for obtaining the number of unsaturations (U) in a composition is the following:

$$N = 2U = \sum_{\substack{\text{all} \\ \text{types} \\ \text{of} \\ \text{atoms}}} \left(\begin{array}{l} \text{number of atoms} \\ \text{of this type} \end{array} \right) \times \left(\begin{array}{l} \text{valence of this} \\ \text{type of atom} \end{array} - 2 \right)$$

Examples of composition lists are ((U . 0)(C . 4)) and ((U . 1)(C . 2)(N . 1)(O . 1)).

Some compositions imply molecular structures while others imply radicals. If $N = 2U$ is an odd integer, then the structures will be radicals. If $N = 2U$ is even (including zero), the structures will be molecules. If $N = 2U$ is negative, there are no structures possible for that composition.

In the process of generating all structures corresponding to a chemical formula, the program starts with the formula given as a composition list. The program then constructs the molecule* of lowest Dendral value for this composition. Once the molecule of lowest

* The current discussion will concentrate on molecular structures. Radicals are generated in an analogous fashion.

value is established, the process of generating all the isomers consists of incrementing this molecule to the next higher Dendral structure, then taking the new molecule and repeating the process until it is no longer possible to make a molecule of higher Dendral value than the previous one.

The following computer program is a precise statement of the generating algorithm mentioned in the Dendral Report, with the following added features:

- a. The number and type of radicals that can be made from a given composition may be determined in advance and placed in a dictionary list for that composition. When a dictionary list is encountered (during structure generation) for a composition, the algorithm will generate only those radicals on the dictionary list. Thus a dictionary is a "memory" for past work. The program knows how to make use of its memory efficiently.
- b. At any given node it is possible to represent the efferent radical in an implied format (by composition rather than by structure).
- c. Several options are available which may limit the output by eliminating or bypassing some structures which are topologically possible but which are not of interest or perhaps are not chemically meaningful.

3. THE DENDRAL PROGRAM

The computer program to implement Dendral is written in

LISP. The program is made up of over a hundred separate LISP functions which call each other to perform certain tasks relevant to different parts of the structure generation. Most of the functions are simple utility programs. The main logic is contained in eight or ten primary functions which are described in the following pages.

The LISP function called GENMOL will make the molecule of lowest Dendral value from a given composition. If the composition contains an even number of atoms, GENMOL will attempt to make a central-bond molecule by making two radicals with equal count and identical afferent links. If this fails or if the composition contains an odd number of atoms, then a central node is selected, starting with the atom of lowest Dendral value ($C < N < O < P < S$) in the composition. MAKERADS is then given the remainder of the composition and instructed to make two efferent radicals of lowest Dendral value. If this fails, then GENMOL selects the next possibility for the central node and the process of generating two efferent radicals is attempted again. If no atom in the composition leads to a structure with degree two, then the lowest atom is again chosen for the central node, and MAKERADS attempts to make three efferent radicals, and so forth.

The function MAKERADS takes a composition and produces a list of a specified number of radicals. These radicals have the lowest Dendral value which is possible in view of the valence requirements specified in the arguments to MAKERADS. Each radical except the last must have a specified number of atoms, and the radicals must be listed in increasing Dendral order. First the total composition is split into

the proper number of compositions by MAKELSTCLS. These separate compositions (listed in increasing Dendral order) are then made into radicals by GENRAD. If any of the compositions cannot be converted to a radical then this composition is incremented to the next greater Dendral value. Any compositions which are greater than the one incremented are reset to compositions greater than or equal to the new one. The process of converting these lists to radicals is then continued as before. Once the compositions are all converted to radicals, the radicals are checked to see if certain valence requirements are met. If not, an attempt is made to decrease the afferent link of each radical in turn starting with the radical of greatest Dendral value. When a radical has the lowest allowable afferent link and the valence requirements are not satisfied, then the compositions of this radical and those of greater Dendral value are incremented by trading atoms among compositions. New radicals are generated from each composition and the process of decreasing afferent links continues.

GENRAD takes a composition and makes the radical of lowest Dendral value within valence limitations specified. If the composition consists of a single atom then this atom with an afferent link accounting for all the unsaturations in the composition will be the desired radical. Otherwise, the lowest apical node is selected from the composition, the lowest possible afferent link is chosen (considering the atoms and unsaturations in the remaining composition), and GENRAD is used (recursively) to generate a single efferent radical

from the remaining composition. If this fails, the afferent link will be increased and GENRAD tried again. If this too fails, then the composition of the apical node will be increased, the afferent link reset, and the whole process repeated until there are no more possibilities for a new apical node. If this process fails, then it is repeated allowing GENRAD to generate two efferent radicals. The number of efferent radicals continues to be increased if necessary and is limited only by the valence of the apical node.

UPRAD increments a given radical to its next highest Dendral value. This is done by first trying to increment one of the efferent radicals from the apical node of the radical (if there are any efferent radicals). Starting with the efferent radical of greatest Dendral value, UPRAD is applied to obtain a new efferent radical and to reset all of the efferent radicals of greater Dendral value so that they are as low as possible but still higher than the one just generated. If this proves unsuccessful, then an attempt is made to, first, raise the afferent link of the radical (using the function UPLINKNODE), or to increase the composition of the apical node (using the function UPCOMPNODE), or, finally, to increase the degree of the apical node (using the function UPDEGNODE). If none of these is successful, and there are radicals of higher Dendral value concurrent with the one being incremented, then MAKERADS is used to make a new set of efferent radicals in which the composition of the first radical is greater than that of the present radical being incremented.

UPMOL uses UPRAD in an attempt to step up one of the

efferent radicals from the center of a molecule. These efferent radicals are given to UPRAD one at a time starting with the one of greatest Dendral value. If it is not possible to step up any of the efferent radicals, then an attempt is made to find a central node of greater compositional value and then use MAKERADS with the remaining composition to find the set of efferent radicals for this central node. (If this is a molecule with a bond as its center, then instead of finding a new central node UPMOL tries to increase the value of the central bond.) If these steps fail, then UPMOL tries to raise the degree of the central node while resetting this node to its lowest value.

The three functions UPLINKNODE, UPCOMPNODE, and UPDEGNODE operate on a radical and attempt to make the next higher radical out of the same elements. In each case the function MAKERADS is the primary tool. UPLINKNODE is asked to return a radical with a higher afferent link. UPCOMPNODE is asked to return a radical with an apical node of higher composition. UPDEGNODE is asked to return a radical in which the apical node has an increased number of efferent radicals.

The main control function for the structure generation is a function called WORKLIST. WORKLIST causes generation of all structures from all composition lists which are subsets of a given composition list. The user specifies this composition list and requests either molecules or radicals to be generated. WORKLIST calls TESTDENDRAL (for molecules) or TSTRAD (for radicals). TESTDENDRAL first calls GENMOL to obtain the lowest molecular structure for that composition list. The resulting structure is printed, and TESTDENDRAL then alternately calls UPMOL and OUTDEN (a print function) until UPMOL

returns the value NIL, indicating that no higher structures can be generated. TSTRAD calls GENRAD and UPAD in an analogous fashion for radicals.

The usual case is that the user has in mind a single chemical formula for which he wishes to see all the allowable structures. The function CHNOPSXVQW takes a chemical formula and converts it to a composition list by calculating the number of unsaturations. CHNOPSXVQW also determines whether molecules or radicals will result. Then WORKLIST is called and instructed to generate structures from this single composition list.

The input to the function CHNOPSXVQW is a list of ten elements, corresponding to the letters in the name of the function. The first six elements are the number of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur atoms in the formula. X is the name of any other atom in the formula. (If X is the word NIL, then no other atoms are present). V is the valence of X; Q is the number of X atoms in the formula; and W is the atomic weight of X.

Other functions have been supplied to accept different forms of input and call CHNOPSXVQW after constructing the appropriate list of arguments. The function ISOMERS takes a single argument which is a composition name (i.e., C₃H₁₀, C₂H₅NO, CH₃COOH, etc.). The function DENDRAL takes no arguments, but later requests the user to specify desired options and to input a composition name. The function STRUCTURES alternately requests composition names and calls ISOMERS so that it is easy for the user to examine the isomers of many compositions in succession.

4. MODIFICATIONS TO THE DENDRAL PROGRAM

The program described above does indeed produce all the structural isomers from a chemical composition. Sometimes, however, the number of isomers is so large that a user may not want to see all of them. Thus, the program will pause after each N structures (N may be set by the user) and ask the user whether or not to continue generating structures.

The "model" of chemistry in basic Dendral includes the following subjects:

- a) Atoms; there are seven distinct atoms (C,H,N,O,P,S and X). Of these, all are treated the same except H.
- b) Valence of an atom; valence specifies the number of attachments an atom may acquire.
- c) Unsaturation; the program knows that unsaturations indicate multiple order attachments and the program knows how to calculate the number of unsaturations for any composition.

Using the above concepts and the Dendral rules for building structures, the program constructs all of the topologically possible structures for any given composition. However, any chemist inspecting the output list for a composition would realize that the program knows little chemistry, since many chemically meaningless structures are included in the class of topologically possible structures. So additional information and programming was added to the Dendral Program in order to eliminate certain types of structures which are always chemically impossible.

This was done in two ways. The first uses the notion of illegal attachment. For instance, an oxygen atom cannot be attached to another oxygen atom. (Other such conditions hold for nitrogen and sulfur atoms.) Whenever the program picks an atom to be the apical node of a radical, it checks the partially built structure to be sure that this atom will not be attached to a forbidden atom or structure. The second and more general way of avoiding certain chemically impossible structures is by presenting the program with a list of impossible substructures. Each generated structure is then examined and rejected if it contains any of the forbidden substructures. The process of checking for certain substructures led to the incorporation of a graph matching algorithm in the Dendral program. The graph matching process is described in detail in later sections of this report.

Eliminating chemically impossible structures from the program output means that the program no longer creates all structural isomers of a chemical composition. Rather, the program now contains some knowledge of chemistry. The on-line user of the program may want to impart more knowledge to the program. A small step toward this goal is provided in three ways.

a) Supervising structure generation:

The program can be made to pause every time it is about to add an atom to a partially generated radical. The program prints out the current status and requests permission to continue. A "no" answer at this point should invoke some (as yet unwritten) learning dialog

to find out why not to continue.

b) Examining molecular partitions:

Certain groupings of atoms are more plausible (chemically speaking) than others. The program can be made to print (in advance of any structure generating) all the partitions of the composition. The user may then rearrange the list of partitions, eliminating any in which he is not interested. The program will then generate structures in the order specified by the list of partitions. The learning process which could be inserted at this point would interrogate the user as to why some partitions were left out and why some are more plausible than others.

An option is included at this point which allows the user to rearrange the partition list on the basis of a built-in plausibility function. This function calculates a "plausibility score" for each partition and then rearranges the partition list, putting the partitions with the highest scores at the beginning of the list. The criteria which are considered in calculating the score of a partition are built into the program.

The function GETSCORE examines a partition (a list of compositions) and returns a number between 1 and 10 which is an assessment of the value of the partition. If any of the compositions with more than one atom contains no carbon atoms, then the score for the whole partition is 1. Otherwise the score is the average of two numbers:

The first number is obtained by considering the proposed center of the structure. A central bond is assigned the value 10. A non-carbon central atom is assigned the value 3. A central carbon with degree two is assigned the value 10. A central carbon with degree three is assigned the value 6. And a central carbon with degree four is assigned the value 1.

The second number is the average of N values, where N is the number of compositions in the partition (aside from the central node, if present). If the composition is COOH then the value is 10. Otherwise the value is a number between 0 and 10 measuring how closely the proportion of carbon to non-carbon atoms matches the overall proportion for the whole partition.

c) Using other data (spectra):

Spectral information may be used to further shorten the list of isomers of a chemical formula. A spectrum is a list of numbers corresponding to weights of substructures. The Dendral program "knows" the atomic weight of each atom, and can calculate the weight of any structure or substructure. If a spectrum is present, then no structure or substructure will be generated unless its weight appears in the spectrum. A function called HSTGRM will calculate the spectrum of a structure in list notation. Another function called USESPECTRUM requires the user to input a spectrum or structure from which to generate a spectrum. This function then sets SPECTRUM and calls structure generating functions.

When structure generation is complete,

USESPECTRUM sets SPECTRUM to NIL .

Future plans for the Dendral program include expanding the dialog, partition, and spectral facilities, streamlining the graph matching algorithm, allowing several levels of memory, and providing more utility functions for the benefit of the program user.

5. GRAPH MATCHING IN DENDRAL

The Dendral program may be made to generate all topologically possible structures from a chemical formula, restricted only by the valence limitations on the atoms. Many of the structures generated by the Dendral algorithm are not chemically meaningful because they contain certain impossible substructures. The forbidden substructures were few enough in number that they could be listed; and a graph matching algorithm was introduced to check each generated structure against the list of forbidden substructures. If a structure generated by the rules of Dendral is found to contain even one of the forbidden substructures, it is not acceptable output, and the Dendral program attempts to find the next higher structure which does not contain a forbidden substructure.

The first graph matching algorithm that was implemented to compare Dendral-generated structures was essentially that of E. H. Sussenguth Jr. of Harvard. This algorithm is described in detail in an appendix to this report. After using this algorithm for some time, however, it was determined to be inefficient in the sense that previous work was constantly being repeated. This resulted from the

fact that graph matching was performed at each level of structure generation, whenever an atom was added to a partially built structure. The Sussenguth method considers the total structure and calculates characteristics of each node (atom). Yet, the nature of the structure generating algorithm implied that if any forbidden substructures were to be present, they would have to include the most recently added portion since the remainder would have been checked previously.

Thus, a simpler graph matching algorithm is now being employed. The structure to be checked is first put into the appropriate format. This format is identical with the list notation representation of the structure except that hydrogen atoms must be included. The function ADDH(S) converts the structure S to the appropriate format. The function NEWCHECK causes this conversion to be made, and then compares the structure with each element of BADLIST to determine whether any forbidden substructure is present.

BADLIST is merely a list of forbidden substructures. Each element is the list notation representation of a structure, with the following alterations:

- 1) The substructures have no afferent links.
- 2) If any node can be one of several types of atoms, the list of atoms is put in place of the usual single atom name.

NEWCHECK examines BADLIST and extracts those structures whose apical node is the same as the apical node of the structure being checked.* Then the structure is searched for each of the appendant

* Because of this, BADLIST must contain several entries for some structures, one entry for each possible apical node. Thus the elements of BADLIST are not all in canonical form.

radicals of the forbidden substructure. If all are found, then the substructure is present, and NEWCHECK returns the value T.

The primary tool used in this process is the function COMPAR. COMPAR takes two arguments, S and L. Each argument is a list of radicals representing some or all of the efferent radicals of the two graphs being matched. The radicals on L come from the forbidden substructure. If not all elements of L are on S, then COMPAR returns NIL. If all elements of L are on S then COMPAR returns the dotted pair (T.R) where R is the remainder of list S after the elements of L have been removed.

This graph matching algorithm is far simpler than the one described in the appendix. It requires much less code (and consequently uses much less core memory) and should prove to be quite a bit more efficient for the types of problems encountered in structure generation.

6. SUMMARY AND DISCUSSION

The Dendral program is constantly being modified as new and better ways of doing things are conceived. The basic structure generating functions are written independently in such a way that new supervisory functions can be easily inserted into the system. It is hoped that the Dendral program will eventually be able to benefit from the user-on-line characteristics of the time-sharing system in order to "extract knowledge" from chemists and other users of the program.

One goal is to be able to give the program a real mass spectrum and have the program predict only a small number of structures which are most plausible. Considerable work must be done before this

goal can be realized. One problem is the determination of which composition was represented by the original substance. Another problem is "cleaning up" a real spectrum. A third problem is obtaining a prediction of the spectrum of a given structure for purposes of comparing with real data. The present spectrum predictor is only a crude, first-order attempt at generating the spectrum of a structure.

The process of generating the most plausible structure from a given composition can be improved by making use of a chemist's knowledge about commonly occurring substructures. Soon the program will contain a GOODLIST which can keep track of likely combinations of atoms. Then, during the initial consideration of the composition, the program can remove certain groups of atoms and replace them by a symbol representing the substructure. Then, using the arbitrary atom X in the function CHNOPSXVQW, the program can treat this substructure as an atomic unit and insure that all generated structures will contain the desired substructure.

Other proposed modifications to the program include the following items:

- 1) Revising the method of remembering past work so that the "dictionary of structures" requires less core storage space.
- 2) Putting the plausibility parameters within the reach of the user so that different schemes for rating plausible partitions may be compared.

- 3) Introducing more efficient "tree pruning" methods so that the search space for plausible structures will be made smaller.
- 4) Introducing "mood items" for BADLIST so that the program can interrogate the user as to what general class of compounds he expects. Certain structures will be added to or removed from BADLIST as a result of the user's "mood".

It should be evident that the Dendral program is constantly expanding and becoming more sophisticated. In evaluating the efforts performed up until now, the importance of the basic Dendral notation for providing unique, non-ambiguous, and algorithmic representation of chemical structures cannot be overstated. Not only has the Dendral notation allowed all the isomers of a chemical formula to be quickly, accurately, and completely generated; but also it has provided a base for studying a certain "inductive and enquiring" system.

APPENDIX 1

The Sussenguth Method of Graph Matching as Implemented in the Dendral Program

1. GRAPH MATCHING IN DENDRAL

The Dendral program may be made to generate all topologically possible structures from a chemical formula, restricted only by the valence limitations on the atoms. Many of the structures generated by the Dendral algorithm are not chemically meaningful because they contain certain impossible substructures. The forbidden substructures were few enough in number that they could be listed; and a graph matching algorithm was introduced to check each generated structure against the list of forbidden substructures. If a structure generated by Dendral is found to contain even one of the forbidden substructures, it is not acceptable output and the Dendral algorithm attempts to find the next higher structure which does not contain a forbidden substructure.

The graph matching algorithm that has been implemented to compare Dendral-generated structures against the BADLIST is essentially that presented in the PhD thesis of E. H. Sussenguth, Jr. at Harvard, 1964. This algorithm considers two graphs, made up of nodes (atoms) and connections (bonds), and compares sets of nodes with equal properties. Using set operations such as union and intersection, and making assignments of node correspondences where sufficient information is not

present for a unique determination, the algorithm avoids a time-consuming direct node-by-node comparison of the two graphs. This algorithm is especially efficient in determining when no match exists. The process of actually finding an isomorphism takes somewhat longer.

2. NOTATION FOR GRAPH MATCHING

The form of a graph G is a list of elements, each element representing a node of the graph.

Each node-element has three parts:

- 1) a node number- an integer. Node numbers must be unique. And each node number is equal to the position of that node-element in the list that forms the graph.
- 2) a node type- the name of the atom at that node (usually C, N, O, P, or S -- but sometimes an abbreviation for a whole structure that is to be treated as a unit - such as OH, NHR, COOH, etc.)
- 3) A list of connections- This list has as many elements as the node has bonds connected to it. Thus the length of this list cannot (theoretically) be greater than the valence of the node. (In practice this restriction is irrelevant.) Each connection is a two element list of the form (CN CB). CN is the node number of the connected node and CB denotes the type of connection:

CB=T (triple bond), CB=D (double bond), CB=S (single bond)

For example, the Dendral notation $. C = O N.C. = C C$

would be represented by ((1 C(2 D)(3 S))(2 O(1 S))(3 N(1 S)(4 S))(4 C(3 S)(5 S)(6 D))(5 C(4 S))(6 C(4 D))) for purposes of graph matching.

Operations comprising the graph matching algorithm depend heavily on the use of node numbers, both alone and in list (sets). An isomorphism (answer) is a pair of lists of node numbers, one from each of the two structures being considered, and where nodes in corresponding positions are isomorphic.

A property for a graph is a list of node numbers headed by an atom denoting the value common to all the nodes. Various property values are used in matching the graphs. They are described briefly below:

- 1) Node value. All nodes with value C are grouped together and headed by the atom "C". (etc. for N, O, P, ..)
- 2) Branch value. All nodes adjacent to single bonds are grouped together and headed by "S". (similarly for D and T)
- 3) Gamma degree. The gamma set of a node n is the list of nodes reachable from n by a path of a specified length (i.e., by traversing a specified number of bonds). Thus, the gamma=1 set for a node n is a list of all nodes immediately adjacent (connected) to node n. The gamma=2 set for node n is the list of all nodes connected to the nodes in the gamma=1 set of node n. The gamma=m degree of node n is the number of nodes (length) of the gamma=m set for node n.
- 4) Quasi-order. The quasi-order of node n is the number of bonds (connections) that must be traversed before getting back to the given node n. Direct backtracking is forbidden.

Thus, the quasi-order for any node in a non-ringed or tree structure must be zero. (Quasi-order is not included in the current version of the Dendral program since ring structures are not included.)

5) Connectivity. Connectivity of a set of nodes is another set of nodes representing all nodes which can be reached by a path of given length from any one of the original nodes. Thus, connectivity of a set of nodes is the union of the gamma sets of all its nodes. Connectivity is usually found for paths of length 1 or 2, but it could be calculated for longer paths until redundant answers begin to be generated.

3. DESCRIPTION OF THE GRAPH MATCHING ALGORITHM

The computer program for the graph matching algorithm is written in the language LISP. The operation of the algorithm is based on comparing corresponding properties from each of two graphs in an attempt to find pairs of nodes with identical properties. During this pairing process, careful checks are made to determine whether otherwise similar nodes have contradictory properties. Such a situation indicates "no match", and the algorithm terminates immediately.

The algorithm makes heavy use of variables which are global to all LISP functions. These variables are certain lists which are set, changed, and translated by the major functions. Eight of these lists deserve special mention:

- 1) G -- one of the two graphs to be matched. This

variable is set as input to the algorithm and never changed.

2) GS -- the other graph to be matched. GS must be the larger of the two (in the sense of having more nodes).

3-4) -- L and LS are lists whose elements are sets of nodes with corresponding properties from G and GS respectively.

5-6) -- V and VS are edited versions of L and LS, being lists of sets of nodes of G and GS with corresponding properties. Non-informative sets have been removed, and pairs of nodes that are known to correspond are removed from the sets in which they appear.

7-8) -- K and KS are lists of corresponding (known) nodes for G and GS. The first element of K is the node corresponding to the first element of KS. If the length of these sets equals the number of nodes in graph G, then an isomorphism is determined. There may be more than one isomorphism, but the current version of the algorithm is satisfied with a single one.

There are two classes of functions making up the Dendral graph matching algorithm. The first class contains all functions that are an integral part of the pure subgraph matching. The second class of functions contains all those needed to use the graph matching from within Dendral. This class depends on the presence of both pure

subgraph matching and pure Dendral.

The important functions in the first class are ISOLISP (the supervisory function), SETFORM (which extracts properties of graphs), FF (a function which obtains families of corresponding nodes), PAR (the function which obtains correspondents of the nodes of the smaller graph), and NEWABLK and NEXTASSIGN (which provide for systematic assignment of correspondences in cases where previous work has failed to find a unique correspondent for some node).

The function CHECKMATCH is the link between Dendral and the graph matching algorithm. It converts a structure to notation suitable for graph matching and calls ISOLISP for each element of BADLIST which may be contained in the test structure.

BADLIST is a global variable which contains information about all of the "forbidden" chemical structures. Each element of BADLIST contains the following pieces of information about a forbidden structure:

- 1) A predicate which indicates whether (T) or not (NIL) the structure to be matched with this subgraph should be checked for terminal OH and NHR before matching.
- 2) A number indicating the number of atoms (nodes) in the subgraph.
- 3) The composition list for the subgraph.
- 4) The subgraph itself in the notation required by the graph matching algorithm.

The function ISOLISP is called only if the number of atoms in the subgraph is less than or equal to the number of atoms in the

test structure and if the composition of the subgraph is contained in the composition of the test structure. The value of CHECKMATCH is T if any forbidden subgraph is found in the test structure. Otherwise the value of CHECKMATCH is NIL.

4. PRIMARY FUNCTIONS FOR GRAPH MATCHING

ISOLISP - 2 arguments

G - a graph

GS - a graph

ISOLISP is the control function for the algorithm. It calls other functions which construct and use the lists L, LS, V, VS, K and KS. ISOLISP recognizes when an isomorphism has been found or denied and calls for assignments to be made if necessary.

The first action of ISOLISP is to set up lists L and LS from the properties of graphs G and GS by calling the function SETFORM. Next, the elements of L and LS are examined and non-redundant sets of nodes are placed on V and VS by the function FF which checks for possible contradictions in properties of G and GS. The function CONNEC places new sets on L and LS. These sets are obtained by applying the property of connectivity to all sets on V and VS. The function PAR obtains the set of correspondents for each node of G which is not on list K. These sets are added to V and VS by the function PAR.

If the sets V, VS, K and KS are unchanged after the functions FF and PAR have both been executed, then an isomorphism cannot be determined without making an assignment. The function NEWABLK causes a node assignment to be made and added to the lists K and KS. If no

assignment is possible, then a previous assignment must be contradicted. (If no previous assignment was made, then no isomorphism is possible). The variable called ASTACK contains a record of current assignments. NEXTASSIGN revises the most recent assignment if possible. Otherwise, the most recent assignment is discarded and the state of the system prior to that assignment is retrieved in order to revise the previous assignment.

FF-0 arguments

FF uses the lists L and LS to form families of corresponding nodes. Each pair of elements (L_i, LS_i) is examined for useful information. Known corresponding nodes are removed (by the function REMK) from L_i and LS_i . If exactly one node remains in each of L_i and LS_i , then this becomes a new known correspondence, the nodes are placed on K and KS and removed from the elements of V and VS (by the function REMNEWK). Otherwise, the elements L_i and LS_i are placed on lists V and VS (by the function MERGESET), provided they do not contradict (no isomorphism) or duplicate any pair of elements (V_j, VS_j) .

After FF has considered all elements on L and LS, the length of lists K and KS determines the future action of the algorithm. If the number of correspondences (length of K) is equal to the number of nodes of graph G, an isomorphism is found and FF terminates with a value of 1. If the length of K is greater than the number of nodes of G then a contradiction must exist and no isomorphism will be possible. In such a case FF exits with a value of 0. Otherwise FF terminates with a value of 2 indicating that more work has to be done.

PAR - 0 arguments

PAR examines elements of V and VS and combines sets (using operations similar to union, intersection, and complement) to find single node correspondences. PAR returns 0 if an isomorphism is impossible, 1 if an isomorphism is found, and 2 if more work needs to be done.

In doing this, PAR takes each node of G and constructs the list of its possible correspondents by considering each pair of sets V_j and VS_j . Initially the known correspondents of node n are all the non-known nodes of graph GS. If node \underline{n} is in V_j then VS_j is intersected with the set of possible correspondents for node \underline{n} . If \underline{n} is not in V_j then the complement of VS_j is intersected with the set of possible correspondents for node \underline{n} . If the set of correspondents has length 1 then node \underline{n} becomes a known node and is added to list K, otherwise \underline{n} is added to V and its set of correspondents is added to VS.

SETFORM - 0 arguments

SETFORM uses set generating (property generating) functions to set L and LS to lists of corresponding nodes of G and GS. The properties are: gamma degree = 1, gamma degree = 2, node value and branch value.

NEWABLK - 0 arguments

NEWABLK adds a new assignment block to ASTACK. An assignment block is a list of the form (XS X SS V VS K KS) where the correspondence (assignment) X:XS was made from set SS which is part of LIST VS. Values of V, VS, K, and KS are prior to assignment so they can be restored if the assignment fails. X is added to ASLIST, and

lists K, KS, V, and VS are all updated using the new assignment.

ASSIGNI - 1 argument

A - a number or NIL

ASSIGNI finds a possible correspondent of node X in set SS. (Both X and SS are global to ASSIGNI, being set within NEWABLK and NEXTASSIGN which are the functions which can call ASSIGNI.) If A is not NIL, then it is the last correspondent used for X, and the search for a new correspondent starts with the successor of A in list SS. The new correspondence X:XS is checked for validity against the sets of V and VS.

NEXTASSIGN - 0 arguments

NEXTASSIGN makes the next assignment in the current block. The current block is the first element of ASTACK, which has kept track of all assignments made in the current attempt to locate an isomorphism.