

RESOURCE - RELATED RESEARCH

COMPUTERS AND CHEMISTRY

yes -
Keep original
file here.

(RR-00612 COMPETING RENEWAL APPLICATION)

Submitted to

BIOTECHNOLOGY RESOURCES BRANCH

OF THE

NATIONAL INSTITUTES OF HEALTH

May, 1976

site visit 1/7/77

DEPARTMENTS OF

CHEMISTRY, GENETICS, AND COMPUTER SCIENCE

STANFORD UNIVERSITY

SECTION I

| | | | |
|--|-----------------------|---------|---------------|
| DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE PUBLIC HEALTH SERVICE GRANT APPLICATION | LEAVE BLANK | | |
| | TYPE | PROGRAM | NUMBER |
| | REVIEW GROUP | | FORMERLY |
| | COUNCIL (Month, Year) | | DATE RECEIVED |

TO BE COMPLETED BY PRINCIPAL INVESTIGATOR (Items 1 through 7 and 15A)

1. TITLE OF PROPOSAL (Do not exceed 53 typewriter spaces)
RESOURCE-RELATED RESEARCH - COMPUTERS AND CHEMISTRY

2. PRINCIPAL INVESTIGATOR

2A. NAME (Last, First, Initial)
Djerassi, Carl

2B. TITLE OF POSITION
Professor of Chemistry

3. DATES OF ENTIRE PROPOSED PROJECT PERIOD (This application)
 FROM **5/1977** THROUGH **4/1982**

4. TOTAL DIRECT COSTS REQUESTED FOR PERIOD IN ITEM 3
\$1,463,940

5. DIRECT COSTS REQUESTED FOR FIRST 12-MONTH PERIOD
\$250,650

6. PERFORMANCE SITE(S) (See Instructions)
 Department of Genetics,
 Department of Chemistry, and
 Department of Computer Science
 Stanford University

2C. MAILING ADDRESS (Street, City, State, Zip Code)
 Department of Chemistry
 Stanford University
 Stanford, California 94305

2D. DEGREE
Ph.D.

2E. SOCIAL SECURITY NO.
 [REDACTED]

2F. TELEPHONE DATA
 Area Code **415**
 TELEPHONE NUMBER AND EXTENSION
 [REDACTED]

2G. DEPARTMENT, SERVICE, LABORATORY OR EQUIVALENT (See Instructions)
Department of Chemistry

2H. MAJOR SUBDIVISION (See Instructions)
Department of Humanities and Sciences

7. Research Involving Human Subjects (See Instructions)
 A. NO B. YES Approved: _____ Date _____
 C. YES - Pending Review

8. Inventions (Renewal Applicants Only - See Instructions)
 A. NO B. YES - Not previously reported
 C. YES - Previously reported

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

9. APPLICANT ORGANIZATION(S) (See Instructions)
 Stanford University
 Stanford, California 94305
 IRS No. 94-1156365
 Congressional District No. 12

10. NAME, TITLE, AND TELEPHONE NUMBER OF OFFICIAL(S) SIGNING FOR APPLICANT ORGANIZATION(S)
c/o Sponsored Projects Office
 Telephone Number (s) **(415) 497-2883**

11. TYPE OF ORGANIZATION (Check applicable item)
 FEDERAL STATE LOCAL OTHER (Specify)
Private, non-profit University

12. NAME, TITLE, ADDRESS, AND TELEPHONE NUMBER OF OFFICIAL IN BUSINESS OFFICE WHO SHOULD ALSO BE NOTIFIED IF AN AWARD IS MADE
K. D. Creighton
 Deputy Vice Pres. for Business & Finance
 Stanford University
 Stanford, California 94305
 Telephone Number **(415) 497-2251**

13. IDENTIFY ORGANIZATIONAL COMPONENT TO RECEIVE CREDIT FOR INSTITUTIONAL GRANT PURPOSES (See Instructions)
20 School of Humanities and Sciences

14. ENTITY NUMBER (Formerly PHS Account Number)
1941156365A1

15. CERTIFICATION AND ACCEPTANCE. We, the undersigned, certify that the statements herein are true and complete to the best of our knowledge and accept, as to any grant awarded, the obligation to comply with Public Health Service terms and conditions in effect at the time of the award.

| | | |
|---|---|----------------------|
| SIGNATURES (Signatures required on original copy only. Use ink, "Per" signatures not acceptable) | A. SIGNATURE OF PERSON NAMED IN ITEM 2A | DATE See page 55A |
| | B. SIGNATURE(S) OF PERSON(S) NAMED IN ITEM 10 | DATE |

BIOGRAPHICAL SKETCH

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

| | | |
|--|--|---|
| NAME CARL DJERASSI | TITLE Professor of Chemistry | BIRTHDATE (Mo., Day, Yr.) 10/29/23 |
| PLACE OF BIRTH (City, State, Country) Vienna, Austria | PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) U.S. citizen | SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female |

EDUCATION (Begin with baccalaureate training and include postdoctoral)

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

HONORS National Medal of Science ('73); Perkin Medal ('75); Am.Chem.Soc. Awards: Pure Chemistry ('58), Baekeland Medal ('59), Fritzsche Award ('60), Award for Creative Invention ('73); Freedman Found. Patent Award ('71) and Chem. Pioneer Award ('73) of Am.Inst.Chem.; Intrascience Res. Found. Award ('69); Hon. Member and Centenary Lecturer, Chem.Soc.(London);

| | | |
|---|--|-------------------|
| MAJOR RESEARCH INTEREST Natural Products Chemistry and chemical applications of physical methods | ROLE IN PROPOSED PROJECT Principal Investigator | (continued below) |
|---|--|-------------------|

RESEARCH SUPPORT (See instructions)

See attached.

HONORS (continued from above): Member of National Academy of Sciences, American Academy of Arts and Sciences, Royal Swedish Academy of Sciences, German Academy of Natural Scientists (Ieopoldina), Honorary D. Sc. Kenyon, Mexico, Rio de Janeiro, Worcester Polytechnic, Wayne State.

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

Academic Experience

Professor of Chemistry, Stanford University, 1959-present
Assoc. Professor ('52-'54) and Professor ('54-'59), Wayne State University

Industrial Experience

Zoecon Corp., Palo Alto, Calif. Chairman of the Board and Chief Exec. Officer, '68-present
Syntex Corp.: Various positions in Mexico City ('49-'52, '57-'60) and Palo Alto, Calif. ('60-'72) ranging from Assoc. Director of Chemical Research to President of Syntex Research
Ciba Pharmaceutical Co., Summit, N.J., Research Chemist, '42-'43, '45-'49.

Miscellaneous

Chairman of AAAS Gordon Res. Conf. on Steroids and Nat. Prod. ('52-'54). Member Amer. Pugwash Committee ('68-'75); Chairman, Latin American Science Board of National Academy of Sciences ('66-'68); Member ('68-'72) and Chairman ('73-'75) of Board on Science and Technology for International Development of National Academy of Sciences; Member, President's Advisory Group on Contributions of Technology to Economic Strength.

Publications

Author or co-author of six books (four dealing with organic mass spectrometry) and over 800 scientific publications. A selection of those dealing with mass spectrometry is given in the Bibliography.

RESEARCH SUPPORT: CARL DJERASSI

Agency: National Institutes of Health

Grant No.: GM-06840-18

Title of Grant: Marine Chemistry with Special Emphasis on Steroids

Period of Grant: 5/1/73-4/30/78

Current Budget: \$101,490

Fraction of time committed: 15%

Agency: National Institutes of Health

Grant No.: AM-04257

Title of Grant: Mass Spectrometry in Organic and Biochemistry

Period of Grant: 10/1/75-9/30/79

Current Budget: \$278,400

Fraction of time committed: 10%

BIOGRAPHICAL SKETCH

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

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

EDUCATION (Begin with baccalaureate training and include postdoctoral)

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

HONORS

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

| | |
|---|--|
| MAJOR RESEARCH INTEREST Molecular Genetics; Artificial Intelligence | ROLE IN PROPOSED PROJECT Investigator |
|---|--|

RESEARCH SUPPORT (See instructions)

Please see attached list.

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

1959 - present Professor and Chairman, Department of Genetics
Stanford University School of Medicine

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

1947 - 1957 Professor of Genetics
University of Wisconsin

Selected publications appear in Bibliography Section.

Privileged Communication - Section II

Lederberg, Joshua

RESEARCH SUPPORT

| GRANT NO. | TITLE OF PROJECT | CURRENT YEAR | PROJECT PERIOD | % OF EFFORT | GRANT AGENCY |
|--|--|------------------------|---------------------------|-------------|--------------|
| Dr. Lederberg: personal research commitments | | | | | |
| 5ROI CA16896-18 | Genetics of Bacteria | \$70,000 5/76-4/77 | \$195,000 5/74-4/77 | 15 | NIH |
| NAS1-9692 | Viking Mission Participation | \$42,500 1/76-6/76 | \$62,572 1/75-3/77 | 5 | NASA |
| Dr. Lederberg also functions as Principal Investigator ex officio on the following program-projects and training grants: | | | | | |
| NGR-05-020-632 | Analytical Methodology for Biochemical Monitoring | \$60,000 5/75-4/76 | \$180,000 5/73-4/76 | 2 | NASA |
| NO1 CB 43902 | Biomedical Markers that May Presage the Presence of Cancer | \$95,000 6/75-6/76 | \$183,108 6/74-6/76 | 5 | NIH |
| 3TOI GM00295 | Genetics Training Grant (graduate research training) | \$121,000 7/75-6/76 | \$916,637 7/74-6/79 | 10 | NIH |
| 1T22 GM00198-02 | Postdoctoral Training Medical Genetics | \$48,133 7/75-6/76 | \$144,133 7/74-6/77 | 5 | NIH |
| 1PO7 RR00785-03 | Stanford University Medical Experimental Computer: National Computer Resource for Research on AI in Medicine | \$358,000 8/75-7/76 | \$3,092,226 10/73-7/78 | 10 | NIH |
| NGR-05-020-004 | Instrumentation for Planetary Exploration | \$110,000 9/75-8/76 | \$110,000 9/75-8/76 | 5 | NASA |
| GM20832-02 | Genetics Research Project | \$241,432 5/76-4/77 | \$1,292,113 5/74-4/79 | 10 | NIH |

BIOGRAPHICAL SKETCH

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

| | | | |
|--|---|---|---------------------------|
| NAME EDWARD A. FEIGENBAUM | TITLE PROFESSOR OF COMPUTER SCIENCE | BIRTHDATE (Mo., Day, Yr.) 1-20-36 | |
| PLACE OF BIRTH (City, State, Country) Weehawken, New Jersey, U.S.A. | PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) U.S. citizen | SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female | |
| EDUCATION (Begin with baccalaureate training and include postdoctoral) | | | |
| INSTITUTION AND LOCATION | DEGREE | YEAR CONFERRED | SCIENTIFIC FIELD |
| Carnegie Institute of Technology Pittsburgh, Pennsylvania | B.S. | 1956 | Electrical Engineering |
| Carnegie Institute of Technology | Ph.D. | 1959 | Industrial Administration |
| HONORS | | | |
| MAJOR RESEARCH INTEREST Artificial Intelligence | | ROLE IN PROPOSED PROJECT Investigator | |
| RESEARCH SUPPORT (See instructions) | | | |

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

Stanford University, Stanford, California
 Chairman, Computer Science Department, 9/1976-
 Professor of Computer Science, 1969-
 Associate Professor of Computer Science, 1965-68
 Director, Stanford Computation Center, 1965-68
 University of California, Berkeley
 Associate Professor, School of Business Administration, 1964
 Assistant Professor, School of Business Administration, 1960-63
 Research Appointment, Center for Human Learning, 1961-64
 Research Appointment, Center for Research in Management Science, 1960-64
 Editor, Computer Science Series, McGraw-Hill Book Company, New York, 1965-
 Member, Computer and Biomathematical Sciences Study Section, National Institutes
 of Health, Bethesda, Maryland, 1968-72
 Ad-Hoc Mail Reviewer, National Science Foundation (various)

Selected Papers, 1965-76

1. J. Lederberg and E. A. Feigenbaum, "Mechanization of Inductive Inference in Organic Chemistry", in B. Kleinmuntz (ed.), *Formal Representations for Human Judgment*, (Wiley, 1968). (Also Stanford Artificial Intelligence Project Memo No. 54, August 1967).
2. J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference I. The Number of Possible Organic Compounds: Acyclic Structures Containing C, H, O and N". *Journal of the American Chemical Society*, 91:11 (May 21, 1969).
3. B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, "Toward an Understanding of Information Processes of Scientific Inference in the Context of Organic Chemistry", in *Machine Intelligence 5*, (B. Meltzer and D. Michie, eds.) Edinburgh University Press (1970). (Also Stanford Artificial Intelligence Project Memo No. 99, September 1969.)
4. E. A. Feigenbaum, B. G. Buchanan, and J. Lederberg, "On Generality and Problem Solving: A Case Study Using the DENDRAL Program". In *Machine Intelligence 6* (B. Meltzer and D. Michie, eds.) Edinburgh University Press (1971). (Also Stanford Artificial Intelligence Project Memo No. 131.)
5. B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, "A Heuristic Programming Study of Theory Formation in Science." In *Proceedings of the Second International Joint Conference on Artificial Intelligence*, Imperial College, London (September, 1971). (Also Stanford Artificial Intelligence Project Memo No. 145.)
6. B. G. Buchanan, E. A. Feigenbaum, and N. S. Sridharan, "Heuristic Theory Formation: Data Interpretation and Rule Formation". In *Machine Intelligence 7*, Edinburgh University Press (1973).
7. D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference X. INTSUM. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids". *Tetrahedron*, 29, 3117 (1973).
8. E. A. Feigenbaum, "Computer Applications: Introductory Remarks," in *Proceedings of Federation of American Societies for Experimental Biology*, Vol. 33, No. 12 (Dec., 1974) 2331-2332.

Other papers in Information Processing Psychology (18)

Books and Monographs

1. *Computers and Thought*, co-editor with Julian Feldman, McGraw-Hill, 1963.
2. *Information Processing Language V Manual*, Englewood Cliffs, N.J., Prentice-Hall, 1961 (with A. Newell, F. Tonge, G. Mealy, et.al.).
3. *An Information Processing Theory of Verbal Learning*, Santa Monica, The RAND Corporation Paper P-1817, October 1959 (monograph).

RESEARCH SUPPORT AND PENDING APPLICATIONS: Edward A. Feigenbaum

Agency: Advanced Research Projects Agency
Contract Number: DAHC 15 73 C 0435
Title of Contract: Heuristic Programming Project
Period of Contract: July 1975-June 1977
Annual Budget Level: \$203,000
Fraction of time committed: 40% Academic Yr.

Agency: National Science Foundation
Grant Number: MCS 76-11649
Title of Grant: MOLGEN: A Computer Science Application to Molecular Genetics
Period of Grant: 6/1/76-5/31/78
Annual Budget Level: \$110,700 (2 yr. amount)
Fraction of time committed: 10% Academic Yr.; 100% Summer

PENDING:

Agency: National Library of Medicine
Title: Training Program in Biomedical Computing
Period: 6/77-5/82
Annual Budget Level: \$334,193 (direct cost)
Fraction of time committed: 20%

BIOGRAPHICAL SKETCH

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

| | | |
|--|--|---|
| NAME BRUCE G. BUCHANAN | TITLE Adjunct Professor | BIRTHDATE (Mo., Day, Yr.) 7-7-40 |
| PLACE OF BIRTH (City, State, Country) St. Louis, Missouri, U.S.A. | PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) U.S. citizen | SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female |

EDUCATION (Begin with baccalaureate training and include postdoctoral)

| INSTITUTION AND LOCATION | DEGREE | YEAR CONFERRED | SCIENTIFIC FIELD |
|---------------------------|--------|----------------|------------------|
| Ohio Wesleyan University | B.A. | 1961 | Mathematics |
| Michigan State University | M.A. | 1966 | Philosophy |
| Michigan State University | Ph.D. | 1966 | Philosophy |

HONORS
Recipient of National Institutes of Health Career Development Award (1971-1976);
Invited Speaker: 1975 NATO Advanced Study Institute on Machine Representation of
Knowledge; 1974 Gordon Conference on Scientific Information Problems in Research.

| | |
|-------------------------|--|
| MAJOR RESEARCH INTEREST | ROLE IN PROPOSED PROJECT Associate Investigator |
|-------------------------|--|

RESEARCH SUPPORT (See instructions)

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

- 1976 - Adjunct Professor, Computer Science Department
Stanford University, Stanford, California
- 1972-1976 Research Computer Scientist, Computer Science Department
Stanford University, Stanford, California
- 1966-1971 Research Associate, Artificial Intelligence Project
Stanford University, Stanford, California

Selected Publications:

1. B. G. Buchanan, "Applications of Artificial Intelligence to Scientific Reasoning." In Proceedings of Second USA-Japan Computer Conference, August, 1975.
2. E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, "Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System," Computers and Biomedical Research, 8, 303-320 (1975).
3. E. H. Shortliffe and B. G. Buchanan, "A Model of Inexact Reasoning in Medicine", Mathematical Biosciences, 23, 351-379 (1975).
4. D. Michie and B. G. Buchanan, "The Scientist's Apprentice" in Computers for Spectroscopy (ed. R.A.G. Carrington) London: Adam Hilger, 1974.
5. B. G. Buchanan and N. S. Sridharan, "Rule Formation on Non-Homogeneous Classes of Objects". Proceedings of the Third International Joint Conference on Artificial Intelligence (1973).
6. D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi, and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference X. Intsum. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids". Tetrahedron, 29, 3117 (1973).
7. D. H. Smith, B. G. Buchanan, R. S. Engelmores, H. Aldercreutz and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference IX. Analysis of Mixtures Without Prior Separation as Illustrated for Estrogens". Journal of the American Chemical Society, 95, 6078, 1973.
8. B. G. Buchanan, Review of Hubert Dreyfus' "What Computers Can't Do: A Critique of Artificial Reason", Computing Reviews (January, 1973).
9. B. G. Buchanan, E. A. Feigenbaum and N. S. Sridharan, "Heuristic Theory Formation: Data Interpretation and Rule Formation". Machine Intelligence 7, Edinburgh University Press (1972).
10. C. W. Churchman and B. G. Buchanan, "On the Design of Inductive Systems: Some Philosophical Problems". British Journal for the Philosophy of Science, 20 (1969), 311-323.
11. B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, "Toward an Understanding of Information Processes of Scientific Inference in the Context of Organic Chemistry", Machine Intelligence 5 (B. Meltzer and D. Michie, eds.), Edinburgh University Press (1970). (Also Stanford Artificial Intelligence Project Memo No. 99, September 1969.)

RESEARCH SUPPORT AND PENDING APPLICATIONS: Bruce G. Buchanan

Agency: Advanced Research Projects Agency
Contract Number: DAHC 15 73 C 0435
Title of Contract: Heuristic Programming Project
Period of Contract: July 1975-June 1977
Annual Budget Level: \$203,000
Fraction of time committed: 25%

Agency: National Science Foundation
Grant Number: MCS 76-11649
Title of Grant: MOLGEN: A Computer Science Application to Molecular Genetics
Period of Grant: 6/1/76-5/31/78
Annual Budget Level: \$110,700 (2 yr. amount)
Fraction of time committed: 25%

PENDING:

Agency: National Library of Medicine
Title: Training Program in Biomedical Computing
Period: 6/77-5/82
Annual Budget Level: \$334,193 (direct cost)
Fraction of time committed: 20%

BIOGRAPHICAL SKETCH

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

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

RESEARCH SUPPORT (See instructions)

N/A

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

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

Publications: See attached list.

1. H.G. Langer, R.S. Gohlke, and D.H. Smith, "Mass Spectrometric Differential Thermal Analysis," Anal. Chem., 37, 433 (1965).
2. S.M. Kupchan, J.M. Cassady, J.E. Kelsey, H.K. Schnoes, D.H. Smith, and A.L. Burlingame, "Structural Elucidation and High Resolution Mass Spectrometry of Gaillardin, a New Cytotoxic Sesquiterpene Lactone," J. Amer. Chem. Soc., 88, 5292 (1966).
3. D.H. Smith, Ph.D. Thesis, "High Resolution Mass Spectrometry: Techniques and Applications to Molecular Structure Problems," Dept. of Chemistry, University of California, Berkeley, California (1967).
4. H.K. Schnoes, D.H. Smith, A.L. Burlingame, P.W. Jeffs, and W. Dopke, "Mass Spectra of Amaryllidaceae Alkaloids: The Lycorenine Series," Tetrahedron, 24, 2825 (1968).
5. A.L. Burlingame, D.H. Smith, and R.W. Olsen, "High Resolution Mass Spectrometry in Molecular Structure Studies. XIV. Real-time Data Acquisition, Processing and Display of High Resolution Mass Spectral Data," Anal. Chem., 40, 13 (1968).
6. A.L. Burlingame and D.H. Smith, "High Resolution Mass Spectrometry in Molecular Structure Studies. II. Automated Heteroatomic Plotting as an Aid to the Presentation and Interpretation of High Resolution Mass Spectral Data," Tetrahedron, 24, 5749 (1968).
7. W.J. Richter, B.R. Simoneit, D.H. Smith, and A.L. Burlingame, "Detection and Identification of Oxocarboxylic and Dicarboxylic Acids in Complex Mixtures by Reductive Silylation and Computer-Aided Analysis of High Resolution Mass Spectral Data," Anal. Chem., 41, 1392 (1969).
8. The Lunar Sample Preliminary Examination Team, "Preliminary Examination of Lunar Samples from Apollo 11," Science, 165, 1211 (1969).
9. S.M. Kupchan, W.K. Anderson, P. Bollinger, R.W. Doskotch, R.M. Smith, J.A. Saenz-Renaud, H.K. Schnoes, A.L. Burlingame, and D.H. Smith, "Tumor Inhibitors. XXXIX. Active Principles of Acnistus arborescens. Isolation and Structural and Spectral Studies of Withaferin A and Withacnistin," J. Org. Chem., 34, 3858 (1969).
10. A.L. Burlingame, D.H. Smith, T.O. Merren, and R.W. Olsen, "Real-time High Resolution Mass Spectrometry," in Computers in Analytical Chemistry (Vol. 4 in Progress in Analytical Chemistry series), C.H. Orr and J. Norris, Eds., Plenum Press, New York, 1970, pp. 17.
11. The Lunar Sample Preliminary Examination Team, "Preliminary Examination of Lunar Samples from Apollo 12," Science, 167, 1325 (1970).
12. D.H. Smith, "Mass Spectrometry," Chapter X in Guide to Modern Methods of Instrumental Analysis, T.M. Gow, Ed., Wiley-Interscience, New York, 1972.
13. D.H. Smith, R.W. Olsen, F.C. Walls, and A.L. Burlingame, "Real-time Mass Spectrometry: LOGOS--A Generalized Mass Spectrometry Computer System for High and Low Resolution, GC/MS and Closed-Loop Applications," Anal. Chem., 43, 1796 (1971).

14. A.L. Burlingame, J.S. Hauser, B.R. Simoneit, D.H. Smith, K. Biemann, N. Mancuso, R. Murphy, D.A. Flory, and M.A. Reynolds, "Preliminary Organic Analysis of the Apollo 12 Cores," Proceedings of the Apollo 12 Lunar Science Conference, E. Levinson, Ed., M.I.T. Press, Cambridge, Mass., 1971, p. 1891.
15. D.H. Smith, "A Compound Classifier Based on Computer Analysis of Low Resolution Mass Spectral Data," Anal. Chem., 44, 536 (1972).
16. D.H. Smith and G. Eglinton, "Compound Classification by Computer Treatment of Low Resolution Mass Spectra-Application to Geochemical and Environmental Problems," Nature, 235, 325 (1972).
17. D.H. Smith, N.A.B. Gray, C.T. Pillinger, B.J. Kimble, and G. Eglinton, "Complex Mixture Analysis - Geochemical and Environmental Applications of a Compound Classifier Based on Computer Analysis of Low Resolution Mass Spectra," Adv. in Org. Geochem., 1971, p. 249.
18. P. Longevialle, D.H. Smith, H.M. Fales, R.J. Highet, and A.L. Burlingame, "High Resolution Mass Spectrometry in Molecular Structure Studies. V. The Fragmentation of Amaryllis Alkaloids in the Crinine Series," Org. Mass. Spectrom., 7, 401 (1973).
19. B.R. Simoneit, D.H. Smith, G. Eglinton, and A.L. Burlingame, "Applications of Real-time Mass Spectrometric Techniques to Environmental Organic Geochemistry. II. San Francisco Bay Area Waters," Arch. Env. Contam. Tox., 1, 193 (1973).
20. G. Loew, M. Chadwick, and D.H. Smith, "Applications of Molecular Orbital Theory to the Interpretation of Mass Spectra. Prediction of Primary Fragmentation Sites in Organic Molecules," Org. Mass Spectrom., 7, 1241 (1973).
21. J.H. Block, D.H. Smith and C. Djerassi, "Mass Spectrometry in Structural and Stereochemical Problems, CCXXXVIII. The Effect of Heteroatoms upon the Mass Spectrometric Fragmentation of Cyclohexanones," J. Org. Chem., 39, 279 (1974).
22. D.H. Smith, C. Djerassi, K.H. Maurer, and U. Rapp, "Mass Spectrometry in Structural and Stereochemical Problems. CCXLII. Analysis of Mixtures Based on the Distribution of Fragment Ions Arising from Unimolecular Decomposition of Metastable Parent Ions," J. Amer. Chem. Soc., 96, 3482 (1974).
23. D.H. Smith, "The Scope of Structural Isomerism," J. Chem. Inf. Comp. Sci., 15, 203 (1975).
24. B.R. Simoneit, D.H. Smith, and G. Eglinton, "Application of Real-Time Mass Spectrometric Techniques to Environmental Organic Geochemistry. I. Computerized High Resolution Mass Spectrometry and Gas Chromatography-Low Resolution Mass Spectrometry," Arch. Environ. Cont. Tox., 3, 385 (1976).
25. T.R. Varkony, R.E. Carhart, and D.H. Smith, "Computer-Assisted Structure Elucidation. Modelling Chemical Reaction Sequences Used in Molecular Structure Problems," in "Computer-Assisted Organic Synthesis," W.T. Wipke, Ed., American Chemical Society, Washington, D.C., in press.

26. D.H. Smith and R.E. Carhart, "Structural Isomerism of Mono- and Sesquiterpenoid Skeletons," Tetrahedron, in press.
27. L.L. Dunham, C.A. Henrick, D.H. Smith, and C. Djerassi, "Mass Spectrometry in Structural and Stereochemical Problems. CCXLVI. Electron Impact Induced Fragmentation of Juvenile Hormone Analogs," Org. Mass Spectrom., in press.

See also Bibliography.

BIOGRAPHICAL SKETCH

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

| | | |
|---|--|---|
| NAME RAYMOND EDGAR CARHART | TITLE RESEARCH ASSOCIATE | BIRTHDATE (Mo., Day, Yr.) 10/4/46 |
| PLACE OF BIRTH (City, State, Country) Evanston, Illinois, U.S.A. | PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) U.S. citizen | SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female |

EDUCATION (Begin with baccalaureate training and include postdoctoral)

| INSTITUTION AND LOCATION | DEGREE | YEAR CONFERRED | SCIENTIFIC FIELD |
|------------------------------------|--------|----------------|----------------------------|
| Northwestern University | B.A. | 1968 | Chemistry |
| California Institute of Technology | Ph.D. | 1973 | Physical Organic Chemistry |

HONORS

Phi Beta Kappa; Sigma Xi; Phi Lambda Upsilon; NSF pre-doctoral fellowship 1968-72; NIH post-doctoral fellowship 1972-74.

| | |
|--|--|
| MAJOR RESEARCH INTEREST Applications of Computer Science to Organic Chemistry | ROLE IN PROPOSED PROJECT Research Associate |
|--|--|

RESEARCH SUPPORT (See instructions)

N/A

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

1974- Research Associate, Department of Computer Science, Stanford University
 1972-1974 NIH Post-doctoral Fellow, Department of Computer Science, Stanford University
 1969(summer) Visiting Scientist, IBM Research Laboratory, San Jose, California

Recent Publications:

R. Carhart and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference XI: The Analysis of C13 NMR Data for Structure Elucidation of Acyclic Amines", Journal of the Chemical Society (Perkin II), 1753 (1973).

L. Masinter, N.S. Sridharan, R. Carhart and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XIII. Labeling of Objects Having Symmetry". Journal of the American Chemical Society, 96, 7714 (1974).

R. E. Carhart, D. H. Smith, H. Brown and N. S. Sridharan, "Applications of Artificial Intelligence for Chemical Inference. XVI. Computer Generation of Vertex Graphs and Ring Systems". Journal of Chemical Information and Computer Science, 15, 124 (1975).

R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XVII. An Approach to Computer-Assisted Elucidation of Molecular Structure". Journal of the American Chemical Society, 97, 5755 (1975).

R. E. Carhart, S. M. Johnson, D. H. Smith, B. G. Buchanan, R. G. Dromey, J. Lederberg, "Networking and a Collaborative Research Community: A Case Study Using the DENDRAL Program," in "Computer Networking and Chemistry", P. Lykos, Ed., American Chemical Society, Washington, D.C., 1975, p. 192.

R. E. Carhart and D. H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XX. 'Intelligent' Use of Constraints in Computer-Assisted Structure Elucidation," Computers in Chemistry, in press.

T. R. Varkony, R. E. Carhart, and D. H. Smith, "Computer-Assisted Structure Elucidation. Modelling Chemical Reaction Sequences Used in Molecular Structure Problems," in "Computer-Assisted Organic Synthesis," W. T. Wipke, Ed., American Chemical Society, Washington, D.C., in press.

D. H. Smith and R. E. Carhart, "Structural Isomerism of Mono- and Sesquiterpenoid Skeletons," Tetrahedron, in press.

R. E. Carhart, "A Model-Based Approach to the Teletype Printing of Chemical Structures," Journal of Chemical Information and Computer Science, in press.

BIOGRAPHICAL SKETCH

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

| | | | |
|--|---|---|--------------------------------------|
| NAME Gretchen Maria SCHWENZER | TITLE Research Associate | BIRTHDATE (Mo., Day, Yr.) 2/6/49 | |
| PLACE OF BIRTH (City, State, Country) Buffalo, New York, U.S.A. | PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) U.S. | SEX <input type="checkbox"/> Male <input checked="" type="checkbox"/> Female | |
| EDUCATION (Begin with baccalaureate training and include postdoctoral) | | | |
| INSTITUTION AND LOCATION | DEGREE | YEAR CONFERRED | SCIENTIFIC FIELD |
| State University of New York at Buffalo University of California, Berkeley Institute in Quantum Chemistry, Solid State Physics & Quantum Biology, Uppsala, Sweden | B.A. Ph.D. | 1971 1975 Summer 1973 | Mathematics & Chemistry Chemistry |
| HONORS Phi Beta Kappa, Pi Mu Epsilon, Alpha Lambda Delta Graduated Magna Cum Laude with Highest Distinction; Allied Chemical Scholar, 1971; Award of American Institute of Chemists for Scholastic Achievement. | | | |
| MAJOR RESEARCH INTEREST Application of Computers in Chemistry | ROLE IN PROPOSED PROJECT Direct C13 NMR with attention to the structural nature of the problem | | |
| RESEARCH SUPPORT (See instructions) | | | |

N/A

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

Stanford University 1976 -
Computer Science Department, Stanford, Calif.

IBM, San Jose Research Division, San Jose, Calif. 1975

University of California, Berkeley, Calif. 1971-1975
Thesis: The Excited Electronic States of HCN and HNC; a New Method to Obtain Wave Functions of SCF Quality
Configuration Interaction Wave Functions to Obtain Optimized Minimum Basis Set Potential Surfaces

State University of New York at Buffalo, Buffalo, N.Y.

"Photochemical Substitution Reactions of Substituted Group VI Metal Carbonyls," G. Schwenger, M.Y. Darensbourg, and D.J. Darensbourg, Inorganic Chemistry, 11, 1967 (1972).

"Photochemical Substitution Reactions of Substituted Group VI Metal Carbonyls," G. Schwenger, D.J. Darensbourg, M.Y. Darensbourg, ACS Meeting, New York, Aug. (1972).

"Use of nonrelativistic wavefunctions for the prediction of properties of molecules containing atoms of high Z. PbO as a test case," Gretchen M. Schwenger, Dean H. Liskow, and Henry F. Schaefer, The Journal of Chemical Physics, Vol. 58, No. 8, 15 April 1973.

"Geometries of the excited electronic states of HCN," Gretchen M. Schwenger, Stephen V. O'Neil, and Henry F. Schaefer, The Journal of Chemical Physics, Vol. 60, No. 7, 1 April 1974.

"The Hypervalent Molecules Sulfurane (SH₄) and Persulfurane (SH₆)," Gretchen M. Schwenger and Henry F. Schaefer, The Journal of the American Chemical Society, 97, 1393 (1975).

"Excited Electronic States of HNC, Hydrogen Isocyanide," Gretchen M. Schwenger, Henry F. Schaefer, and Charles F. Bender, The Journal of Chemical Physics, 63, 569 (1975).

"Confirmation of the Discrepancy Between Theory and Experiment for the B¹A" state of HCN," Gretchen M. Schwenger, Henry F. Schaefer and Charles F. Bender, Chemical Physics Letters, Vol. 36, No. 2, 179 (1975).

"Documentation of ALCHEMY", Gretchen M. Schwenger, IBM Report.

BIOGRAPHICAL SKETCH

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

| | | |
|--|---|---|
| NAME HAROLD D. BROWN | TITLE Research Associate | BIRTHDATE (Mo., Day, Yr.) 7-12-34 |
| PLACE OF BIRTH (City, State, Country) South Bend, Indiana, U.S.A. | PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date) U.S. citizen | SEX <input checked="" type="checkbox"/> Male <input type="checkbox"/> Female |

EDUCATION (Begin with baccalaureate training and include postdoctoral)

| INSTITUTION AND LOCATION | DEGREE | YEAR CONFERRED | SCIENTIFIC FIELD |
|--------------------------|--------|----------------|------------------|
| University of Notre Dame | M.Sc. | 1963 | Mathematics |
| Ohio State University | Ph.D. | 1966 | Mathematics |

HONORS

| | |
|-------------------------|--|
| MAJOR RESEARCH INTEREST | ROLE IN PROPOSED PROJECT Research Associate |
|-------------------------|--|

RESEARCH SUPPORT (See instructions)

Pending, "Computer-Assisted Molecular Structure Elucidation", 12-month grant.
Proposed Amount: \$42,733
Period: 11/1/75-10/31/77
Source: National Science Foundation

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

1971-72 Associate Professor, Computer Science Department, Stanford University
1973-

1973- Research Associate, Medical School, Stanford University

1963-75 Instructor/Assistant Professor, Assistant Chairman/Associate Professor, Mathematics, The Ohio State University

Winter 1971, 73 and 75 Visiting Professor, Mathematics, Rhine. Westf. Tech. Hoch., Aschen

1964-70 Director/Associate Director, National Science Foundation SSTP

1967-68 Visiting Member, Courant Institute of Mathematical Sciences, New York University

1960-63 Assistant to the Chairman, Mathematics, University of Notre Dame

Publications:

Near Algebras, Ill. J. Math. 12(1968), p. 215.

Distributor Theory in Near Algebras, Comm. Pure Appl. Mat. XXI(1968), p. 535.

An Algorithm for the Determination of Space Groups, Math. Comp. 23(1969), p. 499.

Some Empirical Observations on Primitive Roots (with H. Zassenhaus), J. Number Theory 3(1971), p. 306.

A Generalization of Farey Sequences (with K. Mahler), J. Number Theory 3(1971), p. 364.

Basic Computations for Orders, Stanford CS Report STAN-CS-72-208.

An Application of Zassenhaus' Unit Theorem, Acta Arith. XX(1972), p. 154.

Integral Groups I: The Reducible Case (with J. Neubuser and H. Zassenhaus), Numer. Math. 19(1972), p. 386.

Integral Groups II: The Irreducible Case (with J. Neubuser and H. Zassenhaus), Numer. Math. 20(1972), p. 22.

Integral Groups III: Normalizers (with J. Neubuser and H. Zassenhaus), Math. Comp. 27(1973), p. 167.

Constructive Graph Labeling via Double Cosets (with L. Hjelmeland and L. Masinter), Discrete Math. 7(1973), p. 1; and Stanford CS Report STAN-CS-72-318.

An Algorithm for the Construction of the Graphs of Organic Molecules (with L. Masinter), Discrete Math. 8(1974), p. 227; and Stanford CS Report STAN-CS-73-261.

The Crystallographic Groups of 4-dimensional Space (with J. Neubuser, H. Wondratschek and H. Zassenhaus), Wiley Interscience (in preparation).

Molecular Structure Elucidation III: Fragment Embedding, Soc. Industrial and Applied Math. J. on Computing (submitted), and Stanford CS Report STAN-CS-74-469.

Applications of Artificial Intelligence for Chemical Inference XVII. Computer Generation of Vertex Graphs and Ring Systems (with R. Carhart, N. Sridharan and D. Smith), J. Chem. Inf. Comp. Sci. (in press).

Applications of Artificial Intelligence for Chemical Inference XVIII. An Approach to Computer-Assisted Elucidation of Molecular Structure (with R. Carhart and D. Smith), JACS (in press).

Table of Contents

| Section | | Page |
|---------|--|------|
| | Subsection | |
| 1. | Introduction | 1 |
| | 1.1 Objectives | 1 |
| | 1.2 Background and Rationale | 2 |
| | 1.3 Existing Capabilities | 4 |
| | 1.4 Relationship to Mass Spectrometry and AIM-SUMEX Resources | 10 |
| 2. | Specific Aims | 11 |
| | 2.1 Add More "Intelligence" to Existing Programs | 11 |
| | 2.2 Develop New Computer Programs that Assist in Biomolecular Structure Elucidation | 12 |
| | 2.3 Develop New Programs that Aid in Rule Formation | 13 |
| | 2.4 Apply the Structure Elucidation Programs and GC/HRMS | 14 |
| | 2.5 Increase the Availability of the Structure Elucidation Techniques | 14 |
| | 2.6 Maintain and Improve the GC/HRMS System | 14 |
| 3. | Methods | 15 |
| | 3.1 Extra Intelligence in Existing Programs | 15 |
| | 3.2 New Programs for Structure Elucidation | 18 |
| | 3.3 New Programs for Theory Formation | 30 |
| | 3.4 Applications | 35 |
| | 3.5 Increased Availability | 42 |

Table of Contents

| | | |
|-----|--------------------------------|----|
| 3.6 | The GC/HRMS Resource | 46 |
| 4. | BIBLIOGRAPHY | 48 |
| 5. | Appendix I | 56 |
| 6. | Appendix II | 61 |

III. Research Plan

1 Introduction

1.1 Objectives

Our principal objectives are

- A. Developing an integrated approach to computer-assisted elucidation of biomolecular structures; and
- B. Applying the techniques of computer-assisted structure elucidation to a wide range of biomolecular structural problems.

To those ends, we will endeavor:

To extend the current DENDRAL programs and write new programs

To explore other aspects of the structure elucidation problem.

To provide the capability for general analysis of mass spectra and C13 NMR spectra.

To extend the capability of the programs to deal with structural inferences derived from many sources of data, including our existing combined gas chromatography/high resolution mass spectrometry (GC/HRMS) resource, other spectroscopic techniques (e.g., ¹H NMR, IR, UV), chemical reactions applied to the sample and other physical or chemical measurements.

To continue to design these programs to be widely disseminated tools for working laboratory scientists.

To apply our programs to a wide variety of biologically interesting problems selected from our laboratories and those of our collaborators.

To serve our present community of collaborators and extend that community by: a) developing more intelligent and helpful interfaces to programs to make them easier to use; b) soliciting additional users of our programs on SUMEX, either directly via computer networks or indirectly by solving problems sent to us by persons who do not have access; c) making the programs more transportable so others can gain access on machines besides SUMEX.

Application of our techniques also requires some improvements and maintenance of the GC/HRMS system so that users

of this resource can have more routine access to the system. The entire proposal reflects diminished emphasis on new developments in the GC/HRMS data acquisition and reduction system and increased emphasis on problem-solving programs for more general applications to structure elucidation.

1.2 Background and Rationale

1.2.1 The Structure Elucidation Problem

The elucidation of molecular structures is fundamental to the application of chemical knowledge to areas of critical importance to biology and medicine. Areas where we and our collaborators maintain active interest include: a) identification of natural products isolated from terrestrial or marine sources, particularly those products which demonstrate biological activity or which are key intermediates in biosynthetic pathways; b) verification of the identity of new synthetic materials; c) identification of drugs and their metabolites in clinical studies; and d) detection of metabolic disorders of genetic, developmental, toxic or infectious origins by identification of organic constituents excreted in abnormal quantities in human body fluids.

Structure elucidation can be accomplished in one of two ways. X-ray crystallography is now automated to a point where it can be considered relatively routine. A successful analysis of molecular structure using x-ray crystallographic techniques requires, however, that: 1) a sufficient quantity of material exists; and 2) the material can be crystallized. In most circumstances, however, especially in the areas of interest summarized above, we are faced with structural problems where sufficient material is not available and/or the material cannot be crystallized. In these circumstances we must resort to structure elucidation based on data obtained from a variety of physical, chemical and spectroscopic methods.

The latter approach involves a sequence of steps which is roughly approximated by Figure 1. An unknown structure is isolated from some source. The source of the sample and the isolation procedures employed already provide some clues as to the chemical constitution of the compound. A variety of chemical, physical and spectroscopic data are collected on the sample. Interpretation of these data yields structural hypotheses in the form of functional groups or more complex molecular fragments. Assembling these fragments into complete structures provides a set of candidate structures for the unknown. These candidates are examined and experiments are designed to differentiate among them. The experiments, usually collecting additional spectroscopic data and executing sequences of chemical reactions, result in new structural hypotheses which

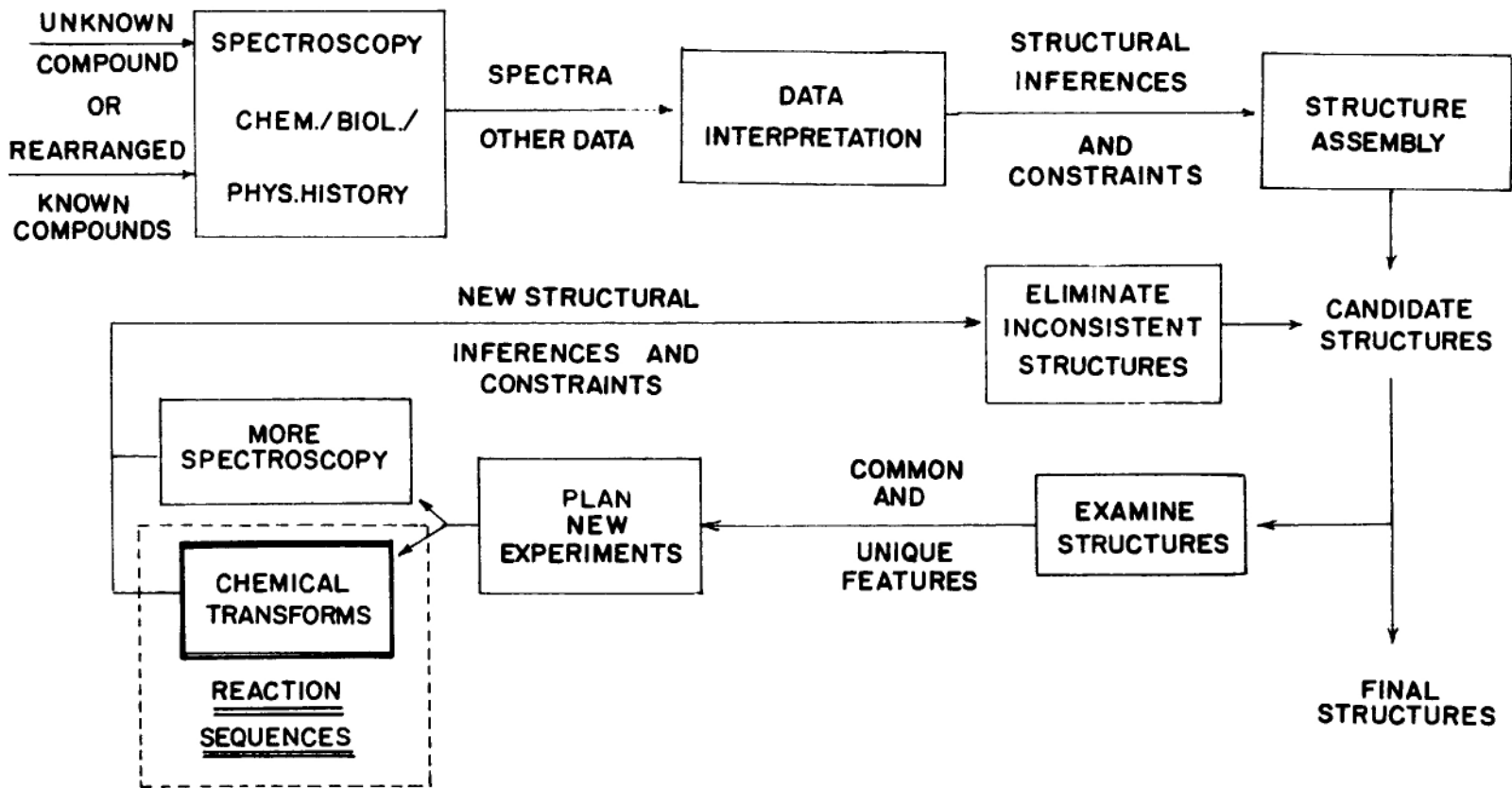


Figure 1

serve to reduce the set of candidate structures, eventually yielding the correct structure.

This approach to structure elucidation has been carried out manually since the beginnings of chemistry as a science. As long as time permits and the number of unknown structures is small, a manual approach will usually be successful. In our opinion, however, the manual approach is amenable to a high degree of computer assistance. Such assistance is increasingly necessary for both practical and scientific reasons. One need only examine current regulatory activities in fields related to chemistry, or the rate at which new compounds are discovered or synthesized to gain a feeling for the practical need for rapid identification of new structures. More importantly, however, is the contribution such computer assistance can make to scientific creativity in structure elucidation in particular and chemistry in general. The automated approaches discussed in this proposal provide a systematic procedure for verifying hypotheses about chemical structure and ensuring that no plausible alternatives have been overlooked.

In our experience, because the user of DENDRAL computer programs is in control of the program, or can at least determine why certain steps were taken, our programs are valuable assistants and foster creativity in at least two ways. The programs suggest alternatives to personal biases which must be accepted or rejected on experimental grounds. Also, the programs have been designed to work with problem solving scientists, to perform the combinatorial tasks that humans find tedious and difficult. These advantages will be elaborated below.

This proposal has as its primary focus the development of high performance programs for computer-assisted structure elucidation. One current program, CONGEN (48), is designed to perform structure assembly, under constraints, based on the structural inferences derived by a user, and to provide some capabilities for examining the candidate structures and removing undesired structures based on new data (Fig. 1). Part of our proposal is to increase the power of CONGEN to improve its performance and make it easier to use in order to promote widespread dissemination of the program to other researchers. A second part is to provide additional programs to perform other tasks outlined in Fig. 1, including some automated examination of experimental data, experiment planning and chemical reaction sequences. Operating together, these programs will provide tools for structure elucidation that will, in our opinion, eventually become as routinely used as conventional spectroscopic methods.

1.2.2 Historical Background

This work was begun over ten years ago as an ARPA-sponsored project exploring scientific inference by computers, together with NASA-sponsored work on GC/MS instrumentation for a planned

automated planetary lander laboratory. At that time we were mostly concerned with the conceptual problems of designing and writing complex symbol manipulation programs containing any scientific knowledge at all. As the programs developed we began to see that we could make them flexible enough to accommodate more and more knowledge of chemistry and mass spectrometry.

Initial funding by the NIH (1971-74) provided the opportunity to add the specific knowledge needed for serious biomedical research problems. In addition, it provided significant improvements in the instrumentation that could be used for structure elucidation problems. Continuation of NIH funding for 1974-77 allowed substantial progress on bringing the computer programs and instrumentation into service on structure elucidation problems of biomedical interest. The last annual report of progress (for 1975-76) is appended to this proposal for more background (Appendix II). It shows the extent to which NIH funding has provided new, sophisticated tools for working biomedical scientists as well as the responsiveness of the DENDRAL project to the goal of sharing the fruits of this research.

Initially our focus was entirely on mass spectrometry, first as a means of demonstrating that a computer could interpret any scientific data and then as a tool for structure elucidation. Some of the programs have been extended beyond mass spectrometry; other programs have yet to be generalized.

Our programs have followed an evolutionary progression. Initial concepts were translated into a working program, the program was tested and improved by confronting simple test cases and finally a production version of the program including user interaction facilities was released for real applications. We expect this progression to continue with our current and proposed efforts. This intertwining of short-term pragmatic goals and long-term development of new science is an important theme throughout this proposal.

1.3 Existing Capabilities

1.3.1 CONGEN

The CONGEN (48) program represents a significant extension of a program which has developed over the last several years, the cyclic structure generator (40,41). The purpose of CONGEN is to assist the chemist in determining the chemical structure of an unknown compound by 1) allowing him to specify certain types of structural information about the compound which he has determined from any source (e.g., spectroscopy, chemical degradation, method of isolation, etc.) and 2) generating an exhaustive and non-redundant list of structures that are consistent with the information. The generation is a stepwise process, and the

program allows interaction at every stage: based upon partial results the chemist may be reminded of additional information which he can specify, thus limiting further the number of structural possibilities.

At the heart of CONGEN are two algorithms whose accuracy has been mathematically proven and whose computer implementation has been well tested. The structure generation algorithm (31,40,41) is designed to determine all topologically unique ways of assembling a given set of atoms, each with an associated valence, into molecular structures. The atoms may be chemical atoms with standard chemical valences, or they may be names representing molecular fragments ("superatoms") of any desired complexity, where the valence corresponds to the total number of bonding sites available within the superatom. Because the structure generation algorithm can produce only structures in which the superatoms appear as single nodes (we refer to these as intermediate structures), a second procedure, the imbedding algorithm (37,48) is needed to expand the superatoms to their full chemical identities.

A substantial amount of effort has been devoted to modifying these two basic procedures, particularly the structure generation algorithm, to accept a variety of other structural information (constraints), using it to prune the list of structural possibilities. Current capabilities include specification of good and bad substructural features, good and bad ring sizes, proton distributions and connectivities of isoprene units (62). Usually, the chemist has additional information (if only some general rules about chemical stability, of which the program has no concept) that can be used to limit the number of structural possibilities. For example, he may know that because of a compound's stability, it cannot contain a peroxide linkage (O-O) and thus the program need not consider such structures when there are two or more oxygens in the "building block" list.

To make CONGEN accessible to research chemists, the program has been provided with an easily used, interactive "front end". This interface contains EDITSTRUC, an interactive structure editor, DRAW, a teletype-oriented structure display program, and the CONGEN "executive" program which ties together the individual subprograms and aids the user with various tasks, such as defining superatoms and substructures, creating and editing lists of constraints or superatoms, and saving and restoring superatoms, constraints and structures from secondary storage (disc). The resulting system, for which comprehensive user-level documentation has been prepared, is running on the SUMEX computing facility and is available nationwide over the TYMNET and ARPANET networks. Several researchers are currently using CONGEN to assist them in structure elucidation problems.

1.3.2 Meta-DENDRAL

The present Meta-DENDRAL program (56) interactively helps chemists determine the dependence of mass spectrometric fragmentation on substructural features, under the hypothesis that molecular fragmentations are related to topological graph structural features of molecules. Our goal is to have the program suggest qualitative explanations of the characteristic fragmentations and rearrangements among a set of molecules. We do not now attempt to rationalize all peaks nor find quantitative assessments of the extent to which various processes contribute to peak intensities.

The program emulates many of the reasoning aspects of manual approaches to rule discovery. It reasons symbolically, using a modest amount of chemical knowledge. It decides which data points are important and looks for fragmentation processes that will explain them. It attempts to form general rules by correlating plausible fragmentation processes with substructural features of the molecules. Then, as a chemist does, the program tests and modifies the rules.

The Meta-DENDRAL program is organized as three subprograms called INTSUM, RULEGEN and RULEMOD.

The INTSUM program (named for data interpretation and summary) interprets spectral data of known compounds in terms of possible bond cleavages. For each molecule in a given set, INTSUM first produces the plausible bond cleavage processes which might occur, i.e., breaks and combinations of breaks, with and without transfer of hydrogens and other neutral species. These processes are associated with specific bonds in a portion of molecular structure, or skeleton, that is chosen because it is common to the molecules in the given set. Then INTSUM examines the spectra of the molecules looking for evidence (spectral peaks) for each process.

Because INTSUM does not recognize that different cleavages (of the skeleton or substituents) may represent fragmentation processes which are similar Meta-DENDRAL next attempts to correlate the fragmentations with substructural features of molecules. The RULEGEN program is a generator of plausible rules. Based on guidance from the INTSUM interpretation of the mass spectra, the rule generator searches a space of rules. It starts from the most general hypothesis "every bond breaks" and systematically searches ways of making the hypothesis more specific. It does this by adding descriptive features, one at a time, to the subgraphs that define the environments of cleavages. For example, the types of atoms in the subgraph may be important, or the degree of substitution. These features, and others, are added to the nodes of an expanding subgraph in ways that fit the improvement criteria of the RULEGEN program. As long as the expanded rule is an improvement over its more general parent, the search for better rules continues. Each time the program's

search for an improvement comes to an end, the program writes the candidate rule on a file and tries the next likely path.

RULEMOD, the third of the Meta-DENDRAL subprograms, tests and modifies the rules that are produced by the generator. There are many ways to improve the rules: the most important are to make them more specific to get rid of counterexamples and to merge pairs of similar rules. This step can be thought of as "fine-tuning" the candidate rules to improve their explanatory power and to reduce the total number of rules.

1.3.3 Gas Chromatography/High Resolution Mass Spectrometry Resource

A major portion of the previous proposal on which this renewal is based was for development of a combined GC/HRMS system. This system is designed to provide data on empirical formulas of molecular ions and fragmentation products thereof, recorded from the effluent of a gas chromatograph used to separate complex mixtures. These data are critical to many of our structure elucidation problems: problems which involve complex mixtures of closely related compounds such as encountered in the marine sterol and the urine analysis work (see Applications, Section 3.4). Limitations in amounts of material complicate use of conventional separation and analysis procedures, making mass spectrometry the technique of choice in these problems. In nearly every case, mass spectrometric data are required to establish the molecular weights and formulas of our unknown compounds and to provide fragmentation evidence to supplement structural hypotheses derived from other spectroscopic techniques.

The increased specificity of high resolution mass spectrometric data obtained from gas chromatographic fractions together with conventional library search procedures and automated analysis of the mass spectra has provided a unique resource which represents the foundation for our resource-related research. The current capabilities of the system, now in routine use, and some examples of recent applications are summarized in the accompanying annual report (Appendix II). In the remainder of the period covered by our current grant, we foresee increased dependence on the GC/HRMS facility for providing spectral data which can be acquired for ourselves and our collaborators in no other way. The current performance of the system together with requested developments will provide a routine tool to support our research. Our more general approaches to computer-assisted structure elucidation as described in this proposal will make maximal use of the GC/HRMS data in structure problems. But because structure elucidation draws on many other sources of data besides mass spectrometers we must provide the facilities to accommodate structural inferences derived from other methods. Thus, our proposal reflects diminished emphasis on new developments in hardware and software

for GC/HRMS analysis and increased emphasis on problem-solving programs for more general applications to structure elucidation.

1.3.4 Related Computer Programs

Our present grant has led to development of several ancillary computer programs which support our efforts in research in mass spectrometry and computer-assisted structure elucidation. These programs have been summarized in detail in last year's annual report and the current annual report (Appendix II). Briefly, the more important of these programs in the area of processing of high resolution mass spectral data include: a) routines for detailed evaluation of the performance of the mass spectrometer to ensure optimum performance when unknown samples are run; b) data reduction programs based on a computed model of the characteristics of the mass spectrometer; c) real-time resolution of overlapping mass spectral peaks; d) rapid determination of elemental compositions; and e) CRT display reporting of instrument operating characteristics both during calibration and actual runs. Together these routines provide a basis for rapid, reliable reduction of the large volumes of data acquired during GC/HRMS runs. In the area of processing of low resolution mass spectral data we have developed the CLEANUP program (70) which, given complete GC/low resolution mass spectral (GC/LRMS) data consisting of repetitive scans of mass spectra, detects the elution of components, removes background contributions and resolves overlapping GC elutants to arrive at mass spectra which more closely represent the spectra of pure components. In collaboration with Professor Lederberg's group in the Department of Genetics, we have also implemented a library search program based in part on the methods of Biemann, et.al (H.S. Hertz, R.A. Hites, and K. Biemann, Anal. Chem., 43, 681 (1971)). The program allows rapid screening of the spectra to remove those components which are known structures, thus focusing our attention on those which have not been previously identified.

We have also developed a program, called MOLION, for prediction of molecular ions in a mass spectrum (45). This program predicts plausible candidates for the molecular weight (or formula, given a high resolution mass spectrum) independent of the presence or absence of the molecular ion in the spectrum. The PLANNER program (28) has been converted to an interactive version available on SUMEX. This program analyzes a mass spectrum in terms of molecular structure based on the spectrum and fragmentation rules for the class of compounds to which the unknown belongs.

In our efforts to provide an interactive program for computer-assisted structure elucidation which is useful outside our own community, we have provided a variety of additional functions for CONGEN. Some of these functions are part of the program itself and were discussed above. Additional examples include: a) support of a wide variety of computer terminals from

simple teletypes to complex graphics terminals, so that remote users can access CONGEN and use it effectively with any terminal they possess; b) a "gripe" system for reporting problems to us; and c) a "bugout" system to save a copy of the entire program when a user encounters a supposed program error, thus allowing us to examine the problem as it occurred.

1.3.5 Collaborative Research Environment

The previous sections summarized those capabilities and facilities which are the direct products of our past research. However, the collaborative nature of our research efforts among the Departments of Chemistry, Computer Science and Genetics is a unique environment, which a brief summary cannot describe adequately. We can call upon the expertise and facilities of a large number of research groups which are involved in work which is at least peripherally related to our own efforts. By doing so we discover common problems and can work in concert toward common solutions. We identify new application areas by encouraging others to use our programs, usually resulting in improvement of the programs as they confront new problems. Although it is difficult to convey the spirit of such close collaboration, suffice it to say that the continuing interest of a variety of people from a variety of backgrounds provides far more facilities available to us than directly supported by this grant. For example, outside researchers on other related projects provide valuable comments, criticisms and assistance; our collaborators share special laboratory, instrument and computer facilities.

This collaboration requires both sustained interest and a critical mass of people who are devoted to making the instrumentation and programs work more effectively. Because we have had both, tremendous savings of time and effort have resulted and should continue to do so. For example, we have been able to provide access to CONGEN via the SUMEX resource to help outside persons solve structure elucidation problems (see Appendix II). Portions of our programs, e.g., the Omnigraph display routines, were developed elsewhere (Omnigraph at NIH). We were able to incorporate them into our programs saving us considerable time by avoiding duplication of effort. Availability of our programs on a public computer network means that they are readily accessible to scientists across the nation. This constitutes a mode of resource sharing and publication of programs in a way that is nearly unique for software. Such sharing not only increases the programs' use to others but provides sources of critical refinement for our own scientific progress.

1.4 Relationship to Mass Spectrometry and AIM-SUMEX Resources

1.4.1 Mass Spectrometry Resource

We have over the past two years, under NIH support, developed a specialized resource for combined gas chromatography/mass spectrometry. Our special interest was in operation of the mass spectrometer at resolving powers sufficiently high to permit accurate mass measurement and, thus, determination of empirical formulas for each ion detected in the spectrum of each elutant from the gas chromatograph. The idea of operating a mass spectrometer at high resolving power in conjunction with a gas chromatograph (GC/HRMS) is not new (Section I in "Biochemical Applications of Mass Spectrometry," G.R. Waller, Ed., John Wiley and Sons, Inc., New York, N.Y., 1972). But because of difficulties with the technique and expense of facilities to provide these data, whether from photographic plates or from on-line recording of spectra, such GC/HRMS systems are not routinely available.

We developed a GC/HRMS system because we recognized its utility in our own research and the research of collaborators, most of whom are engaged in characterization of small amounts of complex mixtures. We recognized some of the problems with earlier efforts by other workers and designed our system to alleviate these problems. First, we recognized that the system (computer and mass spectrometer) must be capable of measuring and validating its performance prior to the introduction of valuable samples. We recognized that data acquisition and reduction must be completely automated because, with limited personnel, there is not time to process parts of the large volume of data manually. We have accomplished our design goals and propose further developments to increase the utility of the system.

The importance of the mass spectrometer resource to our efforts in computer-assisted structure elucidation cannot be underestimated. Structure elucidation cannot be successful unless the empirical formula of the compound has been determined. Mass spectrometry, particularly high resolution mass spectrometry, is the technique of choice for determining this key datum. As summarized in Section 3.4, many of our applications require GC/MS for separation of components and acquisition of their respective mass spectra. We plan, together with our collaborators, to make extensive use of this resource in new applications.

1.4.2 AIM-SUMEX Resource

AIM-SUMEX (NIH RR-00785, Oct. 1, 1973, thru July 31, 1978, Principal Investigator, J. Lederberg) is a national facility for applications of artificial intelligence in medicine (AIM). Our

own use of this facility will include SUMEX PDP-10 computer time and file storage necessary to run the DENDRAL programs. This support will be furnished without charge to the present proposal as it has been in past years. It represents an annual investment of about \$100,000 in computer time, system software and specialized consultation for new system development.

The AIM-SUMEX computing facility is shared equally between a national user community (AIM) and a Stanford Medical School community. The DENDRAL research is supported out of the Stanford portion. The AIM service is administered under the policy control of a national advisory committee and is implemented over a national computer network. AIM-SUMEX provides the means for members of the national user community interested in structure elucidation to access the DENDRAL programs.

2 Specific Aims

2.1 Add More "Intelligence" to Existing Programs

By adding extra intelligence to the DENDRAL programs we mean giving the programs the ability to reason about the chemistry of a problem statement in addition to the program syntax. We believe this will increase their problem solving power and make them easier for scientists to use. There are two specific areas for development: i) adding inferential knowledge to the interface between scientist and program; ii) adding smart assistance capabilities to guide the scientist to productive use of the problem solving programs.

We propose to add inferential knowledge to the CONGEN program which will interpret the scientist's description of the structural problem in terms that are best suited for the program's efficient solution. This extension, which we call the "constraints interpreter" remove any requirement of knowing CONGEN's algorithm for solving the problem.

We propose the development of a help system for CONGEN ("CGHELP") to assist the user in making optimum use of the basic CONGEN program. Though specifically related to CONGEN, CGHELP will be formulated in general terms. CONGEN is the best target

program for this project because, of the current user-level DENDRAL programs, it has both the greatest potential for widespread use among research chemists and the most complex and logically exacting input requirements. We will develop these ideas to include five specific aids:

1) On-line documentation system 2) Tutorial error handling
3) Internal model of the user 4) Error correction aids 5)
Extension of "error" concept to cover strategy, helpful suggestions, perception aids

2.2 Develop New Computer Programs that Assist in Biomolecular Structure Elucidation

CONGEN provides a mechanism for solving the "jigsaw puzzle" aspect, the assembly of structures which are consistent with structural information inferred manually from many sources. It does not help the chemist with two other key steps (Fig. 1): 1) deciding what a good "next step" would be in a partially completed problem; and 2) inferring structural information directly from chemical or spectroscopic data. To become a well-rounded facility for biomolecular structure elucidation, we wish to focus upon these other steps, and to this end we propose four new programs which use chemical knowledge in novel ways.

1) **Experiment Planning.** The first program relates to experiment planning and will draw upon an internal knowledge base of experimental techniques, chemical and spectroscopic, of modern structure elucidation. The application of this knowledge involves recognizing which functional groups and structural relationships in a given problem can be practically deduced, and by what methods. A chemist draws upon such information when he/she has a partially solved problem and needs to decide which experiment will most effectively limit the remaining possibilities. It is this process which we intend to model in the experiment planner. This program will fit logically at the end of a CONGEN run which has yielded a large number of structures consistent with the given constraints, and will provide the chemist with guidance to fruitful new experiments.

2) **Reaction Sequences.** We propose work on a new program called REACT, which will carry out chemical reaction sequences (61). Chemical reactions constitute an important source of structural information for unknowns. Our aim in the further development of REACT is to provide a mechanism for using this information in computer-aided structure elucidation problems. REACT, like the experiment planner, fits logically at the end of a CONGEN run, allowing the chemist to eliminate from consideration candidate structures which are inconsistent with data derived from laboratory experiments involving chemical treatments of an unknown.

3) **General Analysis of Mass Spectra.** We propose a program for the analysis of mass spectra which uses general (as opposed to class-specific) knowledge of allowed mass spectral (MS) fragmentation processes. These rules will come either from expert mass spectroscopists or from the Meta-DENDRAL program, and the user will be able to tailor them to his specific cases as necessary. MS data are currently under-utilized in structure elucidation problems because of the complexity of combining together the structural implications of each observed ion. The new program will embody algorithms for dealing directly with this complexity. The program can be viewed as a data-driven generation scheme, one which will allow the incorporation of MS data from the very beginning of a problem. It will complement the existing generation scheme in CONGEN, where fragmentation rules can only be used now as post-tests to trim a list of structural candidates obtained using other structural data.

4) **C13 Spectral Analysis.** We propose a C13 NMR analysis program paralleling the MS program described above. Here, the rules which guide the analyses relate local structural environments of carbon atoms to their observed chemical shifts. Some rules exist for certain classes of organic compounds while others are expected to result from the C13 Meta-DENDRAL effort (see below). Like mass spectrometry, C13 NMR is now under-utilized as a structure-elucidation tool, partly because of the difficulty of manually combining into complete structures the substructural possibilities corresponding to each peak, and partly because the technique is new enough that the rules themselves have not been exhaustively explored.

2.3 Develop New Programs that Aid in Rule Formation

The Meta-DENDRAL programs have been developed to be conceptually sound; recently they have been improved to be productive research tools. We propose to improve their usefulness and to explore ways of generalizing the concepts.

The quality of rules will improve, we believe, when the program can make incremental improvements to rules. Thus we propose adding feedback loops to the current "single pass" system. In the long term, we also believe the program's rules will need to be improved through the exploration of different models in terms of which the rules are written. We intend to move the program farther away from the current "fixed model" system.

Generalization of these programs will be carried out in steps. The first step toward generalization will be to work in a domain with some similarities to mass spectrometry but many differences. We believe C13 NMR spectroscopy is a promising domain for application of these ideas, and one that is as important for structure elucidation as mass spectrometry. In

rewriting the programs to form rules in this second domain, we will make them as general as possible. We then intend to find another domain of biomedical science in which to test the programs' generality. In the end our aim is to have a knowledge-based rule formation program that can be applied to many types of domains and whose limitations are well understood.

2.4 Apply the Structure Elucidation Programs and GC/HRMS System to Biomedical Problems at Stanford and Elsewhere

We intend to apply the combined gas chromatography/high resolution mass spectrometry system and our programs for computer-assisted structure elucidation to biomedical structure problems at Stanford and elsewhere. Details of the collaborative efforts are presented in the Applications section. Such applications include: a) identification of marine natural products, especially sterols and other terpenoid systems; b) identification of new metabolites in patients with birth defects of genetic origin; c) exploration of mechanisms of cyclization and rearrangement of complex systems; and d) structural problems of biomedical importance posed by outside users of our programs.

2.5 Increase the Availability of the Structure Elucidation Techniques

We intend to increase the availability of our programs for computer-assisted structure elucidation. We will accomplish this in two ways. First, we will improve the performance of the current programs by making them more intelligent and easier to use. This will allow us to serve a larger community of users via the SUMEX computer resource. Second, we will rewrite CONGEN and continue its maintenance in another computer language, more exportable than INTERLISP. This will enable other persons to use the program at their own computer facilities.

2.6 Maintain and Improve the GC/HRMS System

We will maintain and improve the GC/HRMS resource. Maintenance of the mass spectrometer and the associated computer system is obviously essential to guarantee that high quality data are available to us in support of our research. We will improve the system by writing improved data reduction software which will allow us to scan the mass spectrometer at lower resolving powers, thus improving the sensitivity of the system. We will devote more attention to the user interfaces to the data presentation programs so that users can peruse their data in the off-hours at their leisure.

3 Methods

3.1 Extra Intelligence in Existing Programs

3.1.1 Constraints Interpreter for CONGEN

There are generally many different ways to express a structure elucidation problem to CONGEN; some are practical, others are impossible to solve. For example, it is efficient to specify known aggregates of atoms (superatoms) to be used as building blocks. It is inefficient to generate all structures of an empirical formula and test each one for the presence of known superatoms. A scientist cannot be expected to know all efficient ways of specifying a problem. Our experience is that the first few sessions with CONGEN are spent developing a feeling for the combinatorial complexity of structural problems and ways to constrain the problem efficiently. We wish to shift the burden of learning about efficiencies in CONGEN from the scientist to an interface program.

We propose, based on our experience with helping new users, to develop a "smart" constraints interpreter for CONGEN. The interpreter would: 1) examine the information supplied as input and automatically translate that information where possible into additional superatoms or constraints implied by the input data; 2) ask about translations which are questionable; 3) determine the most efficient way to solve the problem beyond efficiencies gained by (1) and (2).

The constraints interpreter is so critical to efficient use of CONGEN that we wish to reemphasize the preceding paragraphs and give some examples to illustrate how the problem solving capabilities of CONGEN will be improved. A typical scientist comes to CONGEN with an unknown structure on which considerable data have been acquired. He/she probably has a few candidate structures for the unknown in mind. Known information is supplied to CONGEN, usually incompletely because knowledge of the problem introduces biases which are not given to the program (e.g., forgetting to forbid certain unfavorable substructures or functionalities such as peroxides). Without knowledge of the best ways to express the problem to CONGEN the known information is seldom input in a way which is optimum for rapid solution. The result is a problem which is too large to solve. Reexamination of the problem with our assistance results in better ways to solve it. The program should provide this assistance automatically to avoid discouraging false starts. The following are some functions of a constraints interpreter which will provide that assistance.

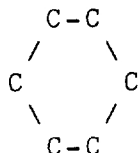
Input Translation. An input translator will determine the implications of the input data and find a new internal definition

of the problem to solve it more efficiently. Several heuristics which we use manually will be given to the program. For example, we know that tremendous reductions in the scope of a problem are achieved when even a single atom or unsaturation is included in a superatom rather than allowing the atom or unsaturation to adopt any of several different environments. Constraints on a problem frequently contain substructures which imply larger or additional superatoms. A single carbonyl group on GOODLIST (14,48), for example, should be used as a superatom to construct structures rather than retrospectively testing for the presence of the carbonyl functionality.

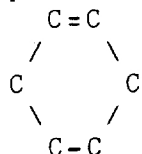
In other cases, substructures appear on GOODLIST either because they are too imprecise to be superatoms (i.e., they may contain atoms or bonds multiplicities which can take on a range of values) or because they may overlap other superatoms (superatoms are required by CONGEN to be atom-disjoint fragments). In many cases, it is possible to remove the imprecision by considering each of the possible values in a range to be a separate subcase. For example, a C13 NMR spectrum might indicate the presence of a carbon atom doubly bonded to either an oxygen atom or a nitrogen atom. The corresponding GOODLIST entry would be C=(N O) where (N O) is a "polynome" meaning "either N or O". This could be broken down into two subcases, one in which C=O is used as a superatom and one in which C=N is used. Each subcase could be solved independently and the results combined to give the full result.

The expression of GOODLIST items as superatoms is just one example of the kind of input translation we foresee. We will explore the automation of several other manual techniques we have used to maximize the efficiency of constraint expression.

User Communication. A second function of the interpreter will be recognizing circumstances where the input data imply a small number of choices about the interrelationships of superatoms and constraints. Such circumstances would result in questions to elicit additional, specific information about a problem. For example, suppose a user gives the superatom



to CONGEN, specifying that there may be one additional bond connecting atoms within the superatom. If GOODLIST also contains C=C, then one possible interpretation would yield the superatom



Because this increases the order of one of the bonds in the

original superatom, it may not be what the user had in mind. A request for clarification at this point could rule out the above possibility and reduce the number of cases considered by CONGEN.

Efficient Problem-Solving within CONGEN. A third function of the interpreter will be to attempt to order the various steps required for solution to solve the problem more efficiently. Currently we require the user of CONGEN to carry out each step in construction of structures explicitly because different constraints have different implications at each step. This is an artificial barrier which will be removed by the constraints interpreter. Given the input data, the program will decide which constraints are applicable at each step and the optimum order of steps to arrive at solutions.

For example, one useful manual strategy is to recognize features of a problem which are not heavily influenced by the constraints, to solve a constrained sub-problem in which those features are removed, and to reintroduce them at the end. We have seen problems in which several monovalent atoms or superatoms were present which were not referred to by the constraints. Such a problem can be solved most efficiently by removing the monovalents from consideration, constructing molecular skeletons under the given constraints, then including the monovalents in a final node labelling step. This is much more efficient than carrying out the full structure generation with constraints.

3.1.2 Intelligent Help System

As programs such as CONGEN and INTSUM have moved closer to routine use, we have become aware of a new kind of computer science problem: How can users at different levels of experience obtain useful results with a minimum of effort and frustration? Historically, the bulk of effort in developing the DENDRAL programs has gone into the underlying algorithms which allow these programs to solve extremely complex symbol manipulation problems. Interfaces to these programs have been designed to give a knowledgeable user (i.e., one who understands the algorithmic structure of the program) access to the basic functions available, not to help a less experienced user understand how these functions can be fit together in solving a larger problem.

This approach is often appropriate for a program which is to be used locally because knowledgeable users are available either to submit runs for others or to guide others in learning the subtleties of operating the program. However, the remote community of DENDRAL users, is growing, so the need to explore the interface problem as a separate research topic becomes increasingly obvious.

We have solved interface problems until now in a piecemeal

fashion. For example, we responded to the psychological problem of unduly long computation times without visible results (in large structure elucidation problems) by providing interrupt facilities to monitor the progress of the problem. Making these interrupts available to researchers gives them control over the frequency of progress summaries printed by the program and puts them in closer touch with the problem solving steps of CONGEN. We now seek to undertake a unified, consistent approach to the interface problem.

We propose to develop a help system for CONGEN (called CGHELP) to assist in making optimum use of the basic CONGEN program. We will approach the development of CGHELP incrementally through development of the following facilities:

- 1) On-line documentation system
- 2) Tutorial error handling
- 3) Internal model of the user
- 4) Error correction aids
- 5) Extension of "error" concept to cover strategy, helpful suggestions, perception aids

Details of the the individual sections of CGHELP, the proposed intelligent help system for CONGEN, are provided in Appendix I.

3.2 New Programs for Structure Elucidation

3.2.1 Experiment Planning Program

The problem of intelligent planning by computers is currently receiving attention in the artificial intelligence community [e.g., E. Sacerdoti, Ph.D. thesis, Stanford] and in application areas such as molecular genetics here at Stanford. In the context of elucidation of molecular structures experiment planning plays a crucial role (Fig. 1). One can consider the overall problem of structure elucidation (as done manually) as the construction and testing of a series of hypotheses (candidate structures). CONGEN gives us the capability of constructing all plausible candidates under an initial set of constraints; the next problem is how to provide the researcher with some assistance in the problem of rejecting incorrect candidates to focus in on the correct structure.

This problem is attacked manually by examining the candidates, determining their common and unique structural features and designing experiments to differentiate among them. When there are dozens or hundreds of structural candidates, manual examination and intercomparison for structural features and, consequently, experiment design become extremely difficult. We propose to automate some aspects of the manual methods to assist the chemist in designing new experiments.

The methodology for a computer-based approach to this problem will involve two major steps: 1) examination of functional groups and other substructures in the set of candidates in view of knowledge of available spectroscopic and chemical techniques and the type of information provided by each technique; and 2) presentation to the researcher of an ordered list of experiments to be performed to reduce the set of candidates.

We will draw on our experience in helping design a similar knowledge base for experiment planning in molecular genetics. As in that domain, the basic item of information to represent about each experimental technique is a transformation of a molecular structure (or partial description thereof) into data points. We also need to store information about the precision of the technique, its necessary preconditions (sample size, volatility, etc.) and its likely sources of error. If complete enough, the information in this knowledge base can be used to simulate a sequence of experiments.

The capability for experiment planning will be developed in three parts, the first two to carry out structure intercomparison in the context of the knowledge base and the problem, the third to determine an optimum strategy for the new experiments.

1) **Comparison of Structures.** The first step is to develop an efficient method for intercomparison of structures to determine the key features which allow differentiation among them. We will improve and extend our current, limited capabilities for surveying a set of structures for the occurrence of each member of a specified set of structural descriptors. The extensions required include a solution to a subset of the general problem of determining differences between two graphs (it is a subset in that both structures possess the same number of atoms of each type).

As the knowledge base of experiments grows (see (2) below), we can begin guiding the intercomparison according to the types of substructural features which can be distinguished by experiments described in the data base. We will retain other distinguishing features and report them also because the knowledge base will never be complete and an undescribed test may exist for special cases. However, there are other considerations which will be used to guide strongly the procedure for intercomparison; the context of the problem provides this guide. For the procedure to display any degree of chemical common sense, it must be aware of the input superatoms and constraints (see also section on Constraints Interpretation), because all structures will have the features of these input substructures in common.

2) **Knowledge Base of Experiments.** Proper organization of the knowledge base which contains information on spectroscopic and chemical procedures and the structural inferences which can

be derived therefrom is very important. To be general and reasonably efficient to search it must be organized hierarchically in terms of structural information. It must also be cross-referenced to take advantage of the knowledge of both the set of inferences which can be obtained from a particular technique and the possibility of reinforcing an hypothesis by examination of data from more than one technique.

Our proposals for this organization are as follows. Considering the substructural organization of the knowledge base (which provides the keys which can be searched for in intercomparison of structures) we assume a hierarchy of structural descriptors, from broad descriptions to specific items. Broad descriptors include one category for functional groups, one category for proton distributions (e.g., from ^1H NMR data), one for carbon distributions (e.g., from ^{13}C NMR data), one for ring size and type distributions, and so forth. Each category will be further subdivided as appropriate. For example, the functional group category can be subdivided according to heteroatom, local functional environments for each heteroatom, and "extended" environments which include the functionality and more remote structural features. As each descriptor becomes more specific and an experiment exists which can provide some information about the descriptor, the experiment will be included as part of the information associated with the substructure. Associated with each experiment will also be qualifiers on sample requirements, interfering functionalities, and preconditions for the experiment (e.g., solubility, etc.). Of course, the experiments will become more specific also. For example, an initial suggestion for an experiment might be to obtain a ^1H NMR spectrum if one has not been obtained. The next suggestions would depend on how the structures differed in those characteristics which are normally easy to determine from a ^1H NMR spectrum, e.g., number of methyl groups, vinyl and aromatic protons, etc. At the most detailed level, specific proton decoupling experiments would be proposed if the candidate structures differed in appropriate ways.

Cross referencing of the knowledge base can be used effectively. Frequently, the same substructural information can be derived in a variety of ways. If a chemical experiment suggests the presence of an hydroxyl group, then confirmatory evidence should be available from NMR and IR spectral data. Knowledge that these spectra are available, or are about to be suggested as the next experiments to be performed can be used to search the knowledge base for other relevant substructural information which is routinely obtainable from these techniques. Then the substructures can be examined to determine if they have any discriminatory power among the candidate structures. Thus, an experiment suggestion can take the form "determine the NMR spectrum to check for 'xyz'; also, the same spectrum should reveal whether or not 'zyx' is present". The knowledge base will therefore be used in two complementary modes. The first is keyed according to a hierarchy of substructures. The second is keyed

through the cross-indexing of experiments which might be performed.

3) **Proposed Experiments.** The above procedures examine structural candidates and make decisions on what experiments might be done. The final procedure is to determine which experiments are feasible and to develop a strategy for carrying them out in an efficient sequence. We know of several heuristics to guide this procedure. Feasibility is related to sample size and physical and chemical characteristics of the sample. The knowledge base will have qualifiers relating to specific requirements for each experiment. Where necessary the researcher will be queried about the amount of sample available and other characteristics to help the program determine feasibility. For those experiments which are feasible, there are several heuristics which will guide determination of a good strategy for carrying out the experiments. Information which might be obtained from available data should be considered first. Experiments which would reject only a small number of structures should have lower priority than those which would yield a higher reduction. Experiments which are simple and non-destructive of sample may be given higher priority. Certain combinations of experiments will have greater discriminatory power than other combinations. We will develop decision criteria based on these considerations. Based on our experience with the MYCIN program [57,58] we will provide the capability for the researcher to query the system to determine why certain experiments were proposed, and to alter the strategy for experiment selection where he/she deems it necessary.

3.2.2 Reaction Chemistry Program

Knowledge of reaction chemistry can provide important analytic information for structure determination problems. In addition, we believe it is important for the success of CONGEN to understand the fundamental graph-theoretical questions raised by reaction transformations. We will develop a program, called REACT, which uses knowledge of chemical reactions to carry out reactions in the computer and thus enables us to explore these two important areas. Some preliminary exploration of these ideas (61) convinces us of their feasibility. Since some of these ideas overlap with those of T. Wipke in the area of chemical synthesis by computer, we will continue to work closely with him. His research group also uses the SUMEX computer.

3.2.2.1 Use of REACT in Structure Elucidation

Reactions can play a key role in structure elucidation problems in several different ways. Chemical reactions may:

a) test for a specific functional group;

- b) simplify the problem by decomposing the unknown into smaller, more easily characterized molecules;
- c) modify the skeleton or functional groups to define more accurately their respective environments or make the unknown more amenable to analysis (e.g., increase its volatility); or
- d) unambiguously relate the unknown to a previously characterized compound.

In all of these cases, measurements on the products of the reaction are used to limit structural possibilities for the original material. In many cases such new information can be expressed directly as constraints on the possible structures for the unknown. There is, however, an important class of reactions in which the translation of observations on the products into direct constraints on the structural possibilities is difficult if not impossible. In these cases it is essential to consider the application of the reaction to each structural candidate and the relationship of these candidates to their respective products. The most common examples of this class are reactions in which a given product or set of products may be obtained from different candidate structures for the unknown (e.g., an oxidative cleavage of several candidate structures might yield proposed products some of which are the same. (See ref. 61 for further examples). Or, stated slightly differently, the class of reactions in which there is more than one way for a given product or set of products to undergo the reverse, or antithetic reaction. Through the REACT program, we intend to give the research chemist the capacity to incorporate this reaction-dependent information into the computer-assisted identification of unknowns. REACT is currently in embryonic form. We are developing it as an extension of CONGEN, using the existing capabilities therein to allow us to focus on the key new concepts. The proposed research on REACT involves separating it from CONGEN, enriching the menu of basic tools available to the user and developing an input language which is flexible and easily used. Our initial experience with REACT indicates that the following topics require investigation.

(i) We intend to add the ability to define a wide range of constraints upon each reaction. We can now specify many features in the reactant for, or the product(s) from, a reaction which either are necessary conditions for the reaction to occur or will prevent it from occurring. Other crucial constraints, however, cannot be specified. Specifically, these are constraints which apply relative to a potential reaction site rather than to the molecule as a whole. For example, while we can say that a hydroxy group (OH), if present anywhere in the reactant molecule, will inhibit a given reaction, we cannot say that such inhibition will take place only if the group is adjacent to the reaction site. Such site-specific constraints are vital to the detailed description of reactions and their inclusion in REACT will

substantially increase its usefulness in real-world chemical problems.

(ii) We foresee improvements in the higher-level control structure of the program to give greater latitude in controlling the grouping of structures and describing required relationships between products and reactants. There are currently only two types of control information which can be given to REACT: 1) Substructural constraints to group the structures within a given list of products into an arbitrarily complex set of interrelated classes; and 2) constraints requiring that only specified numbers of products in any class can be obtained from each molecule in the parent (i.e., reactant) list. The former operation is analogous to chemical separation while the latter is used for eliminating parent molecules which do not give the proper types and numbers of products under a given reaction. There are some structure elucidation problems in which this level of control is not sufficiently detailed. For example, a single-step reaction, when applied to a given structure, may yield multiple products either because it is a cleavage reaction which splits the parent into smaller fragments or because the reaction site appears more than once in the parent, with each occurrence giving rise to a distinct product. We now only count the total number of products, and thus miss the sometimes crucial distinction between multiple pathways for a reaction and multiple products from a given pathway.

(iii) We intend to make REACT a stand-alone interactive program which gives the user a "chemical laboratory" in computerized form. A variety of interactive aids and consistency checks upon input will be needed to make the program understandable and easily used. There will be considerable complexity in both the internal format of defined reactions and the structure of the reaction sequence tree (the central data structure of REACT which holds all lists of chemical structures and the interrelationships them). The challenge of developing the interface will lie in giving the chemist access to this information in as intuitive a language as possible. Fortunately the complexities are ones which are inherent to the chemical problem so most chemists already have the conceptual base and the language necessary to deal with the program's logic. Terms such as "reaction mixture", "cleavage products", "exhaustive reaction" and "separation of products" all have meaning both in laboratory chemistry and in REACT. We intend to draw upon this parallelism as extensively as possible in designing the input language.

3.2.2.2 Importance of REACT for Relating Graph Theory to Chemistry

Our second interest in chemical reactions is mathematical. Reactions bring up a number of graph-theoretical questions which have not previously been formalized concerning what we might call "transformational graph theory" (some of these problems are

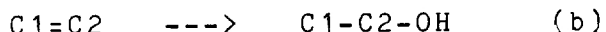
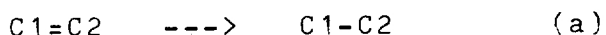
currently under investigation in other laboratories; see W. T. Wipke, et al., in "Computer Assisted Organic Synthesis," W. T. Wipke, Ed., American Chemical Society, Washington, D.C., in press). We will investigate these questions in an attempt to find a theoretical foundation which is consistent with the largely intuitive approach embodied in our preliminary version of REACT. We expect that such an exploration not only will contribute to the mathematical foundations of chemistry in general (and CONGEN in particular) but also will give us a general method for describing graphical transformations that can be applied to other problems, for example, an in-depth study of questions of the mechanisms and rearrangements involved in the formation of terpenoid systems (62).

We see three main areas of mathematical interest in reaction chemistry. First is the question of formally representing graph transformations. For the description of static topological properties of molecules we have made extensive use of graph theory as a foundation, but there is no analog for the process of graph interconversion which is at the heart of reactions. In REACT, as in programs developed elsewhere dealing with chemical transformations, representations for transformations have been chosen primarily on an ad hoc basis with guidance not from underlying mathematical principles but from specific requirements of the program and/or the problem domain. We will investigate other representations for chemical transformations, including: a) subgraph substitution, in which a reaction consists of two subgraphs one of which (the "product site") is substituted for the other (the "reactant site") wherever the latter is found; b) subgraph plus modifications, in which the reactant site is described as above but is accompanied by a standardized list of elemental graph transformations which describe the overall graph modifications. (This is similar in concept to the current implementation in REACT); and c) subgraph plus "difference graph", which is similar to (b) above except that the modifications are expressed as a special kind of graph rather than as a list of elemental transformations. By exploring the relative advantages of these and perhaps other descriptions, we hope to arrive at one which will not only be amenable to formal mathematical reasoning but also gives an adequate descriptive language for chemistry.

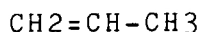
A second mathematical question, which has import for both the theory and the efficiency of REACT, concerns duplication among the products of a reaction. There are two sources of duplication: a given molecule can undergo a reaction by different pathways (i.e., different instances of the site) which yield isomorphic products (or sets of products for cleavage reactions); or two structures within the parent list may undergo reaction to give isomorphic products. In REACT we eliminate duplicates by casting each product into a canonical, or standard, form as it is created and comparing it directly with each previously obtained product. Not only does this approach imply redundant effort within REACT, but it is also an unsatisfying "brute force" method

which we feel is amenable to mathematical refinement. In the first case mentioned above, part of the problem relates to the symmetry of the reacting molecule and the "symmetry" (still an ill-defined concept for this problem) of the reaction. We now have a theoretical model for using these symmetries to avoid symmetry duplicates before generating them, a model which is distantly related to the "double coset" algorithm which plays an important role in CONGEN.

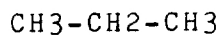
Third, we intend to explore and formalize the concept of symmetry as it applies to graph transformations. While symmetry of individual graphs is well defined, the symmetry of transformations is not, although chemists have an intuitive concept of reaction symmetry which they apply as second nature when deducing the products of a reaction. For example, consider the two reactions (a) and (b) below, which respectively represent a hydrogenation and a hydration of a double bond.



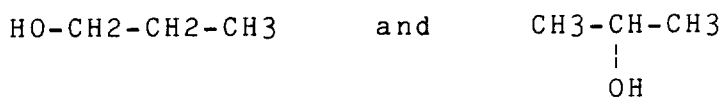
It is easy to see that in (a) the carbons (atoms 1 and 2) play equivalent roles but in (b) they do not. In applying these reactions to the molecule



a chemist will automatically consider only one occurrence of the reaction site (C1=C2) for (a) and will obtain only one product



For (b) he/she will "see" two instances of the site and will write down two products



These two instances of the site use the same atoms and bond in the parent molecule but for (b) the two fittings are not equivalent as they are for (a). The difference in symmetry of these reactions is obvious in this simple example, but there are more complex cases in which intuition gives little help. Only through a firm understanding of the principles behind the intuition can we hope to model it successfully in a program.

3.2.3 General Mass Spectrum Analysis Program for Unknowns

PLANNER, which is currently the only program we have for

interpreting mass spectrometry (MS) data directly for an unknown, is restricted to class-specific rules and although it is quite useful in some domains (e.g., locating possible positions of substituents in a compound whose skeleton is known), it is not well suited to the general structure elucidation problem. The general pattern of use of mass spectral data in problems where class-specific information has not proven useful, and the compound's spectrum is not in a library, has been to use the data to determine molecular weight (or formula) with detailed structure/spectrum correlations left for retrospective rationalization. But we know that the mass spectrum contains a great deal of more specific structural information. Every ion observed is a fragment of the original molecule and because rearrangement of atoms or groups other than hydrogen is a very unfavorable process, except for certain special cases, every ion observed contains atoms which were linked together in the intact molecule. Every spectrum contains from a few to perhaps hundreds of unique ions. Even granting considerable redundancy, a spectrum should yield more useful information than is usually obtained. One approach of limited generality to extraction of structural information from a spectrum has been presented for analysis of so-called "sequence" molecules (A. Kunderd, R.B. Spencer, and W.L. Budde, *Anal. Chem.*, 43, 1086 (1971)). This is a generalization of work by Biemann and McLafferty on peptide sequencing by MS. See M. Senn, R. Venkataraghavan, and F.W. McLafferty, *J. Amer. Chem. Soc.*, 88, 5593, (1966); K. Biemann, C. Cone, B. R. Webster, and G.P. Arsenault, *ibid*, 5598 (1966)). narrow category and one cannot always assign a unique structure for each of the sequence ions.

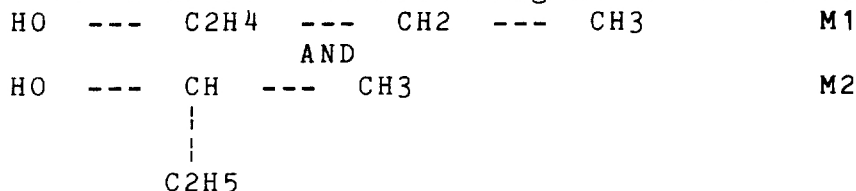
We have recently begun to explore a generation scheme which may be viewed as a generalization of sequence analysis. It makes use of a new concept called a mass distribution graph (MDG). An MDG is an entity related to (topological) chemical structures except that partitions of atoms (e.g., $C_3H_7-OCH_3$) are linked instead of individual atoms. Thus, one MDG may represent a whole family of topological isomers. Being a graph, it is composed of nodes interconnected by edges. Each edge in an MDG stands for one single or multiple bond, The restriction upon MDG's is that there must be some way of assembling the atoms in each partition into a connected molecular fragment (superatom) and some way of linking superatoms (using the MDG edges) into a connected chemical structure. Corresponding to each MDG is a family of structures which can be created by these two assembly steps. Within CONGEN we have the algorithms necessary for carrying out these steps.

To illustrate this, we will use a very simple example. Suppose a high resolution mass spectrum for a compound shows four major peaks, corresponding to compositions of C_4H_{100} (M^+), C_3H_70 ($M^+ - CH_3$), C_4H_9 ($M^+ - OH$) and C_2H_50 ($M^+ - C_2H_5$). Further suppose that the MS theory in this case is the simplest possible one; an allowed fragmentation involves the cleavage of just one single bond with no transfers of hydrogen or other neutral

species into or out of a fragment. The M^+ peak defines the overall composition which, together with the MS theory, allows us to represent each remaining peak as an MDG in the form peak composition - complement, as follows:

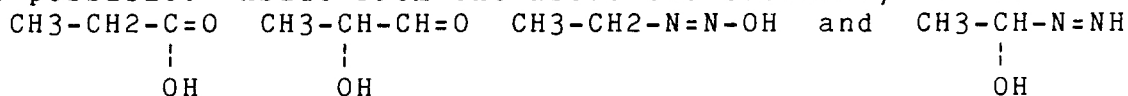
| Peak | Composition | Corresponding MDG |
|------|-------------|-------------------|
| P1 | C3H7O | C3H7O ---- CH3 |
| P2 | C4H9 | C4H9 ---- OH |
| P3 | C2H5O | C2H5O ---- C2H5 |

The MDG generation scheme revolves around the combination of these 1-peak MDG's into more detailed MDG's each of which simultaneously accounts for several peaks. We define an "overlap operator", @, which represents this combination. Thus P1 @ P2 is the set of MDG's which have two bonds, one of which splits the overall composition according to P1 and the other according to P2. Each of these can then be "overlapped" with P3, and the resulting MDG's can be expanded into full chemical structures using structure generation and imbedding techniques already developed for CONGEN. The resulting MDG's are M1 and M2.



Only two possible structures result in this simple case, 1-butanol and 2-butanol.

MDG's can be formulated in terms of low-resolution MS data as well as high resolution data. In this case the nodes correspond to masses rather than compositions and the final expansion of MDG's to structures is accompanied by an extra step, the determination of all compositions which account for the mass of each MDG node. If the above example is treated as a low resolution problem (peaks 79[M+], 59, 57 and 45), then assuming only C, H, N and O as possible constituent atoms, six structures are possible. Aside from the above two butanols, we have:



As an initial exploration of the use of MDG's we have implemented a program, MDGGEN, which can deal with single-step fragmentations in which one or two single bonds are allowed to break, and a user-specified number of neutral hydrogen transfers are allowed into or out of the charged fragment. Because the MS theory used in MDGGEN is so simple the program has limited practical utility, but the work has demonstrated the feasibility of the MDG approach and has helped us to define the major mathematical and algorithmic advances upon which we must focus to arrive at a more general program. Two major topics are indicated.

First is the problem of formalizing the @ operator used to combine MDG's. We now have only a special-case implementation in which possible ways of overlapping MDG's based on a new data point (to yield new MDG's) were determined by hand and supplied to the program. This casewise analysis was hand tailored for the simple MS theory and a similar, though much larger, case library will be created to cover up to 3-bond, 2-step processes. These account for a great many observed peaks in typical mass spectra. The casewise approach is not sufficiently general or flexible for long term MDGGEN development, but will give us the means of creating a useful production program for short term experimentation while we explore the more general MDG overlap problem. The general solution, we believe, can be viewed as a "fuzzy" form of graph matching in which one MDG is mapped onto another with each node of the first matching either nodes or "pieces" of nodes or connected subgraphs of nodes in the second. When we have explored this concept in sufficient depth to construct an efficient, general overlapping algorithm, we will substitute it for the casewise process now in MDGGEN.

Second, we will need to increase our capacity to include constraints in the MDG generation process, constraints both on the structural features of the generated molecules and on the bonds broken in each fragmentation process. Constraints of the former type allow for the specification of desired or undesired structural features which the chemist has deduced either from other sources of structural information or from the chemical history of the unknown.

Some of this information could be incorporated at the beginning of the MDGGEN problem by defining a "starting MDG" which contains desired features. Suppose that in the above C₄H₁₀ example we had known (say, from proton NMR) that the molecule contained two methyl groups. Then rather than starting with the one-noded MDG

C₄H₁₀

we could have started with

CH₃ --- C₂H₄ --- CH₃

and incorporated P₁, P₂ and P₃ into this using the @ operator. The testing of substructural requirements which cannot be entered in this way (e.g., BADLIST items or overlapping GOODLIST entries) will be folded into the generation scheme wherever possible.

Constraints on the cleavage processes allow for greater precision in the specification of the MS rules which make up the theory. They will be incorporated into MDGGEN in two ways. Some constraints, primarily those concerning the allowed number of broken bonds, multiplicities of those bonds, neutral transfers and number of steps, are reflected directly in the choice of MDG structures accounting for a given peak. For example, if the simplest MS theory is used (1-bond, 1-step processes only, no

hydrogen transfers) a peak of mass p will have the MDG representation

$$p \text{ --- } q$$

where q is the remaining composition. However, if single hydrogen transfers are allowed, the MDG for the peak will be any one of

$$p \text{ --- } q \quad p+H \text{ --- } q-H \quad p-H \text{ --- } q+H$$

Constraints reflecting required or prohibited substructural environments for cleaved bonds will be tested for each break of each MDG as the MDG's are generated. The testing algorithms will be the ones used for structural constraints, modified to account for the fact that there are two kinds of bonds in a cleavage constraint: the "break" set, which must be tied to a set of MDG edges that corresponds to a single process, and the "ordinary" set (i.e., ordinary chemical bonds) which are free to associate either with existing MDG edges or with the implied edges inside the MDG nodes.

3.2.4 C13 NMR PLANNER

C13 NMR (CMR) is one of the most rapidly developing and potentially useful spectroscopic techniques for structure elucidation today. The chemical shift of a given carbon atom, even one which is far removed from functional groups, is sensitive to features of the local environment such as branching and steric crowding, features which are difficult to detect using other spectroscopic techniques. Yet CMR data are typically used in structure elucidation studies only to determine gross features of the carbons in the structure such as their hybridization and degree of substitution and whether they neighbor electronegative atoms, primarily nitrogen and oxygen. It should be possible to extract a great deal more structural information from a CMR spectrum automatically, and we propose to explore this possibility. Specifically, we intend to create a CMR planning program analogous to the existing PLANNER which interprets mass spectral data for unknowns.

Like PLANNER, the CMR planning program will assume a basic skeleton (a class) for the unknown and will base its analysis upon a set of class-specific CMR rules relating local environments to observed chemical shifts. There are two sources for such rules. On one hand, rules have been manually extracted from CMR spectra for a variety of simple compound classes, such as acyclic alkanes, amines and alcohols, and poly-methyl cyclohexanes. These rules are available from the literature. On the other hand, our own C13 Meta-DENDRAL research proposed in a later section will be directed toward deducing from sets of spectra of known compounds relationships between local carbon environments and observed shifts. Whatever the source of the rules, the purpose of the CMR planner will be to infer from the

spectrum of an unknown the possible skeletal positions and local environments of each carbon, then to assemble full structures consistent with these possibilities. The assembly stage will also be guided by a user-supplied set of constraints similar to those in CONGEN which will allow him to enter structural information he has deduced from other sources.

In our early work on acyclic amines (39), we have demonstrated the feasibility of an automatic CMR planner for acyclic, monofunctional compounds. The research proposed here will be directed toward the much more complex problem of cyclic, polyfunctional compounds, with our primary interest being the automatic identification of polyfunctional steroids. The AMINE program, though too simple to be generalized directly to such complex cases, has given us valuable experience in dealing with CMR data, particularly in the prediction of the spectra of partially resolved structures and in the testing of such predictions against observed spectra.

Studies from both our laboratories and elsewhere have shown that in relatively rigid molecules such as steroids, CMR chemical shifts are quite sensitive to stereochemistry. This sensitivity will be reflected in the rules, and thus our structure-assembly scheme will need to include some representation of the three-dimensional features of the molecule. This will give the CMR planner the unique ability among DENDRAL programs to distinguish stereochemical isomers of a given topological structure, a step which is usually crucial to the complete solution of a structure elucidation problem. By exploring various representations of stereochemistry in the CMR planner, we expect to develop concepts which will also be useful in CONGEN and other DENDRAL programs.

3.3 New Programs for Theory Formation

3.3.1 Theory Refinement

3.3.1.1 Feedback Loops

The Meta-DENDRAL program (56) has been developed as a single pass program -- molecules and mass spectra are accepted as input data, and rules are generated in one pass through the program. A major step toward increasing the proficiency of the program is to include feedback in the control structure. A program which can notice ambiguities and uncertainties in its rule base might request a certain type of additional input (or select input from some data bank) in order to resolve discrepancies. We intend to provide the existing Meta-DENDRAL program with such abilities.

Initially, we will introduce a feedback mechanism to allow

RULEGEN to apply rules to the input data a second time (with different parameters) so that it can ignore already 'understood' data peaks and focus its attention on the more interesting 'new' data. This modification will allow experimentation with the following new strategy of rule formation:

1. Place a cutoff threshold on the intensity of input data peaks to be considered.
2. Apply existing rules (if any) to the molecule-spectrum pairs to remove 'understood' peaks from consideration. (There will be no existing rules on the first pass.)
3. Generate rules to explain peaks which are above the cutoff. Merge these into the rule base.
4. Lower the intensity cutoff threshold.
5. Go to step 2.

It is anticipated that the above strategy will focus the program's attention on the strongest unexplained peaks at each stage of rule formation. This strategy seems to parallel closely the approach taken by mass spectroscopists when analyzing data.

A second major effort to introduce feedback into the Meta-DENDRAL program will involve allowing the program to select new test data in order to

- (1) increase confidence in existing rules,
- (2) resolve discrepancies or ambiguities in the existing rule base, and
- (3) add rules to broaden the applicability of the rule base.

In order to select new test data intelligently, the program must understand the shortcomings of the current rules. We propose to develop a formalism for concisely representing information about evidence used to support each rule. Information about possible alternate versions of the rules will allow the selection of new data to choose among competing versions of the rule. The formalism will also allow updating each rule incrementally on the basis of the correctness of each new prediction.

3.3.1.2 Alternative Representations for Rules

The rules now formed by the Meta-DENDRAL program are satisfactory codifications of the mass spectrometry processes at a given level of description. Within the model of mass

spectrometry given to it, the program finds very plausible rules. However, the success of the program is tied closely to the adequacy of the underlying model. We propose to investigate means of automatic theory formation in the absence of firm, well accepted models of the domain. The existing Meta-DENDRAL program will provide the framework for this investigation.

One way of reducing the program's dependence on a strictly defined model of the domain is to provide it with the union of terms and concepts which might plausibly contribute to explanatory rules. From this superset of terms, then, the program will be expected to select terms for rules in such a manner that the explanatory power of the rules will be maximized. Terms that contribute nothing to rules will be dropped from the model. For example, the program could discover that a potentially useful descriptive term like electronegativity is never used to explain mass spectrometry data for a class of compounds.

We can improve on the selection process by introducing a hierarchy of terms. For example, there are node and edge properties of subgraphs in a connectivity model and there are geometric properties of subgraphs in a three-dimensional model. We expect to extend the current template schema to describe hierarchies of terms and to select and reject terms from these sets.

An approach that is closer to human theory-formation methods is to give the program models of other disciplines and ask it to construct analogous models of mass spectrometry. Since some of the items in the analogy may be unnecessary when applied to mass spectrometry, the program will need to select the subset of terms that are most helpful. There is no guarantee that this method will work. But its charm lies in providing a mechanism for postulating new concepts for a domain without having to provide a generator of new concepts together with heuristics for determining their worth a priori. For this work on analogical reasoning, which we see as long term research, we would expect to draw largely from the model of theory formation in mathematics proposed in a forthcoming PhD thesis by Mr. Doug Lenat [Stanford University Computer Science Dept.].

3.3.2 C13 NMR Rule Formation

To extend the ideas of theory formation and test the generality of the basic concepts (56) we propose to explore a new problem domain outside of mass spectrometry. The domain of C13 NMR provides an excellent testing ground for generalization of the theory formation program since the format of the rules is significantly different from that of mass spectrometry.

C13 NMR has been characterized as the spectroscopic technique of the 1970's [69]. Our laboratories have been

involved in experimental work on C13 NMR spectra of amines, keto and hydroxy steroids (63-65). In addition, we have carried out a preliminary investigation of a Heuristic DENDRAL approach to interpretation of C13 spectra of amines [39]. Other workers have reported a related approach to the interpretation of hydrocarbon spectra [A.L. Burlingame, R.V. McPherron and D.M. Wilson Proc. Nat. Acad. Sci. USA, 70, 3419 (1973)]. Our aim in exploring C13 NMR rule formation is threefold:

- 1) It will greatly assist chemists who are concerned with formation of explanatory rules for C13 NMR.
- 2) It will be useful for assigning C13 NMR peaks in new spectra to specific carbon atoms in known structures.
- 3) The rules generated by Meta-DENDRAL can be used to infer structures (or partial structures) from C13 NMR data (see C13 Planner section).

There are several parallels between rule formation in mass spectrometry and C13 NMR spectrometry. In both techniques the precise reasons for molecular fragmentation (in the former) or NMR absorption (in the latter) are poorly understood. In the absence of a detailed theory capable of accurate prediction of spectra, we seek empirical rules which can relate observed data to measurable structural parameters. Some of the structural parameters presumed relevant, e.g., atom type, bond multiplicities, are shared in both techniques. Some of the current Meta-DENDRAL structural manipulation functions can be used for either technique. An important difference is that the planning phase of Meta-DENDRAL (i.e., INTSUM) necessary in applications in mass spectrometry is not required for C13 NMR because we will deal initially with spectra whose absorption peaks (or "shifts" relative to an internal standard) are assigned to specific atoms in the known structures. Typically scientists have sought an explanation for the C13 NMR shift of an atom in terms of the structural environment of the atom. Searching such structural environments is a problem which is amenable to solution by existing and proposed parts of the Meta-DENDRAL program.

As in applications to mass spectrometry (56) we will propose a set of factors which might affect C13 NMR absorptions. With a description of these factors we will use the Meta-DENDRAL program to produce a set of rules which will reproduce and predict resonance shifts of individual C13 atoms.

The current Meta-DENDRAL program represents a basic framework for studying C13 NMR rule formation. We believe that the program will require little revision to accommodate the differences in data and rules. We have already considered some of the problems of changing the form of rules. The subgraphs in the descriptive ("situation") parts of rules need to be expanded "outward" from a specific C13 atom instead of outward from a bond

broken in the mass spectrometer. The action parts of rules need to take account of an explicit absorption range whereas for mass spectrometry the rules predict much more precise data points (mass positions). We have made a preliminary test of the program's extensibility in the context of alkanes.

We intend to take the following steps in order to apply Meta-DENDRAL to C13 NMR data for complex molecules:

Incorporate three-dimensional relations among atoms as properties in subgraphs, in addition to connectivity and atom properties now used for rule formation. The preliminary studies on alkanes used only properties of connectivity, but we realize the necessity of describing stereochemical features of complex molecules.

Obtain a program to give us reasonable geometric models for known structures. We are currently looking at model building programs written by Wipke and Allinger [N.L. Allinger, M.T. Tribble, M.A. Miller and D.H. Wertz J. Amer. Chem. Soc., 93, 1637-1648 (1971)] to see if they will fit our needs for this problem.

Study the relationship of conformation and C13 NMR shifts. We intend to start by looking at the C13 NMR spectra of simple fused ring systems in cyclohexanes and decalins, and progress toward our long-range goal of understanding the C13 NMR spectra of steroids.

3.3.3 Further Generalization of Meta-DENDRAL

One of the main motivations of this project is to develop programs and ideas that are applicable to more than a single domain. We propose to extend the generality of the Meta-DENDRAL programs to test the applicability of the knowledge-driven rule formation strategy to other data. We believe the Meta-DENDRAL strategy can be shown to be a useful complement to statistical approaches such as clustering and multiple regression.

Part of the effort of extending Meta-DENDRAL into C13 NMR rule formation will be spent on making the program general enough to work with both mass spectra and C13 NMR spectra, especially since C13 NMR spectral data accumulation is becoming rapidly a routine procedure in many organic laboratories. After this we will have a much better idea of how general our original ideas have been and what restrictions there are on the domains of applicability. A general, model-driven rule formation program will be applied to other medical and biomedical domains that will be selected for their medical relevance and their suitability for the program's development.

3.4 Applications

The attached annual report (Appendix II) summarizes our activities to date involving applications of our instrumentation resource and our programs for computer-assisted structure elucidation to chemical structure problems. These activities have included pursuit of our own mass spectrometric and structural problems, those of other members of the Department of Chemistry, collaboration with several groups in the Stanford Medical School and assistance on problems submitted by a wide variety of persons remote from Stanford who have made use of our facilities. We have so far been able to accommodate almost every request which has been made for use of the mass spectrometer and the computer programs, under the guidelines established in our current grant period.

We indicate in this section the directions we see our own interests in chemical applications taking us. On-going work with local and remote collaborators which will presumably continue into the future is also mentioned. We cannot, however, predict the kinds of applications which current or new users of our facility will bring to us. Much of the work summarized in the annual report was undertaken after informal conversations or correspondence with interested persons. We expect this to continue, we encourage it and we are taking steps (see subsequent section on increased availability) to improve our mechanisms for sharing of our resources in new applications areas.

Important research areas which we know will receive our continuing interest are the following:

3.4.1 Marine Natural Products

Professor Djerassi's laboratory is engaged in intensive structural studies of the organic constituents of marine organisms. The attached annual report (Appendix II) describes the use of DENDRAL facilities including the mass spectrometry resource and CONGEN in structural studies in this area (see also Cheer, et al. ref 59). We propose to continue these studies with special emphasis on elucidating individual structures and the sterol content of several marine organisms. We have chosen mixtures of marine sterols as candidates for computer-assisted analysis for a number of important reasons. First, not only are sterols intrinsically interesting compounds in that they are hormone precursors and important membrane constituents, but sterols derived from marine sources are particularly interesting because a number of sterols found only in marine sources possess very unusual, difficult-to-synthesize structures. These sterols are interesting not only from the standpoint of their potential biological activity and biosynthesis, but also as potential sources for starting materials in difficult steroid hormone syntheses. Second, marine sterol compositions have yielded important information which has helped clarify the phylogenetic

and evolutionary relationships among a number of classes of marine invertebrates. Evidence is now accumulating which indicates that many minor sterol constituents from marine animals are exceptionally stable molecules which have been carried intact through the complex marine food chains. A careful systematic study of marine sterols could therefore not only yield new and important compounds, but at the same time help clarify uncertain evolutionary relationships, and help disentangle complex marine food chains which are of considerable economic as well as scientific relevance. Finally, marine sterols are a fairly homogeneous class of compounds in that (1) marine sterols all possess a common nucleus which results in a number of common mass spectrometric properties; (2) marine sterols all possess very similar chromatographic properties and can be quickly and completely isolated as a single complex fraction which is amenable to a rather thorough separation and analysis by GC/MS; (3) because the fractions are generally complex mixtures (it is not uncommon for a single extract to contain upwards of 30 sterols), a great amount of time is required by highly skilled scientists in the analysis of the GC/MS data for these mixtures.

Routine analysis of the sterol content of a new mixture can be carried out with a computerized GC/MS system which includes facilities for data acquisition and reduction and subsequent library search facilities which make use of spectrum matching and GC relative retention indexes. This will quickly screen out known compounds leaving new components whose structures can be investigated further.

We propose to study new structures in a two-pronged attack using our programs for mass spectral analysis and CONGEN. Specifically, we plan to use the Meta-DENDRAL programs INTSUM, RULEGEN and RULEMOD to assist in the discovery of rules of mass spectral fragmentations to supplement available studies on the influence of side chain and skeletal unsaturation and substitution. These rules will then be used in PLANNER to assist in solving new structures. CONGEN will be used to supplement PLANNER as new features for mass spectral analysis are added to CONGEN.

3.4.2 Analysis of Organic Constituents of Body Fluids

We propose to apply our existing and proposed programs for computer-assisted structure elucidation and our GC/HRMS resource to structural problems of our collaborators in the Department of Genetics. A portion of their research is a metabolic screening program aimed at characterization of organic constituents of body fluids of patients with suspected metabolic disorders of genetic origin. (That work is funded separately under a Genetics Research Center grant, Prof. J. Lederberg, Principal Investigator.) Candidate patients are identified by collaborators of the Center grant, drawing from clinics at Stanford and other area hospitals. Urine samples, occasionally blood, cerebrospinal

fluid or amniotic fluid, are collected from these patients and turned over to Prof. Lederberg's laboratory for analysis. Analytical procedures involve chemical fractionation of the fluid into several fractions, including amino acids, organic acids and sugars. Each fraction is derivatized with appropriate reagents and subjected to GC/LRMS analysis.

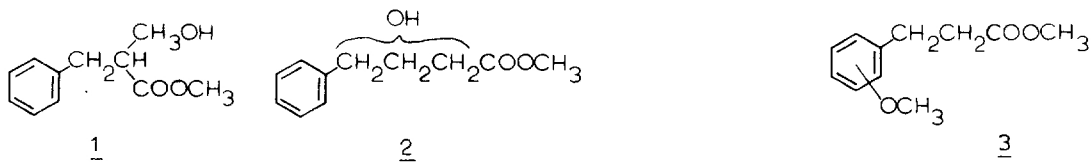
This research is in many ways ideally suited for applications of our techniques. In fact, collaboration with Prof. Lederberg's group has taken place during our current grant period, primarily involving computer programs for processing low resolution mass spectral data subsequent to data collection. Our programs for molecular ion determination and for removal of background and overlapping component interferences (CLEANUP) are part of the standard data processing procedures in Prof. Lederberg's laboratory. We also contributed some effort toward the library search facilities which are common to the mass spectrometry laboratories in Genetics and Chemistry. The analytical procedures and structural identifications in Genetics rely almost exclusively on gas chromatography and on mass spectrometric data. The CLEANUP program produces spectra which compare favorably with spectra present in our library if the compound's spectrum has been previously recorded. However, frequently new components are detected which are not present in the library. There are usually several such components in a given GC/LRMS run. Unidentified components in those experiments present important problems in structure elucidation. They can indicate metabolic abnormalities important to the future treatment of the patients. We feel our current and proposed programs and instrumentation are capable of high enough performance to provide valuable assistance in solution of these problems.

We see collaboration to make use of our facilities proceeding along the following lines: a) GC/HRMS data - empirical formulas are needed to help establish the empirical formula of the compound and of its fragment ions prior to detailed structural analysis. We can provide GC/HRMS data semi-routinely now and will be able to routinely at the outset of our proposed grant; b) CONGEN analysis - CONGEN is now capable of dealing with construction of structural possibilities. We can express many of the constraints which represent knowledge of the biochemical sources of the compounds and the chemistry of the isolation procedures. Improvements and extensions to CONGEN which we propose to implement will simplify analysis of these problems and make it much easier for the person working on a particular problem to use the program; and c) mass spectrum analysis programs - our proposed development of powerful programs for analysis of mass spectra in terms of structure includes a constructive procedure based on mass distribution graphs (see Methods Section 3.2) and the capability for testing candidate structures to determine agreement of predicted spectra with observed. Together, these developments represent a powerful amalgam with CONGEN for study of unknown structures where mass spectrometry is the primary data source.

Recently we have been able to exercise some of the above procedures in the study of unknown compounds as we seek to determine where our instrumentation and programs need further attention. We outline two simple examples which are representative of the approach outlined above. We make no claim for these cases that the results could not be derived manually, but these preliminary studies indicate a strong potential for future applications.

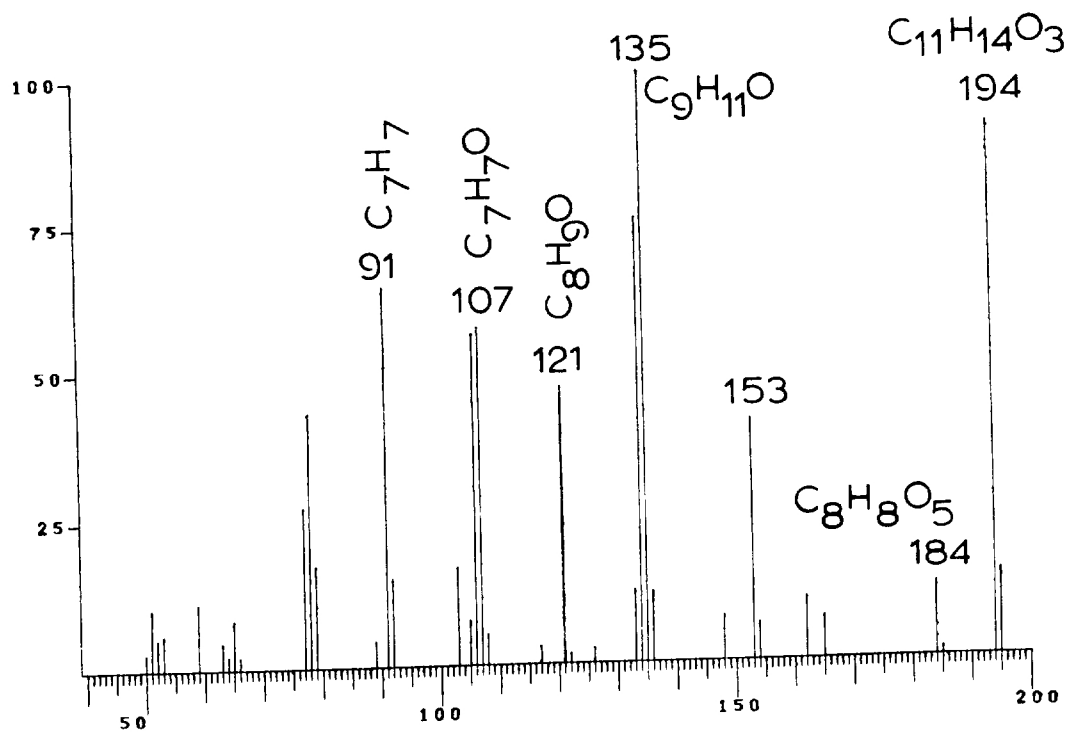
Example 1. The patient was a mentally retarded eleven year old. The organic acid and amino acid fractions of the patient's urine revealed abundant quantities of phenylketo and phenylhydroxy acids and phenylalanine and phenylglutamine. These compounds are characteristic of phenylketonuria (PKU). Further investigation revealed the patient had been born just prior to general screening for PKU and had never been tested subsequently. The organic acid fraction contained several prominent GC peaks which were not identified by library search procedures. Subsequent chemical investigations revealed that some of the unknown GC peaks were artifactual products of the reaction of diazomethane (the derivatizing reagent) and an abundant component, phenylpyruvic acid. A GC/HRMS analysis of this fraction provided the necessary elemental composition information to begin structural analysis of the unknowns.

One new, non-artifactual compound, C₁₁H₁₄O₃, has been analyzed with CONGEN using a variety of constraints and structural fragments inferred from the chemical procedures, the mass spectrum and biochemical knowledge. There are nine plausible structures including branched chain phenylhydroxybutyric acids (e.g., 1) (less likely), straight chain phenylhydroxybutyric acids (2) (questionable) and o, m or p methoxyphenylpropionic acid (3) (all as methyl esters; phenolic hydroxyl groups are etherified under the derivatization conditions).



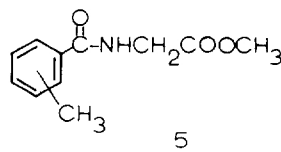
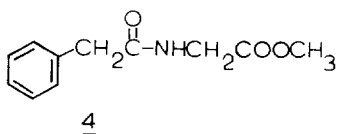
Using recently developed CONGEN functions to predict mass spectra of structures, the set of nine candidates were tested against observed elemental compositions of abundant fragment ions of mass 91 (C₇H₇⁺), 121 (C₈H₉O₁⁺), 135 (C₉H₁₁O₁⁺) and 107 (C₇H₇O₁⁺) (Fig. 2). Only the methoxy-substituted phenylpropionic acids (represented by 3) can yield these ions under reasonable constraints. Comparison of the spectra of authentic standards will soon be carried out to verify our hypothesis. The

Figure 2



biochemical significance of this compound remains to be assessed. Work is continuing on the structures of the artifacts resulting from the derivatization procedure.

Example 2. The urine of a mentally retarded 21 year old was subjected to the same analytical procedures. Abnormal quantities of salicylic acid (*o*-hydroxybenzoic acid), as the *o*-methoxy-methyl ester derivative, were noted in the organic acid fraction. This compound is a metabolite of aspirin so its presence is probably not significant. However, two additional components were present in abundant quantities in this fraction. No record of them was found in our spectral library. The observed low resolution mass spectra, which share similar ions, are presented in Figure 3. GC/HRMS data revealed that the compounds are isomeric, of empirical formula C₁₁H₁₃NO₃. Analysis of structural possibilities with CONGEN yielded 40 structures including a variety of ways of assembling an aromatic ring, a methyl ester and an amide functionality together with two other carbon atoms. Use of mass spectrum prediction functions with a restricted theory of mass spectrometric fragmentation yielded four "most plausible" candidate structures, 4 and three isomers represented by 5.

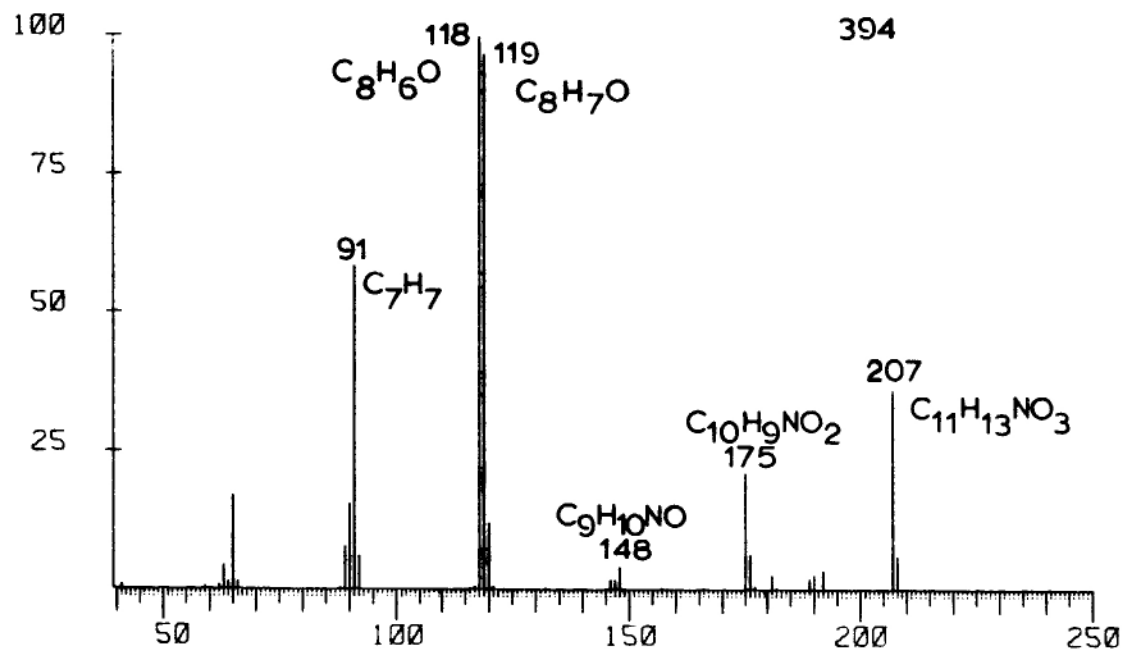


Structure 4 represents a conjugate of phenylacetic acid with glycine, and has been observed in the dog, but never in man. Structures represented by 5 are attractive because closely related isomers, which might yield similar spectra, are possible. However, there are no logical biochemical precursors for such structures. Again, we are attempting to verify our hypothesis by synthesis and comparison of spectra.

In both examples, structures which had not yet been considered by manual interpretation were derived independently by the program. In addition, other, perhaps less plausible, candidates were suggested, which gave the investigator the full set of possibilities to evaluate systematically using whatever additional knowledge or data he/she possessed.

3.4.3 Applications of Reaction Sequences

We have discussed in the Annual Report (Appendix II) our initial steps in development of extensions to CONGEN toward facilities for carrying out in the computer complex sequences of



39A

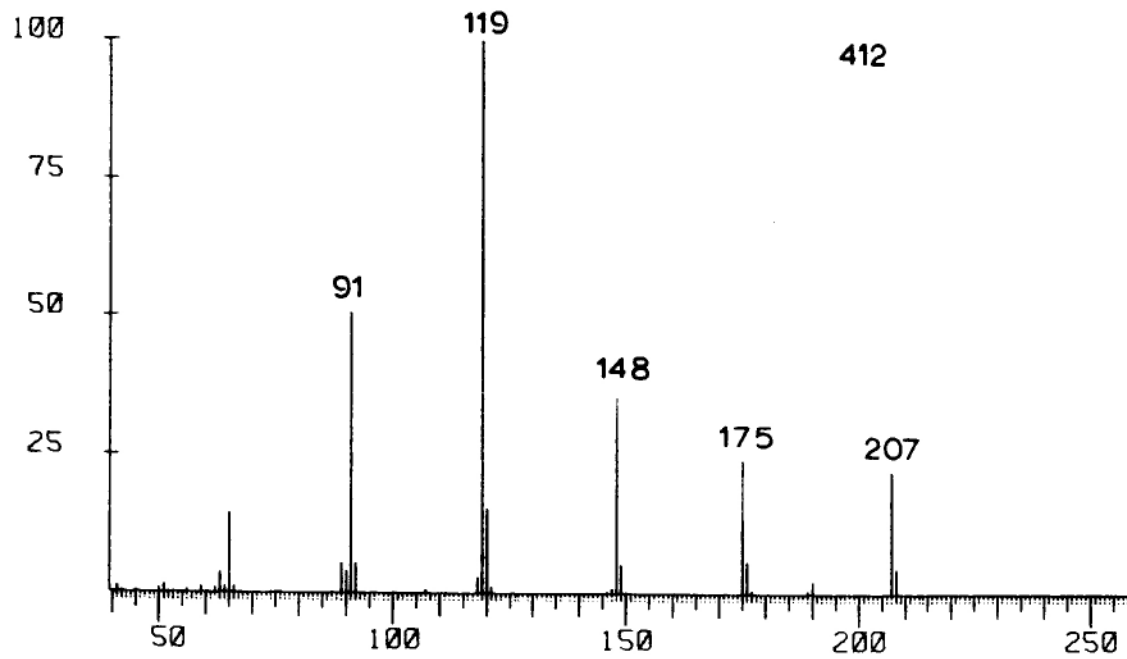
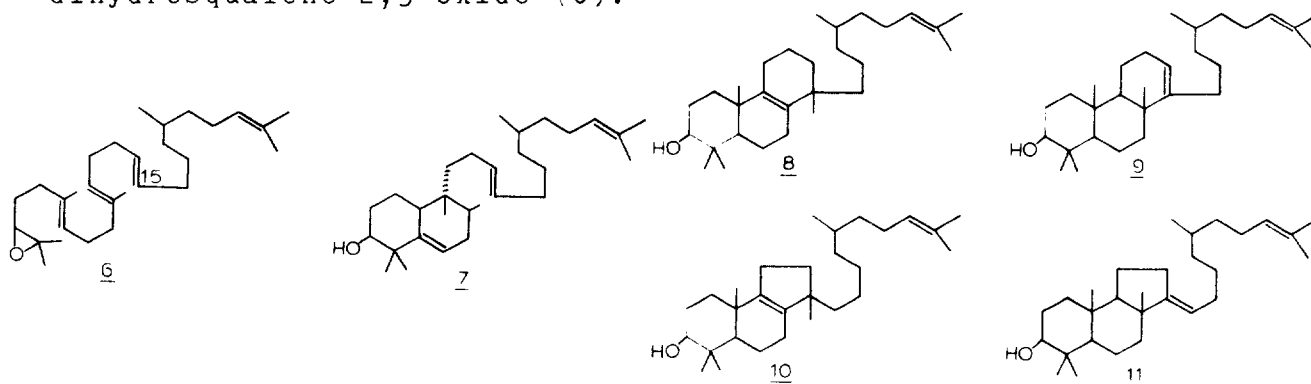


Figure 3

chemical reactions. An initial publication (ref. 61) has described the utility of this approach to structure elucidation problems and to mechanistic studies. A previous section of this proposal has described planned extensions to these facilities for studying reaction sequences.

Structural studies based on reaction sequences open up a broad class of problems of cyclizations and rearrangements to analysis with the assistance of CONGEN. Such studies do not involve assembling structural possibilities from small fragments of the molecule inferred from various data. Rather, the studies are founded on the fact that one begins with a known structure and the products must be related to the known by relatively minor perturbations of that known structure via a set of known reactions. We note that the ability to study reaction sequences also gives us the capability, in principle, to approach structure elucidation by hypothesizing a candidate structure and working toward a closely related solution by judicious manipulation of the candidate. We propose to explore these ideas in areas of current research interest.

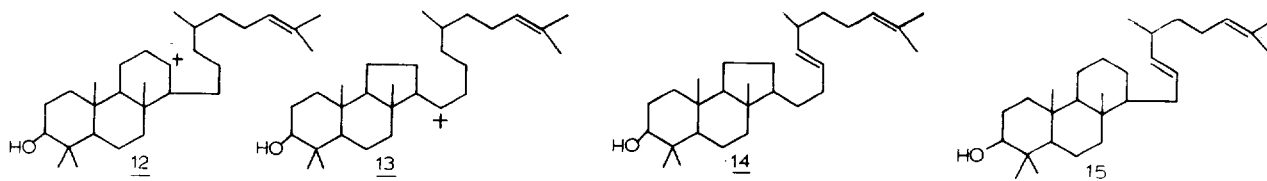
We are currently assisting Prof. van Tamelen's group in a study of unknown cyclization products using our current program. This study, described below, represents a model for an approach which we feel can be extended significantly by new developments proposed in the previous section describing the REACT program. The problem involves unknown structures from both acid and enzyme catalyzed cyclization of a squalene congener, 15'-nor-18,19-dihydrosqualene-2,3-oxide (6).



1) **Acid catalyzed cyclization of (6).** This reaction yielded a complex mixture of bi- and tricyclic alcohols. GC and liquid chromatographic analysis of the mixture yielded ten significant components. The main product was the bicyclic alcohol (7) formed in 25-30 percent yield from (6). In addition to 7, several structures possessing 6-6-5 and 6-6-6 tricyclic ring systems (ring A,B,C, of the steroid nucleus, respectively) were formed. Mass spectral and NMR data gathered on separated unknowns has led to structural suggestions for three of the components, including 8-10.

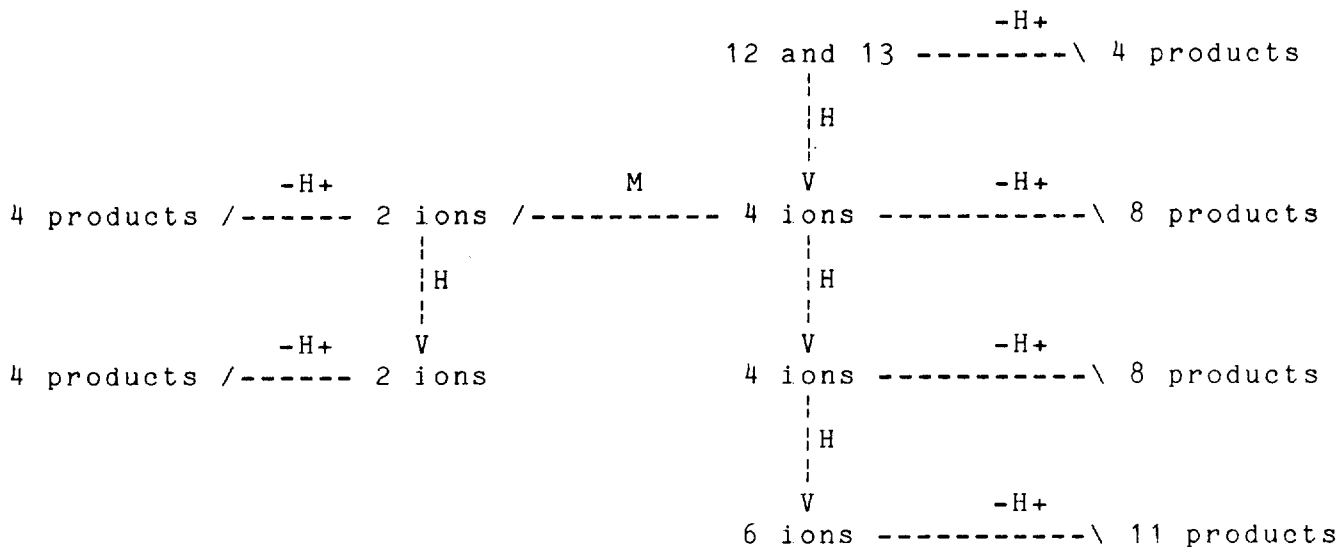
The remaining six structures remain unknowns, although structure 11 has been assigned tentatively to one compound.

We have simulated possible rearrangements of initial cyclization products to yield candidate structures for the remaining unknowns using CONGEN under a variety of constraints. Tricyclic products from such cyclizations almost always yield 6-6-5 or 6-6-6 ring systems. This constraint was used to postulate two tricyclic carbonium ions as starting points for further rearrangement (12 and 13). These carbonium ions were allowed to rearrange via 1,2 shifts of hydrogen atoms or methyl groups, with the terminating condition of loss of H⁺ to yield the observed double bond (all structures are thus tricyclic and possess two additional degrees of unsaturation as two double bonds).



Shifts were allowed only when the resulting carbonium ion possessed the same or higher degree of substitution as its precursor. Collection of products after each step of the following sequence yielded a total of 17 unique final structures, including 8 - 11. The other 13 candidates are under investigation as possible structures for the remaining unknowns.

Rearrangement Processes



Key: -H+ means loss of H+ to yield a double bond in the product.

H means 1,2 hydrogen shift.

M means 1,2 methyl shift.

There are only 17 unique structures; the same structure can be produced in different steps.

2) **Enzymatic Cyclization of 6.** Incubation of 6 with a squalene oxide-lanosterol cyclase preparations obtained from microsomes of rabbit livers yielded several products. One major product was purified by chromatographic techniques and analyzed by mass and NMR spectrometry. The empirical formula was the expected C₂₉H₅₀ and spectral data indicated a tricyclic system.

The unknown was subjected to oxidative cleavage with OsO₄/NaIO₄ to help locate the positions of the double bond. Spectral data collected on the product were strongly suggestive of an aldehyde of molecular formula C₂₀H₃₄O₂, implying loss of a C-9 unit in the oxidative cleavage. This was accompanied by loss of another degree of unsaturation, implying loss of nine terminal atoms in the side chain.

Of the 17 structures from the above simulation of rearrangement processes, only one, 14, has a double bond in a position which would yield a product consistent with the observed data. This structure is an alternative to a manually derived possibility, 15, which necessarily must arise from a more complex rearrangement process. Given these two candidates (14 and 15) it is possible to design experiments to differentiate between them. Further work is now being carried out to solve this problem.

3.4.4 C13 NMR Applications

C13 NMR has been a topic of interest in Professor Djerassi's laboratory for several years and also a subject of some earlier DENDRAL work. Experimental data have been collected for amines, keto steroids and hydroxy steroids (63-65). A computer program was written here [39] which used a set of predictive rules to deduce the structures of acyclic amines. Presently polyhydroxy-steroids are of primary concern in Prof. Djerassi's labs. The formation of rules for the hydroxy-steroids should be an easier task than for the keto-steroids due to the greater structural distortions of the steroid skeleton caused by the latter. A set of rules for the hydroxy-steroids could always be used in the analysis of keto-steroids since these compounds can be chemically reduced to the hydroxy-steroid and analyzed as such. Thus a set of rules for the hydroxy steroids will assist in the analysis of two classes of steroids. A summary of the hydroxy steroid data was given by Eggert (65) which pointed out trends in the data. Presently further studies are being made to assess the effects of steric crowding and skeletal distortions upon the C13 chemical shifts. These studies will aid the Meta-DENDRAL program in the selection of the terms which should contribute to the rule description. This work is being supported in part by NIH grant AM17896.

3.5 Increased Availability

3.5.1 Continued Collaboration and Solicitation of New Efforts

The DENDRAL project, one of the major users of the SUMEX-AIM computer facility, has formed a small community of regular, remote users. This "exodendral" community has continued to provide valuable contributions to program development, although the growth of this community has had to be slowed in response to increasing demands by other projects upon the SUMEX-AIM facility. As an example, for the months of September 1975 to February 1976, the number of CPU hours used by exodendral persons amounted to at least 8 percent of the CPU hours used by the DENDRAL project. There are currently four remote researchers whose groups regularly use CONGEN in their day to day work. Additionally,

there are several remote users who use their accounts on an occasional basis, or who access SUMEX-AIM via the GUEST mechanism.

There have been several applications of our GC/HRMS resource and CONGEN to structural problems of other members of the Stanford community and researchers both in the U.S. and abroad. These collaborations are summarized in the attached annual report (Appendix II). Contacts were made with these people in a variety of ways. We have actively encouraged persons engaged in structure elucidation to consider use of our CONGEN programs (e.g., Drs. Karliner, Nakanishi, Minale). Usually this has involved solution of a previously solved problem to indicate capabilities and limitations, followed by further collaboration on new problems. Informal discussions among scientists at meetings have inspired new applications. We have, in all of our publications, announced that our facilities were available to an outside community of users within the limits of available resources.

We feel our efforts have been successful in encouraging new researchers and solving important problems. To date we have not had to deny use of our facilities to anyone who came to us with a reasonable request. We are currently facing problems of high computer system loading on SUMEX. This restricts the utility of an interactive program like CONGEN due to slow response time. We have requested that people compute in off hours and have added facilities to CONGEN to make this easy to do. Within these resource limitations, we plan to continue existing collaborative efforts and to solicit new collaborators as we have done the past year of our current grant. These collaborations have been an immense benefit in improving our GC/HRMS resource and CONGEN. Our facilities have had to confront real-world problems with all the attendant uncertainties and assumptions characteristic of such problems. This experience has provided the background for our future plans in making our facilities more widely available. With a growing user community, additional burdens are placed on the GC/HRMS and SUMEX resources. Some of our proposed new program developments are directed specifically to easing these burdens. There are, however, several additional ways of increasing availability, especially of CONGEN and new extensions to it, which are discussed in the subsequent section. Recent applications of our programs to structural problems of our collaborators are summarized in Appendix II, Section 3.4.

3.5.2 Program Translation

Translation of CONGEN

Although it has proven to be a useful research tool for chemists, CONGEN is considerably larger and more time-consuming than it could be. Its development has been an evolution involving the work of many people over several years, and most of

it is written in INTERLISP, a language which promotes rapid program development but which is not noted for its run-time efficiency. In retrospect we feel that this was the proper course - we could not have reached the current level of complexity and sophistication in CONGEN otherwise - but the result is a production-level program which, because it was not designed as a single, efficient package, is rather wasteful of computer resources.

Because of the demands which CONGEN places upon SUMEX, we probably will not be able to offer use of the program to all who have fruitful applications. Not only does this deprive the chemical community of a useful tool, but it limits the number and variety of new applications to guide us in further program developments. We propose to ease this problem by recoding CONGEN in a more efficient and exportable computer language. Greater efficiency will increase CONGEN's productivity, allowing us to offer the resource to more users at SUMEX, while exportability will allow others to transfer the program to their facilities, relieving SUMEX of the burden of supplying access for routine, non-developmental use.

The language chosen for the translation is the ALGOL subset of SAIL. This choice was made for four reasons. First, an ALGOL-like language is preferable to FORTRAN because the former allows recursive programming techniques to be used. It would be possible to rework CONGEN in terms of non-recursive algorithms, but recursion is such an integral part of the logic of the program that such a transformation would be quite difficult. Second, although probably not as efficient as FORTRAN, the SAIL compiler creates reasonably fast and compact machine code. We have done some experimental translations of a few key segments of CONGEN and find a 10- to 15-fold improvement in running time, an improvement which will greatly ease the impact of CONGEN on SUMEX. Thirdly, SAIL is designed for the standard TOPS-10 operating system on the PDP-10, a fairly common research computer configuration accessible to a large number of chemists at both universities and industrial research facilities. We believe that such outside users will have relatively little difficulty mounting a SAIL version of CONGEN on their local facilities. Finally, compared to LISP, ALGOL is a more widely known language by itself or as the basis for other languages such as PL/1, PASCAL and SIMULA. In the ALGOL subset of SAIL, CONGEN will be significantly easier both to understand and to modify by interested non-Stanford workers.

One other reason for selecting SAIL warrants special mention. A proposal has recently been submitted as an extension of the SUMEX grant to develop a machine-independent language called MAINSAIL which, in many respects, is quite close to SAIL. Particularly, the subset which we will be using for CONGEN is virtually identical between the two languages, and transferring CONGEN from SAIL to MAINSAIL would not be a major task. One design criterion of MAINSAIL is transferability from one type of

computer to another - all that is needed for a new machine is a "MAINSAIL bootstrap" package to define basic machine operations and input-output characteristics. A preliminary version is now available for the PDP-10 under both the TENEX and TOPS-10 operating systems, and for the PDP 11/45 under the RT-11 system. A bootstrap package is being designed for ORVYL, the local time-sharing monitor for the IBM 370/168, as well. Although we are not specifically proposing the coding of CONGEN in MAINSAIL because the funding of the MAINSAIL effort is not yet certain, we are aware of that effort and will maximize MAINSAIL compatibility as we proceed with the CONGEN translation. When MAINSAIL matures to a stable and widely-available language, we feel that it will provide the ideal mechanism for implementing CONGEN on a variety of other machines including smaller laboratory systems such as the 11/45.

There are two existing facilities which will ease the translation and will enable us to reach a workable balance between the run-time efficiency of SAIL and the program-development aids of INTERLISP. First, the structure of the TENEX operating system allows one to run simultaneously two or more sections of a program written in different languages, with communication between them taking place through a shared file or a shared segment of memory. This means that not all of CONGEN needs to be translated at once. Rather, it can be transferred a piece at a time from INTERLISP to SAIL. Not only will this ease the problems of debugging a large and complex system, but it will allow us to retain the more rapidly-changing developmental portions of CONGEN in INTERLISP for as long as possible as the more stable sections are translated. Even when all of CONGEN has been transferred to SAIL, we expect to maintain a SAIL-INTERLISP interface so that new ideas may be tested easily in the latter. The prototype for this "pipeline" between the two languages already exists in the linkage between CONGEN and the SAIL program responsible for fragment imbedding and structure canonicalization. The SAIL segment contains a monitor program plus a set of modules which the monitor can call. The INTERLISP segment passes data and control information to the monitor, and collects output from it. We will retain this structure so that even when essentially all of the control is given to the SAIL portion, it will still be possible to "call" INTERLISP for specialized or experimental types of processing. Of course, this mechanism will not be used in any export version of CONGEN, but it will substantially enhance the flexibility of the system for local research.

The second existing facility is a cross-compiler we have developed which translates a specialized ALGOL-like subset of INTERLISP into SAIL. The subset, called SAILISP, can be used to create and test ALGOL programs in the highly interactive and well engineered environment of INTERLISP. Once a program or portion thereof has been perfected in SAILISP, the cross-compiler is used to translate it automatically into SAIL code which can be compiled and run in the normal fashion. Though SAILISP does not

provide easy processing of linked lists, a central concept in the LISP language, it does allow a programmer to build a system interactively in small pieces, debugging and modifying each piece using the powerful INTERLISP editor and error handling package. We have found that the ease of programming in INTERLISP results as much from these interactive aids as it does from the basic structure of the language itself, and SAILISP makes these aids available for SAIL programming. In conjunction with the SAIL-INTERLISP interface described above, SAILISP will provide a well balanced system not only during the recoding of the existing algorithms, but for future CONGEN research.

3.5.3 GC/HRMS System

We will continue to run samples under our current guidelines which stress that the facility is to be used for important structural problems of biomedical relevance, but not for obtaining routine mass spectra from crude reaction mixtures. Within these guidelines we have been able to entertain nearly all requests for spectra while continuing our active program of instrument and program development. As this development requires less and less instrument and computer time, additional time will be available for obtaining high resolution and GC/HR mass spectra. We are already taking advantage of this available time in our own research on marine natural products and our collaborations with local persons at Stanford. We should have more flexibility in the future, however, and we will encourage our remote collaborators to make use of our facilities for GC/HRMS to help solve their structural problems.

3.6 The GC/HRMS Resource

In previous sections we discussed the use of the GC/HRMS resource as a tool to provide necessary data for our structural studies. We also discussed the probable increased availability of the system as time required for development decreases. We propose to devote our attention to maintenance of the system and development of a detailed understanding of its performance in a variety of applications. We also propose some further developments to improve the sensitivity and throughput of the system.

Although maintenance of a system may seem trivial, in fact maintenance goes far beyond actually keeping all the parts in working order. It means having a trained operator who can take precautionary measures to avoid down time and who can recognize when performance is deteriorating, however slightly. It means devotion of significant programmer time to carry out modifications to existing software because new chemical problems frequently require new data reduction techniques.

The developments we propose are simple in concept but are potentially very valuable. They are described in the following subsections.

3.6.1 Increased Sensitivity

We propose to develop the data reduction tools required to scan spectra at lower resolving powers. We know from past studies (A.L. Burlingame, D.H. Smith, T.O. Merren, and R.W. Olsen, in "Computers in Analytical Chemistry," (Vol. 4 in Progress in Analytical Chemistry series), C. H. Om and J. Norris, Eds., Plenum Press, New York, N.Y., 1970, p. 17) that mass measurement accuracy (and thus the certainty with which elemental compositions can be assigned) decreases only slightly in scanning at lower resolving powers. The sensitivity change in reducing resolving power can be dramatic, at least a factor of ten in going from a resolving power of 10,000 to 1,000 on the Varian-MAT 711. Obviously, it would be better to operate the instrument at lower resolving powers, except that problems arise because spectral peaks which were resolved at high resolving powers may overlap at low resolving powers. We routinely operate at resolving powers of 4,000 to 5,000 in GC/HRMS mode. We have found it necessary even at these moderate resolutions to implement a scheme for doublet resolution (see Appendix II, Annual Report, for a detailed description) to separate ions from the reference compound from those of GC column bleed and the sample. This approach has generally proven sufficient because in most of our applications, overlapping triplets or higher multiplets of ions are unlikely. At lower resolving powers, however, we know that the simple doublet resolver will be insufficient. Therefore, we propose to implement a multiplet resolver effectively to restore some of the resolution lost by the mass spectrometer.

Multiplet resolution techniques applied to mass spectral and many other types of data have been reported for years. We propose a seemingly minor, but critical, twist to these procedures, namely, using a peak model based on measurement of actual mass spectral peaks immediately previous in the scan as the basis for performing this resolution. Drawbacks to multiplet resolution procedures include the facts that they are time consuming and that almost every procedure employs an assumed "ideal" peak shape. We can do nothing about the extra time required for data reduction, but we think the increased sensitivity more than justifies it. But we have found in all our efforts toward evaluation of instrument performance and doublet resolution that an accurate and reliable system must be based on the measured performance (e.g., peak shape, dynamic resolution, etc.) of the mass spectrometer, not an idealized model. Thus, we will use a peak model based on measured peaks which are presumed singlets as the basis for multiplet resolution.

3.6.2 User Interface

We will improve the facilities for examining the large volumes of data produced in a GC/HRMS experiment so that the person whose sample was run can explore his own results. We have many of the file handling routines to recover easily various experimental results and the display routines to display on a CRT or produce on a hard copy plotter any of a variety of results which can be derived from a scan from calibration data to final assigned elemental compositions. We propose to provide simple procedures for examining these data, doing library searches and performing inter-experiment comparisons of results. This will increase the throughput of the laboratory because the examination of data can be done in the off hours, leaving more prime time available for running additional samples.

4 BIBLIOGRAPHY

DENDRAL PUBLICATIONS

- (1) J. Lederberg, "DENDRAL-64 - A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs", (technical reports to NASA, also available from the author and summarized in (12)). (1a) Part I. Notational algorithm for tree structures (1964) CR.57029 (1b) Part II. Topology of cyclic graphs (1965) CR.68898 (1c) Part III. Complete chemical graphs; embedding rings in trees (1969)
- (2) J. Lederberg, "Computation of Molecular Formulas for Mass Spectrometry", Holden-Day, Inc. (1964).
- (3) J. Lederberg, "Topological Mapping of Organic Molecules", Proc. Nat. Acad. Sci., 53:1, January 1965, pp. 134-139.
- (4) J. Lederberg, "Systematics of organic molecules, graph topology and Hamilton circuits. A general outline of the DENDRAL system." NASA CR-48899 (1965)
- (5) J. Lederberg, "Hamilton Circuits of Convex Trivalent Polyhedra (up to 18 vertices), Am. Math. Monthly, May 1967.
- (6) G. L. Sutherland, "DENDRAL - A Computer Program for Generating and Filtering Chemical Structures", Stanford Artificial Intelligence Project Memo No. 49, February 1967.

- (7) J. Lederberg and E. A. Feigenbaum, "Mechanization of Inductive Inference in Organic Chemistry", in B. Kleinmuntz (ed) Formal Representations for Human Judgment, (Wiley, 1968) (also Stanford Artificial Intelligence Project Memo No. 54, August 1967).
- (8) J. Lederberg, "Online computation of molecular formulas from mass number." NASA CR-94977 (1968)
- (9) E. A. Feigenbaum and B. G. Buchanan, "Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry", in Proceedings, Hawaii International Conference on System Sciences, B. K. Kinariwala and F. F. Kuo (eds), University of Hawaii Press, 1968.
- (10) B. G. Buchanan, G. L. Sutherland, and E. A. Feigenbaum, "Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry". In Machine Intelligence 4 (B. Meltzer and D. Michie, eds) Edinburgh University Press (1969), (also Stanford Artificial Intelligence Project Memo No. 62, July 1968).
- (11) E. A. Feigenbaum, "Artificial Intelligence: Themes in the Second Decade". In Final Supplement to Proceedings of the IFIP68 International Congress, Edinburgh, August 1968 (also Stanford Artificial Intelligence Project Memo No. 67, August 1968).
- (12) J. Lederberg, "Topology of Molecules", in The Mathematical Sciences - A Collection of Essays, (ed.) Committee on Support of Research in the Mathematical Sciences (COSRIMS), National Academy of Sciences - National Research Council, M.I.T. Press, (1969), pp. 37-51.
- (13) G. Sutherland, "Heuristic DENDRAL: A Family of LISP Programs", Stanford Artificial Intelligence Project Memo No. 80, March 1969.
- (14) J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference I. The Number of Possible Organic Compounds: Acyclic Structures Containing C, H, O and N". Journal of the American Chemical Society, 91.
- (15) A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference II. Interpretation of Low Resolution Mass Spectra of Ketones". Journal of the American Chemical Society, 91:11 (May 21, 1969).
- (16) B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, "Toward an Understanding of Information Processes of Scientific

- Inference in the Context of Organic Chemistry", in Machine Intelligence 5, (B. Meltzer and D. Michie, eds) Edinburgh University Press (1970), (also Stanford Artificial Intelligence Project Memo No. 99, September 1969).
- (17) J. Lederberg, G. L. Sutherland, B. G. Buchanan, and E. A. Feigenbaum, "A Heuristic Program for Solving a Scientific Inference Problem: Summary of Motivation and Implementation", Stanford Artificial Intelligence Project Memo No. 104, November 1969.
- (18) C. W. Churchman and B. G. Buchanan, "On the Design of Inductive Systems: Some Philosophical Problems". British Journal for the Philosophy of Science, 20 (1969), pp. 311-323.
- (19) G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference III. Aliphatic Ethers Diagnosed by Their Low Resolution Mass Spectra and NMR Data". Journal of the American Chemical Society, 91.
- (20) A. Buchs, A. M. Duffield, G. Schroll, C. Djerassi, A. B. Delfino, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Applications of Artificial Intelligence For Chemical Inference. IV. Saturated Amines Diagnosed by Their Low Resolution Mass Spectra and Nuclear Magnetic Resonance Spectra", Journal of the American Chemical Society, 92, 6831 (1970).
- (21) Y.M. Sheikh, A. Buchs, A.B. Delfino, G. Schroll, A.M. Duffield, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference V. An Approach to the Computer Generation of Cyclic Structures. Differentiation Between All the Possible Isomeric Ketones of Composition C₆H₁₀O", Organic Mass Spectrometry, 4, 493 (1970).
- (22) A. Buchs, A.B. Delfino, A.M. Duffield, C. Djerassi, B.G. Buchanan, E.A. Feigenbaum and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference VI. Approach to a General Method of Interpreting Low Resolution Mass Spectra with a Computer", Helvetica Chimica Acta, 53, 1394 (1970).
- (23) E.A. Feigenbaum, B.G. Buchanan, and J. Lederberg, "On Generality and Problem Solving: A Case Study Using the DENDRAL Program". In Machine Intelligence 6 (B. Meltzer and D. Michie, eds.) Edinburgh University Press (1971). (Also Stanford Artificial Intelligence Project Memo No. 131.)
- (24) A. Buchs, A.B. Delfino, C. Djerassi, A.M. Duffield, B.G.

- Buchanan, E.A. Feigenbaum, J. Lederberg, G. Schroll, and G.L. Sutherland, "The Application of Artificial Intelligence in the Interpretation of Low-Resolution Mass Spectra", *Advances in Mass Spectrometry*, 5, 314 (1971),
- (25) B.G. Buchanan and J. Lederberg, "The Heuristic DENDRAL Program for Explaining Empirical Data". In proceedings of the IFIP Congress 71, Ljubljana, Yugoslavia (1971). (Also Stanford Artificial Intelligence Project Memo No. 141.)
- (26) B.G. Buchanan, E.A. Feigenbaum, and J. Lederberg, "A Heuristic Programming Study of Theory Formation in Science." In proceedings of the Second International Joint Conference on Artificial Intelligence, Imperial College, London (September, 1971). (Also Stanford Artificial Intelligence Project Memo No. 145.)
- (27) Buchanan, B. G., Duffield, A.M., Robertson, A.V., "An Application of Artificial Intelligence to the Interpretation of Mass Spectra", *Mass Spectrometry Techniques and Appliances*, G. W. A. Milne, Ed., John Wiley & Sons, Inc., 1971, p. 121.
- (28) D.H. Smith, B.G. Buchanan, R.S. Engelmores, A.M. Duffield, A. Yeo, E.A. Feigenbaum, J. Lederberg, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference VIII. An Approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids", *Journal of the American Chemical Society*, 94, 5962 (1972).
- (29) B.G. Buchanan, E.A. Feigenbaum, and N.S. Sridharan, "Heuristic Theory Formation: Data Interpretation and Rule Formation". In *Machine Intelligence 7*, Edinburgh University Press (1972).
- (30) J. Lederberg, "Rapid Calculation of Molecular Formulas from Mass Values". *Journal of Chemical Education*, 49, 613 (1972).
- (31) H. Brown, L. Masinter, and L. Hjelmeland, "Constructive Graph Labeling Using Double Cosets". *Discrete mathematics*, 7, 1 (1974). (Also Computer Science Memo 318, 1972).
- (32) B. G. Buchanan, Review of Hubert Dreyfus' "What Computers Can't Do: A Critique of Artificial Reason", *Computing Reviews* (January, 1973). (Also Stanford Artificial Intelligence Project Memo No. 181)
- (33) D. H. Smith, B. G. Buchanan, R. S. Engelmores, H. Adlercreutz and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference IX. Analysis of Mixtures Without Prior Separation as Illustrated for Estrogens". *Journal of the American Chemical Society* 95, 6078 (1973).

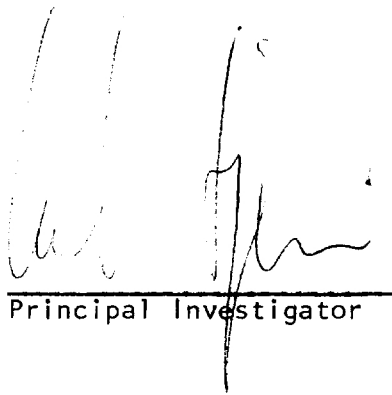
- (34) D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference X. Intsum. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids". *Tetrahedron*, 29, 3117 (1973).
- (35) B. G. Buchanan and N. S. Sridharan, "Rule Formation on Non-Homogeneous Classes of Objects". In proceedings of the Third International Joint Conference on Artificial Intelligence (Stanford, California, August, 1973). (Also Stanford Artificial Intelligence Project Memo No. 215.)
- (36) D. Michie and B.G. Buchanan, "Current Status of the Heuristic DENDRAL Program for Applying Artificial Intelligence to the Interpretation of Mass Spectra", in "Computers for Spectroscopy," R.A.G. Carrington, Ed., Adam Hilger, London, 1973. Also: University of Edinburgh, School of Artificial Intelligence, Experimental Programming Report No. 32 (1973).
- (37) H. Brown and L. Masinter, "An Algorithm for the Construction of the Graphs of Organic Molecules", *Discrete Mathematics*, 8, 227 (1974). (Also Stanford Computer Science Dept. Memo STAN-CS-73-361, May, 1973)
- (38) D.H. Smith, L.M. Masinter and N.S. Sridharan, "Heuristic DENDRAL: Analysis of Molecular Structure," in "Computer Representation and Manipulation of Chemical Information," W.T. Wipke, S. Heller, R. Feldmann and E. Hyde, Eds., John Wiley and Sons, Inc., 1974, p. 287.
- (39) R. Carhart and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference XI: The Analysis of C13 NMR Data for Structure Elucidation of Acyclic Amines", *Journal of the Chemical Society (Perkin II)*, 1753 (1973).
- (40) L. Masinter, N.S. Sridharan, R. Carhart and D.H. Smith, "Application of Artificial Intelligence for Chemical Inference XII: Exhaustive Generation of Cyclic and Acyclic Isomers". *Journal of the American Chemical Society*, 96, 7702 (1974). (Also Stanford Artificial Intelligence Project Memo No. 216.)
- (41) L. Masinter, N.S. Sridharan, R. Carhart and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XIII. Labeling of Objects having Symmetry". *Journal of the American Chemical Society*, 96, 7714 (1974).
- (42) N.S. Sridharan, Computer Generation of Vertex Graphs, Stanford CS Memo STAN-CS-73-381, July, 1973.
- (43) N.S. Sridharan, et.al., A Heuristic Program to Discover Syntheses for Complex Organic Molecules, Stanford CS Memo

- STAN-CS-73-370, June, 1973. (Also Stanford Artificial Intelligence Project Memo No. 205.)
- (44) N.S. Sridharan, Search Strategies for the Task of Organic Chemical Synthesis, Stanford CS Memo STAN-CS-73-391, October, 1973. (Also Stanford Artificial Intelligence Project Memo No. 217.)
- (45) R. G. Dromey, B. G. Buchanan, J. Lederberg and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIV. A General Method for Predicting Molecular Ions in Mass Spectra". Journal of Organic Chemistry, 40, 770 (1975).
- (46) D. H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XV. Constructive Graph Labelling Applied to Chemical Problems. Chlorinated Hydrocarbons". Analytical Chemistry, 47, 1176 (1975).
- (47) R. E. Carhart, D. H. Smith, H. Brown and N. S. Sridharan, "Applications of Artificial Intelligence for Chemical Inference. XVI. Computer Generation of Vertex Graphs and Ring Systems". Journal of Chemical Information and Computer Science, 15, 124 (1975).
- (48) R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XVII. An Approach to Computer-Assisted Elucidation of Molecular Structure". Journal of the American Chemical Society, 97, 5755 (1975).
- (49) B. G. Buchanan, "Scientific Theory Formation by Computer." In Proceedings of NATO Advanced Study Institute on Computer Oriented Learning Processes, 1974, Bonas, France.
- (50) E. A. Feigenbaum, "Computer Applications: Introductory Remarks," in "Proceedings of Federation of American Societies for Experimental Biology," 33, 2331 (1974).
- (51) R. Davis and J. King, "Overview of Production Systems" To appear in Machine Representation of Knowledge, Proceedings of the NATO ASI Conference, July, 1975. (Also Stanford Artificial Intelligence Project Memo .)
- (52) B. G. Buchanan, "Applications of Artificial Intelligence to Scientific Reasoning." In Proceedings of Second USA-Japan Computer Conference, American Federation of Information Processing Societies Press, August, 1975.
- (53) R. E. Carhart, S. M. Johnson, D. H. Smith, B. G. Buchanan, R. G. Dromey, J. Lederberg, "Networking and a Collaborative Research Community: A Case Study Using the DENDRAL Program," in "Computer Networking and Chemistry", P. Lykos, Ed., American Chemical Society, Washington, D.C., 1975, p. 192.

- (54) D. H. Smith, "The Scope of Structural Isomerism" (Paper XVIII in our series of AI Applications in Chemistry). *Journal of Chemical Information and Computer Science*, 15, 203 (1975).
- (55) D. H. Smith, J. P. Konopelski and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIX. Computer Generation of Ion Structures." *Organic Mass Spectrometry*, 11 (1976) 86.
- (56) B. G. Buchanan, D. H. Smith, W. C. White, R. Gritter, E. A. Feigenbaum, J. Lederberg and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XXII. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program." *Journal of the American Chemical Society*, in press.
- (57) E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green and S. N. Cohen, "Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System." *Computers and Biomedical Research* 8, 303-320 (1975).
- (58) R. Davis, B. Buchanan and E. Shortliffe, "Production Rules as a Representation for a Knowledge-Based Consultation Program", accepted for publication by *Artificial Intelligence*. (Also Stanford Artificial Intelligence Project Memo No. AIM-266.)
- (59) R.E. Carhart and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XX. 'Intelligent' Use of Constraints in Computer-Assisted Structure Elucidation," *Computers in Chemistry*, in press.
- (60) C. Cheer, D.H. Smith, C. Djerassi, B. Tursch, J.C. Braekman, and D. Dalozé, "Applications of Artificial Intelligence for Chemical Inference. XXI. Chemical Studies of Marine Invertebrates. XVII. The Computer-Assisted Identification of [+] -Palustrol in the Marine Organism *Cespitularia* sp., aff. *Subviridis*," *Tetrahedron*, in press.
- (61) T.R. Varkony, R.E. Carhart, and D.H. Smith, "Computer-Assisted Structure Elucidation. Modelling Chemical Reaction Sequences Used in Molecular Structure Problems," in "Computer-Assisted Organic Synthesis," W.T. Wipke, Ed., American Chemical Society, Washington, D.C., in press.
- (62) D.H. Smith and R. E. Carhart, "Structural Isomerism of Mono- and Sesquiterpenoid Skeletons," *Tetrahedron*, in press.
- (63) H. Eggert and C. Djerassi, "The Carbon-13 Magnetic Resonance Spectra of Acyclic Aliphatic Amines," *Journal of American Chemical Society*, 95, 3710 (1973).

- (64) H. Eggert and C. Djerassi, "Carbon-13 Nuclear Magnetic Resonance Spectra of Keto Steroids," *Journal of Organic Chemistry*, 38, 3788 (1973).
- (65) H. Eggert, C. VanAntwerp, N. Bhacca and C. Djerassi, "Carbon-13 Nuclear Magnetic Resonance Spectra of Hydroxy Steroids," *Journal of Organic Chemistry*, 41, 71 (1976).
- (66) S. Hammerum and C. Djerassi. "Mass Spectrometry in Structural and Stereochemical Problems CCXLV. The Electron Impact Induced Fragmentation Reactions of 17-oxygenated Progesterones." *Steroids*, 25, 817 (1975).
- (67) S. Hammerum and C. Djerassi. "Mass Spectrometry in Structural and Stereochemical Problems CCXLIV. The Influence of Substituents and Stereochemistry on the Mass Spectral Fragmentation of Progesterone." *Tetrahedron*, 31, 2391 (1975).
- (68) L. L. Dunham, C. A. Henrick, D. H. Smith, and C. Djerassi, "Mass Spectrometry in Structural and Stereochemical Problems. CCXLVI. Electron Impact Induced Fragmentation of Juvenile Hormone Analogs," *Org. Mass Spectrom.*, in press.
- (69) C. Djerassi, Foreword to "13C NMR-Spectroscopy," by E. Breitmayer and W. Voelter, Verlag Chemie GmbH, Weinheim/Bergstr., 1974.
- (70) R. G. Dromey, M. J. Stefik, T. Rindfleisch, and A. M. Duffield, "Extraction of Mass Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography/Mass Spectrometry Data," *Analytical Chemistry*, in press.

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



Principal Investigator

5/26/76

Date

5 Appendix I

Details of Proposed HELP SYSTEM.

1) On-line documentation system

We designed CONGEN without a highly structured interface between a researcher and the program. This provides a great deal of flexibility in the ways the program can be used to solve a given problem. But this lack of structure can result in a feeling of helplessness when a researcher has little idea of what to do next. The printed document is usually inadequate; it is too long to find a necessary piece of information quickly.

A recent formulation of guidelines for humanizing computerized information systems (T.D. Sterling, Science, 190, (1975), p. 1168), places particular emphasis on the importance of permitting scientists to control interaction with the program. Illustrations of the sorts of help we propose are in the programs MLAB and Interlisp-Masterscope. A demonstration of this concept in the current version of CONGEN is found in the "information interrupts" described previously. In keeping with this user-driven form of interaction, it has become obvious that a flexible, on-line help system in the form of access to the information contained in the document is necessary.

We will rearrange the document into a form that will make it serviceable as a help file, as well as more readable as a reference. Requests for assistance to CONGEN will result in accessing the help file for a summary of what is useful to do at that point, or what commands can be used, or what format is necessary for a given command. Options will be provided for a more detailed description if the researcher finds it necessary for clarification.

2) Tutorial error handling

A new tutorial error handling portion of CGHELP will rely heavily upon a flexible error-detection mechanism in CONGEN, one which is significantly easier to work with than the current version. Errors in CONGEN are now perceived in the traditional manner: built into the code at many different points there are various consistency checks, and when one or more of these is violated, an error message is printed and corrective action is taken (this is usually a simple return to the top-level prompt of the program). This approach has become increasingly more cumbersome as CONGEN has been extended. As each new concept is added to the system all possible conflicts with old concepts must be considered and a progressively larger number of new tests must be added throughout the code. To alleviate these difficulties and to lay the foundation for other CGHELP developments, we plan a completely new approach to the problem of error detection, one which draws upon a knowledge base, external to CONGEN itself, of error conditions. Philosophically, this amounts to a realization that error checking can be dealt with as an activity quite apart from the symbol-manipulation algorithms of the main program.

We intend to formulate this separate error-checking program as a production system which will process each input from the user. In this system, all documented knowledge about CONGEN will be represented internally as a set of situation-action rules. The situation of each rule will be a condition which must not occur during a CONGEN session, and the action portion will be executed whenever the control program detects that a given situation is satisfied. In the first implementation, each action will simply cause an error message to be printed and will provide the "tutor" with information concerning pertinent sections of the document. However, further CGHELP developments (see below) will depend heavily on more sophisticated actions, and in fact the production system will form the core of the "intelligent" aspects of the entire CGHELP system.

The flexibility of the production system format here will allow us to approach the error-detection problem in a general context. The rules themselves must, of course, represent specific knowledge about CONGEN, but the elements of the control system (i.e., the portion of the program which controls testing and evaluation of the rules) will contain the protocols for printing messages and guiding the tutorial interaction. To apply the system to a new program, we will need a new knowledge base and on-line document, but many of the details about interacting with users in a tutorial mode will be directly transferrable.

3) Internal model of the user

In order to accomplish the "tutoring" outlined above without seeming overly solicitous or overly presumptuous, the error handling system will clearly need some model of the user to guide it. For example, an experienced user who types poorly would quickly tire of frequent offers to display the menu of available commands, but to a novice such offers could be quite useful. The user model we plan for CGHELP will contain an internal representation of the user's knowledge of various key concepts in the program, and as he gains new information, either through use of the on-line document or via tutorial error handling, the model will be updated. The tutoring process will then be coupled strongly to this model so that access to the document is offered only when CGHELP perceives the topic as one which has not frequently (or recently) been touched upon.

The coupling of CGHELP to the user model will again be accomplished via the production system concept. The action portion of the error-testing rules mentioned above will be modified so that they cause no direct user interaction. Rather, they will cause internal assertions to be made that an error has occurred. A new body of rules, representing knowledge about how to deal with errors in the context of the user model, will be accessed to generate appropriate actions for the current user. Still other rules, invoked whenever the user accesses the document or otherwise indicates an increased knowledge of the program (say, by flawlessly executing a complex input sequence), will be responsible for updating the model itself.

Ideally, the model for a particular user should span several sessions with the program, and rules should be included which account for the normal attrition of user knowledge over a period of weeks or months. This implies that some profile be stored in the computer system on a long-term basis for each CONGEN user. We will design such a system with care, storing profiles only with the express consent of each user, and will provide alternative methods of defining an initial user profile (e.g., a short question-and-answer period at the start of a session) for those who do not wish to have stored profiles.

4) Error correction

So far we have discussed CGHELP as a system primarily for presenting the user with documented information, allowing him to learn the "rules" of CONGEN as easily as possible. There are of course other functions for a help system and at this point we will begin to explore more general CGHELP tasks.

A frustrating aspect of many interactive programs including CONGEN is that when an error occurs, it is frequently necessary for the user to "back up" and restart the program at some earlier point, even when the error is a relatively simple one which could be corrected locally, at the point of detection. For many user errors in CONGEN it is possible to define one or more probable fixes to the problem. These corrections may be either automatic modifications of internal CONGEN variables or minor digressions from the normal input sequence to allow the user to correct the error or omission himself. The next step in CGHELP development will be to incorporate error correction information into the "actions" of the error detection rules and to establish methods of using this knowledge to help the user recover gracefully from error conditions.

Automatic error correction must be approached with care because it will require CGHELP to take an active role in modifying the user's inputs to CONGEN. This can cause serious difficulties when the presumed correction is not appropriate; blatant errors can be transformed into more subtle ones which are extremely hard to detect later. One of the primary design criteria in the CGHELP error correction system will be that no modification is ever carried out unless CGHELP both obtains an explicit OK from the user and determines, from the user model, that he has an understanding of the nature of the problem and its solution. A second problem is that the automatic correction could take substantially longer than the user himself would need to correct the same problem. In CGHELP we will include some estimation of the lengthiness of possible fixes which, together with a measurement of system load, will influence the selection of the appropriate corrective action.

The natural result of including and maintaining a sophisticated error correction facility in CGHELP will be an increased flexibility in the input language. The user will be

allowed to deviate from the normal input protocols and the burden of verifying the overall correctness of the input will fall upon the program. The messages from the program can be phrased in such a way that the user seldom needs to know that technically he has made "errors" - he will use the commands in an order which seems logical to him and CGHELP will establish the dialog necessary to educate him and to query him as detailed information becomes important.

5) Extensions of error correction to "soft errors"

When we interact with new researchers who are learning to work with CONGEN, we find ourselves explaining not only the "rules" of the program but also many other topics such as strategy, helpful hints, details of the algorithms, etc. The clues that a user needs such higher-level help usually come directly from his inputs to the program, augmented by our mental model of the expectations which users bring to the program. The last phase of CGHELP development will be an open-ended exploration into the automation of help on what we term "soft-errors", or errors which are correct statements but show poor strategy, poor use of commands, and so forth.

We plan several new tools for the error detection system which will allow it to perceive these "soft errors". First, we will develop methods of estimating the computer time and storage space required in specific cases by each of the major functions of CONGEN. Currently there are no guidelines to help a user determine whether a given phrasing of a problem is possible to solve with CONGEN, and cases which are too large cause the program to carry out extensive computations before it becomes obvious to the user that the task is impossible. Second, we will incorporate a scanning facility which can examine intermediate results of a computation as they are being produced, looking for unusual or characteristic chemical features which the chemist may not have realized were possible. The chemical knowledge base which will define the criteria of "unusualness" will be distilled from our experience with typical CONGEN cases, and the chemist will have access to these criteria so that he can change them to better suit his needs, if necessary. Finally, we will create a strategy section which, given a problem in a particular state, will rank, in terms of overall problem efficiency, the possible sequences of commands needed to complete the problem. The evaluation will draw upon the estimator described above and upon a set of heuristics concerning "good form" in approaching CONGEN problems. Such evaluations will give us not only a yardstick against which to measure the user's strategy, but also a possible "driver" for automatically carrying out whole problems.

These tools represent measures of "soft error" conditions which we now feel to be important, but it is likely that other tools will become evident as we gain experience with other users. Perception of "soft errors" will be implemented by adding

appropriate situation-action rules to the basic production system, and the on-line document and tutorial systems will be augmented with information about problem size, chemical unusualness (as defined in CGHELP) and strategy. The user model will gain new importance in this process because it will become an integral part of the decision as to whether or not a "soft error" has even occurred; these conditions are defined in terms of the user's expectations and desires. Also, in order to maintain a considerate and useful dialog between CGHELP and the user, we will explore the inclusion of some elements of user psychology into the model. Because the danger of frustrating or boring the user will be substantially increased when CGHELP takes a more active role in the session, a commensurately more accurate model of user irritation or satisfaction will be needed to guide the program.

6 **Appendix II**
 1975-76 Annual Report to NIH

Table of Contents

| Section | | Page |
|---------|--|------|
| | Subsection | |
| 1. | PART 1: ENHANCING THE POWER OF THE M.S. RESOURCE . . . | 1 |
| | 1.1 Introduction | 1 |
| | 1.2 Hardware Acquisition and Development | 3 |
| | 1.3 Software Development | 4 |
| | 1.4 Operating System | 6 |
| | 1.5 Combined Gas Chromatography/High Resolution Mass Spectrometry | 7 |
| | 1.6 High Resolution Spectra Utility Programs | 10 |
| | 1.7 Process Monitor: PMON | 11 |
| | 1.8 METASYS | 12 |
| | 1.9 Summary | 13 |
| 2. | PART 2: DEVELOPING PERFORMANCE AND THEORY FORMATION PROGRAMS TO ASSIST IN BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS | 13 |
| | 2.1 Introduction | 13 |
| | 2.2 CONGEN | 14 |
| | 2.3 PLANNER | 21 |
| | 2.4 Meta-dendral Rule Formation Programs | 21 |
| | 2.5 Results | 26 |
| | 2.6 Heuristic Programming Project Workshop | 27 |
| 3. | PART 3: APPLICATIONS TO BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS | 28 |

| | | |
|-----|--|----|
| 3.1 | Introduction | 28 |
| 3.2 | Applications by Professor Djerassi's Research Group | 29 |
| 3.3 | Utilization of the Mass Spectrometry Resource | 36 |
| 3.4 | Applications of Programs by External Scientists | 38 |
| 3.5 | Export of GC/MS Programs to Other Sites. | 41 |
| | Index | 60 |

II.A. DESCRIPTION OF PROGRESS

OVERVIEW

In the period August, 1975 to July, 1976 the DENDRAL programs and the gas chromatography/mass spectrometry (GC/MS) data system have made significant progress toward the goals stated in the research proposal. This report of progress is organized in three parts, corresponding to the three specific aims of our December, 1973, proposal: (PART 1) Enhancing the power of the mass spectrometry resource, (PART 2) Developing performance and theory formation programs, and (PART 3) Applying the computer programs and instrumentation to biomedically relevant structure elucidation problems.

The DENDRAL project, one of the major users of the SUMEX-AIM computer facility, has been forming its own community of remote users. This "exodendral" community has already provided valuable contributions to program development and both the community and contributions are expected to grow at an increased rate. Our programs are receiving heavy use from local users and outside users who are investigating structure elucidation problems for a variety of different compound classes. Local users include members of Professor Djerassi's group, other chemistry department persons and research groups at the Stanford Medical School. We have continued building a community of outside users who can access our programs at SUMEX through the TYMNET or ARPANET.

1 PART 1: ENHANCING THE POWER OF THE M.S. RESOURCE

1.1 Introduction

Our grant proposal requested funds for significant upgrading of our capabilities in mass spectrometry. The goals of this upgrading were to provide routine high resolution mass spectrometry (HRMS), combined gas chromatography/low resolution mass spectrometry (GC/LRMS) and to develop a combined gas chromatography/high resolution mass spectrometry (GC/HRMS) facility. In addition, this would provide the capability for new experiments in the detection and utilization of data on metastable ions. These capabilities would then be available as required for application to our wider goal, solution of biomedical structure elucidation problems of a community of researchers.

The upgrading included several items of hardware and software development, as follows: 1) Acquire stand-alone computer support for the mass spectrometer because existing facilities were inadequate and very expensive; 2) convert existing software, written in the PL/ACME language into FORTRAN so that it would run on the new system; 3) develop new software as required for the demanding task of GC/HRMS; 4) provide hardware and software for semi-automatic acquisition of data on metastable ions. The initial development phase of this upgrading included performance tests to determine the capabilities and limitations of the GC/HRMS system to define the scope of problems to which it can be applied. The past year's efforts (year two of the grant) have culminated in accomplishment of many of the above goals for development. In the first year, the computer system (a Digital Equipment Corp. PDP 11/45) was purchased, installed and is now operating routinely in conjunction with the mass spectrometer (a Varian-MAT 711) and an auxiliary PDP 11/20 system (see system configuration, Fig. 1). Program conversion and modification for the initial version of the software system was completed and the computer system now provides complete stand-alone support for our experiments in mass spectrometry. Over the past year we have developed further our philosophy of data acquisition and reduction based on computed models of the actual performance of the mass spectrometer. This was and is necessary for routine automated collection and reduction of combined GC/HRMS data with minimal operator intervention in the procedures.

The system development is motivated by two goals. First, the system must be robust in the sense that it continue to operate under a variety of changing conditions, including intermittent misbehavior of the mass spectrometer. This ensures that the system can recover from hardware or software error conditions to prevent fatal "crashes" of the system and resulting loss of data. Second, the system must automate the GC/HRMS task. The volume of data acquired in GC/HRMS experiments can be efficiently handled only when every spectrum can be acquired and reduced for final output by the system without manual intervention. We are successful in these goals because we have written the software to determine the actual performance of the mass spectrometer and to have subsequent calculations based on that measured performance, as opposed to some hypothetical ideal.

We are now providing routine GC/HRMS service on a limited basis as we improve the system. The time required for system development and testing will slowly diminish over the next year, leaving additional time for analysis of mixtures obtained in our own work and that of our collaborators. We have deferred implementation of the metastable system (see below) while the GC/HRMS development is continuing, although we have completed the hardware and much of the software for the system.

1.2 Hardware Acquisition and Development

We have, in the mass spectrometry laboratory, two high resolution mass spectrometers, the Varian-MAT 711, and the AEI MS-9. Development efforts have focussed upon the MAT-711 because this more modern instrument is equipped with the high performance gas chromatograph needed for the GC/MS efforts.

We concurred with the study section's recommendation that stand-alone computer support be provided for efficiency and long-term cost effectiveness, and that such support be provided by the existing PDP 11/20 and a new PDP 11/45 or equivalent. We were able to adjust our first year budget to allow purchase of this computer.

At the time that the processor was ordered the cost of DEC disk drives was nearly double that of other vendor's drives. Accordingly we originally procured dual density top loading drives from System Industries. These drives were not directly software compatible with DEC RK type drives, but System Industries promised a hardware development to develop such compatibility and furnished software patches for DOS 8 so that we could use the drives. Unfortunately the hardware development was not carried out and our software needs expanded beyond DOS 8. We dealt with this problem by returning the System Industries' drive in the spring of 1975 and obtaining an equivalent drive for less money from International Memory Systems (IMS) which was RK compatible. The IMS disk drives have been installed for over a year with no indication of incompatibility with the DEC drives. The current hardware configuration is shown in Figure 1.

The PDP 11/20 processor is directly connected to the mass spectrometers and the gas chromatograph through two interfaces. The Ion Multiplier inter-face is a DR 11-B which provides direct memory access transfer of digitized ion multiplier samples. The direct memory access is necessary to provide a channel of sufficient bandwidth to achieve the requisite sampling rate for GC/HRMS work. The General Interface is our own design for a multiplexed interface used to select between the spectrometers; to manipulate the hardware mass scanner; to control the source voltage, the magnet current, or the analyser voltage; and to read the magnetic field, the source voltage, or the total ion current. All of these functions are slow speed and hence do not require the high data rate of a DMA interface. The general interface has 5 unused channels which are available for future development.

In addition to the instrument interfaces the PDP 11/20 is equipped with 8k of core memory, a KSR-33 terminal, a KW 11/P programmable clock and is tied to the PDP 11/45 via the Inter-Processor Interface (IPI). The IPI is a full duplex single word channel for which we have written a software driver providing user programs with 16 priority driven block transfer unidirectional channels. Thus, though the hardware provides only a single real channel it has proven easy to build mechanisms to

provide a very flexible and convenient mode of communication between the two processors.

The PDP 11/45 is equipped with 28k core, a PC 11 high speed paper tape reader/punch, an LA 30 terminal, a TM 11 industry compatible magnetic tape drive, an LP 11 300 line per minute printer, a KW 11/L line clock, a Loma Linda crt display, a CalComp drum plotter, and a hard line to the PDP 10. The dual Loma Linda / CalComp facility provides for both high speed real-time displays as well as for low speed off-line hardcopy graphics. The TM 11 provides a communications media to other processors and is used by procedures to save data on system failures and to maintain the archival data base.

1.3 Software Development

Conversion of existing PL/ACME programs to FORTRAN was begun on the award of the grant. Conversion of these algorithms also included many system software developments to ensure that previously batch processing programs could function in a real-time environment under the requirements of GC/HRMS operation. This development included not only improvements and extensions to existing algorithms, but building a file management system for facile logging and storage of spectra with the ability for simple recall to examine or recompute old data, and a diverse package of debugging, display and plotting and mass spectrometer evaluation programs. Development of improved capabilities for these tasks is an on-going project.

Because we view GC/HRMS as the most important new capability of our mass spectrometer/computer work, the requirements of GC/HRMS have guided development of the software system. These requirements include continuous automatic monitoring of instrument performance to avoid wasting time collecting poor or erroneous data. Because we have chosen to approach GC/HRMS with an electrical recording system, as opposed to photographic, we are able to monitor the instrument continuously, both during initial setup and during the course of the GC/HRMS experiment. Major sections of the software and how they interact among one another are summarized below.

During the past year the routine production usage of the HRMS data has become a reality. The direct utilization of the system for the acquisition of high resolution mass spectrometry data typically consumes 6 hours per day. This figure does not include time for the post-processing of data, retrieval of data from the archival data base, or for the generation of duplicate print outs of selected data. These demands add 1 to 2 hours of system service each day to the total high resolution system requirements.

Low resolution mass spectral data whether it be derived

from high resolution data or obtained directly as low resolution data, places additional time demands upon the data system. High to low resolution conversion, low resolution plotting, and low resolution spectral library searching have all generated a need for increasing amounts of system time.

In an effort to utilize the data system more completely during non-prime time, batch and spooling mechanisms have been constructed. The high resolution spectral reviewing mechanism may be actuated and then left unattended while the hard-copies are being generated. The high to low resolution conversion process contains a mechanism for the generation of a low resolution plotting spool which can be played without operator intervention. Batch procedures have been written which provide for the archival of newly acquired spectral data in the archival data base.

As with any system the size of the high resolution system there is a continual need for system maintenance and minor software upgrades. As a wider range of data acquisition and analysis becomes available new demands upon the system have developed which require modification of the software.

The net result of the production demands has been to reduce the amount of system time available for the development of new software facilities. Software development and production compete for the available system time reducing the productivity of both the chemical user and the software developer. This competition can be drastically reduced if software development can proceed on a machine separate from that on which production is done. The SUMEX PDP-10 offers an exceptionally attractive environment for software development. The TENEX operating system provides a more tractable medium for development than does the restricted environment provided by PDP-11 operating systems.

A major factor in the ease with which programs can be constructed is the ease with which text can be manipulated. The TV-EDIT program which is available on the PDP-10 has proven to be effective for this task. This program provides an extremely flexible text editing system for display terminals. The mechanics of program construction can be greatly simplified by the utilization of this facility. Typically all major (more than a few changes) text modification of programs are carried out on the PDP-10 using TV-EDIT and then transferred to the PDP-11. Thus even the task of writing FORTRAN programs is simplified even though there exist FORTRAN incompatibilities between the two machines.

While TV-EDIT has reduced development demands on the PDP-11 by eliminating PDP-11 text editing sessions, the problem of program compilation and debugging remain. Clark Wilcox, of the SUMEX staff, has provided an effective solution to this problem with the development of the MAINSAIL (machine independent SAIL) compiler. This compiler provides the user with a powerful,

machine independent, structured language. Not only is the compiler machine independent, but exhibits superior execution speeds and storage requirements as compared to the DOS 9 FORTRAN which has been used previously.

The combination of TV-EDIT and MAINSAIL has proven to be an effective method for the development of software for the PDP-11s within the PDP-10 environment. Most debugging can be carried out on the PDP-10 and then transferred to the PDP-11s for final debugging of machine-dependent facilities. The class of machine-dependent facilities includes device drivers and interaction with the operating system. The class of machine-independent facilities includes analysis algorithms, file manipulation, and most other programs which need development. This means that the amount of time required on the PDP-11 for program development can be reduced significantly using the aforementioned process, leaving more time for production demands.

1.4 Operating System

DOS version 8 was the first operating system to be used. However, this system was abandoned in favor of DOS 9. The major mandate for this conversion is the vastly improved overlay system offered by DOS 9. Overlaid files are maintained as a single, contiguous file on disk as opposed to the DOS 8 method of maintaining a separate linked file for each overlay. The DOS 8 strategy demands that a linked file be opened, read, and closed for each overlay load. DOS 9 allows an overlay to be loaded with a single disk read. Also the DOS 9 overlay facility provides for a tree structuring process which was completely absent from DOS 8. Considering that the version of the system in use at the time of the conversion had 17 overlays, the importance of efficient overlay loading is obvious. In addition to these factors, DOS 9 provides batch processing facilities which make it much easier to do system generation, archive data, etc.

We have been using DOS 9 for the past year. This operating system was chosen as the most suitable system available at the time we started its usage. Unfortunately DOS has many shortcomings. Bugs in many of the system programs and poor recovery from hardware errors on mass storage devices are the most visible defects. More subtle defects exist however when complex real-time processing is desired. These defects are compounded by our lack of monitor or system program sources.

In response to these defects in DOS we initiated an investigation into alternative operating systems. Both RSX-11M and RT-11 were examined in light of our particular demands. RSX-11M was rejected due to its size, poor terminal handling, and its implied dependence upon memory management hardware. RT-11 version 2C has been shown to possess advantages over both DOS and RSX-11M. RT-11 is a small system which comes as either a single

job monitor (1.5k word resident monitor) or a foreground/background monitor (3.5k word resident monitor). This is much smaller than RSX-11M (6k word resident monitor) and somewhat smaller than DOS 9 (about 4k word resident monitor). The foreground/background facility provides a convenient environment for simultaneous processing of plot, print, or filing spools with system program or user program execution. The single job monitor provides a small high speed system suited to real-time instrument control and data acquisition.

Both the I/O facilities and file structure of RT-11 possess advantages over those provided by DOS. RT-11 provides a queue structure for all I/O, leading to a more flexible utilization of peripherals. Additionally a completion routine facility is available which allow user supplied routines to be invoked upon I/O completion, providing interrupt service outside of the device drivers. Adding, deleting, or modifying a device driver is also very easy, amounting to simply replacing a file on the system device. While the file structure is limited to contiguous files the access time to these files is much more rapid than that provided by DOS. The rapid file access is quite evident when running system programs. Assembly, linking, and file transfer operations are significantly faster operations under RT 11 than under DOS 9. This is an important consideration in light of the fact that it takes over 35 minutes to link the GC/HRMS system under DOS 9 and such slow response seriously degrades programmer efficiency.

RT-11 is additionally attractive in light of the development of MAINSAIL. The runtime system for MAINSAIL under RT-11 already exists while none is available for DOS. The RT-11 magnetic tape formats are directly readable and writeable by the PDP-10, eliminating the conversion necessary for DOS magnetic tape files. The RT-11 system will also provide a much cleaner interface for the hardline to the PDP-10. It will be possible to log onto the PDP-10 through the PDP-11 RT-11 system and transfer files directly between the systems via the hardline.

1.5 Combined Gas Chromatography/High Resolution Mass Spectrometry

The gas chromatography/high resolution mass spectrometry (GC/HRMS) system provides for the acquisition, analysis, and archival storage of high resolution mass spectral data of gas chromatographic effluents. The system is composed of a real-time instrument control and data acquisition system, a post-processing system, an archival data base, and various development facilities.

SAQMON is an assembly language real-time instrument control and data acquisition monitor which executes within the PDP 11/20. SAQMON is responsible for controlling and monitoring all

instrument hardware to provide for the acquisition of high resolution data from the mass spectrometer. It contains processes to start and stop mass scanning in both a cyclic and single scan fashion. DC signal level is determined here and peak thresholding and background removal are also done here. A major portion of the memory allocated to SAQMON is dedicated to buffering of the peak profile data, relieving the PDP 11/45 processor of this burden.

SAQMON communicates with the PDP 11/45 through the Interprocessor Interface using the IPIDVR program which provides 16 unidirectional priority driven channels between the processors using the IPI. Such a scheme allows for independent communication between the systems depending on the task being performed and the data being acquired.

REFRUN is a FORTRAN overlaid program which is responsible for the acquisition, filing and post-processing of high resolution calibration spectra. Prior to analyzing a sample of interest the instrument must be calibrated by generating spectra of a reference gas (currently perfluorokerosene) which can be later used to compute the masses of ions acquired in spectra of unknowns. REFRUN uses SAQMON to acquire peak profile (PPF) data from the instrument or the IOLNK program to acquire PPF data from a back up file. PPF data is converted to mass/amplitude pairs and various characteristics of the spectrum are computed. These results are summarized in a CRT display for use by the operator. This summary includes the calibration range, the voltage of the reference base peak, and plots of a model peak, the projection error versus mass and the resolution versus mass. From this summary the operator can gauge the performance of the total system. The model peak plot provides critical information on the instrument set-up so that the operator can optimize the instrument performance. Once the operator can repetitively calibrate using the reference gas a spectrum is filed. Both the PPF and the reduced data are filed so that all system functions can be performed again at a later time. When the data is filed automatic displays are generated of the scan summary and mass/amplitude pairs. REFRUN also provides a reviewing capability so that reduced data files can be used to generate additional copies of the displays.

SAMRUN is an overlaid FORTRAN program which executes within the PDP 11/45 to acquire, analyze, and post-process spectra of samples. SAMRUN uses SAQMON to acquire PPF data from the instrument or the IOLNK to acquire PPF data from a backup file. Spectral analysis of samples requires a reference spectrum previously filed by the REFRUN program. The reference spectrum is used to guide the detection of reference peaks within the spectrum of the sample. The reference peaks which are found in this fashion provide discrete samples giving the time-mass conversion information. Mass values are computed by interpolation between the reference peaks. The spectrum summary presented to the operator for each sample spectrum is similar to

that provided by REFRUN minus the graphics plus information on the amplitude of the sample's base peak. When a set of sample spectra are filed both the PPF and the reduced data are filed, providing the same rerun facilities as REFRUN. The automatically generated displays include the spectra summaries, the mass/amplitude listings and the composition listings. SAMRUN also provides a reviewing capability for generating new copies of displays or new composition listings with different parameters.

The minute quantities of certain samples which have been submitted for analysis prohibit the re-running of any experiments associated with these samples. The system operates in a somewhat hostile environment. The physical laboratory environment dictates that the computer system be located in close proximity to the GC/MS instrument. The instrument can cause severe electromagnetic disturbances (sparks within the source, high voltage shut down, etc.) which can bring down either the entire data system or portions of the system. Static electric discharges from the operator through the system console have also resulted in catastrophic consequences for the data system. These occurrences are quite unpredictable from the software point of view and are difficult to alleviate in the physical environment. Therefore, the software must file data as soon as it is acquired in order that in the event of system failure any data gathered up to that point is maintained intact. A restart facility is also provided so that an experiment can be continued after catastrophic failure, losing only the data associated with the particular mass scan in progress at the time of the failure.

Both raw and reduced data are logged in real-time into a standard system file. The operator has the option of permanently filing this data in a file with an automatically generated name or to ignore the experiment altogether and file none of the data. Filing of both the raw and reduced data is necessary so that later rerunning of the experiment can be carried out. This is desirable in case of difficult data or in cases of software malfunction.

Buffering is a central issue in the system. Due to the uneven distribution of data, high data rates, and slack periods, it is desirable to provide a large amount of buffering between the instrument itself and the reduction processes. It is the case that data from one spectrum can be reduced while another spectrum is being acquired. Currently the PDP 11/20 has sufficient buffer capacity to hold almost a complete spectrum. The PDP 11/45 can concentrate on the conversion of time/intensity information into mass/amplitude information and the generation of displays with little regard to buffering the raw data.

Feedback provided by the real-time displays can be used by the operator to determine the quality of the spectral data. One can disregard scans which are poor and know when one is of high quality. The operator can choose to print out results immediately for critical samples, or defer final output until

later while additional data are being collected. An archival system provides the facility for storing and retrieving old spectral data for review or reanalysis.

High resolution mass spectral data often contain peak complexes consisting of more than one peak not separated by the simple thresholding technique. This problem is aggravated in GC/HRMS experiments because scans are acquired at lower resolving powers to achieve increased sensitivity. In GC operation, a further source of overlapping peak complexes is bleed from the organic phase of the GC column; many components of column bleed have masses similar to those of perfluorokerosene, the reference material. In particular if a bleed peak is so close to a reference gas peak used for calibration that a complex arises the entire calibration mechanism can go awry. In response to this problem we have developed a technique for the analytic resolution of such complexes. The problem has two aspects. First, a reliable detection method for complexes must be available. The computer must be able to tell the difference between single peaks and complexes of peaks. Second, once the computer detects a complex it must be able to provide an estimate of the position and area of the component peaks. After careful examination of the data it was determined that a reliable detection technique could be based upon the second moment of peaks suspected of being complexes. The basic idea is to determine the statistics of a peak which is representative of a single peak in the (mass) region of the suspected complex. A decision can then be based upon a comparison of the observed 2nd moment and the 2nd moment of the representative peak. It should be noted that the representative peak is dynamic within each scan due to instrumental variations in the resolution vs. mass curve. Once a complex is detected it is subjected to an analytic resolution technique developed by our personnel which computes the position and area of two peaks assumed to produce the complex. This technique works on the previously calculated statistics of the representative peak and the actual statistics of the observed complex. This method of resolving peak complexes has made possible the full reduction of GC/HRMS data which is not reducible otherwise. The operator has the option of either normal data reduction or data reduction with the resolution technique.

1.6 High Resolution Spectra Utility Programs

The development of the GC/HRMS system has generated some additional programs for the examination of peak profile data. These programs are not intended for usage by the chemists but rather serve as tools for the software personnel developing new facilities and analyzing failures of existing facilities. PPFSEE is a FORTRAN program which allows the user to plot on a CRT or CALCOMP the profiles of selected peaks from a spectrum. The relevant statistics of the peak area, amplitude, width, 1st, 2nd,

and 3rd moments are also displayed. This program is useful for examining peaks for the occurrence of doublets and comparing peak shapes obtained under differing instrument conditions. PKEXAM is a FORTRAN program which provides the user with various spectral plots. 2nd moment vs time is a typical output which is used to evaluate the performance of the peak complex resolution mechanism.

Often the investigator submitting a sample for GC/HRMS GCMS can obtain useful information from a low resolution plot of the high resolution data. HRTOLR is a FORTRAN program which converts reduced high resolution data to a standard low resolution format. This conversion can be carried out in two modes:

- 1) All peaks which are present in the sample spectra but not in the reference spectra are represented in the low resolution output.

- 2) Only peaks whose masses match a user supplied composition are represented in the low resolution output.

Available in both of these modes are facilities for PFK removal, selected mass removal, scan selection, compound renaming, and spooling for later low resolution plotting.

We currently support two types of low resolution data post-processing. First, the program LRLOT produces plots of low resolution spectra. It is capable of generating plots of individual spectra of a selected file or plotting of all files contained in a spool file which can be produced either by the HRTOLR program or with a text editor. Secondly, the program SEARCH (developed by ourselves and our collaborators in the department of Genetics) can be used to search a library of low resolution spectra for matches to a user supplied spectrum. Thus data acquired from either high or low resolution operation can be plotted and library searched.

1.7 Process Monitor: PMON

Experience has shown that it is very difficult to obtain full processor utilization with a traditional subroutine structure. Such a structure lends itself to predefined static conditions rather than to the dynamic situation presented by a real-time instrument. The operator requires automated functions to be available in unpredictable ways due to the experimental nature of the work being done. What is required is a method for scheduling program execution on a priority demand basis.

PMON is a monitor for scheduling real-time processes in a user definable fashion. Each system which uses PMON simply links PMON as the main segment of the program. The user supplies a process structure, PSTRUC, which describes to PMON all processes

in the system and their relative priorities . The PSTRUC contains a sequential list of priority levels running from the highest priority through the lowest priority. Each priority level is composed of a ring of process descriptors which specify processes at the same level. All processes represented on a given level are guaranteed to receive equal processor attention. Each process has associated with it a list of blocking conditions. These conditions are simply booleans relating to the state (empty/non-empty) of a queue. PMON schedules the highest priority process that has a satisfied blocking condition. Processes communicate with each other by pushing information into queues and by waiting for queues to attain a desired state. The queues are manipulated in a mutually exclusive fashion so that completion routines can send information to processes about I/O completion at an interrupt level. The current implementation of PMON requires less than 512 words of memory, despite its implementation in a high level machine independent language.

1.8 METASYS

METASYS is a data acquisition and analysis system for data on metastable ions. It is constructed around the PMON real-time monitor. It is composed of two autonomous subsystems:

- 1) A High Resolution Metastable Virtual Instrument (MVI)
- 2) A Data Manipulation System (DMS)

The MVI provides the user with the following capabilities:

- 1) Setting of any automated control.
- 2) Reading any automated indicator.
- 3) Scan source voltage, analyser voltage, and magnet current in user selectable fashions.
- 4) Acquire digitized samples of ion multiplier current, magnetic field, source voltage, total ion current, or functions of these samples in a user selectable fashion.
- 5) The generation of the Context Base which is a medium term memory for system events. All operator interaction and all data acquired from the instrument are recorded here.

The DMS provides the user with the following capabilities:

- 1) Permanent filing of data contained in the Context Base into the METASYS Data Base. (MDB).
- 2) Reexamination of data contained within the MDB.

- 3) Generation of both soft and hard copy displays of data contained in the Context Base and the MDB.
- 4) User selectable analysis of any data available to the DMS.

The implementation of METASYS is not yet complete. The PMON and the basic MVI processes are functional, but the DMS development has not yet been completed. This development is proceeding rapidly, however, which can be attributed to the advantages of program development on the PDP 10.

1.9 Summary

As the above hardware and software improvements are being made we will continue evaluation of the GC/HRMS system in parallel with its actual application to real problems. GC/HRMS is a relatively new and difficult technique for routine application. In order to use it effectively, we will have to exert some effort toward determining and optimizing the performance of the many elements of the system, the GC, the MS, and the computer hardware and software.

2 PART 2: DEVELOPING PERFORMANCE AND THEORY FORMATION PROGRAMS TO ASSIST IN BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS

2.1 Introduction

The Heuristic DENDRAL computer programs assist with structure elucidation problems by helping interpret mass spectra and helping generate structures that are consistent with data obtained from a variety of other spectroscopic and physical/chemical sources. The Meta-DENDRAL programs assist with rule formation problems in cases where the rules of mass spectrometry are not known.

Both the interpretation and rule formation programs are written as interactive tools to be controlled by professionals to combine the professional's judgment with the computer's combinatorial power.

2.2 CONGEN

The CONGEN [48,53] program represents a significant extension of a program which has developed over the last several years, the cyclic structure generator [40,41]. The purpose of CONGEN is to assist the chemist in determining the chemical structure of an unknown compound by 1) allowing him to specify certain types of structural information about the compound which he has determined from any source (e.g., spectroscopy, chemical degradation, method of isolation, etc.) and 2) generating an exhaustive and non-redundant list of structures that are consistent with the information. The generation is a stepwise process, and the program allows interaction at every stage; based upon partial results the chemist may be reminded of additional information which he can specify, thus limiting further the number of final structures.

CONGEN fits with the other DENDRAL programs as a "backstop" solution to structure elucidation problems. If the mass spectrum of an unknown compound is available, then CLEANUP and MOLION could be used, but if the general class of the compound is not known, PLANNER has no starting point from which to work. In such cases, structural information can be extracted manually from the spectrum and given to CONGEN for analysis. Because CONGEN makes no assumptions about the source of this information, other spectroscopic or chemical techniques may be used to supply supplemental data.

At the heart of CONGEN are two algorithms whose accuracy has been mathematically proven and whose computer implementation has been well tested. The structure generation algorithm [31,37,40,41] is designed to determine all topologically unique ways of assembling a given set of atoms, each with an associated valence, into molecular structures. The atoms may be chemical atoms with standard chemical valences, or they may be names representing molecular fragments ("superatoms") of any desired complexity, where the valence corresponds to the total number of bonding sites available within the superatom. Because the structure generation algorithm can produce only structures in which the superatoms appear as single atoms (we refer to these as intermediate structures), a second procedure, the imbedding algorithm [48,53] is needed to expand the superatoms to their full chemical identities.

These two routines give the chemist the ability to construct

structures from a given set of molecular "building blocks" which may be atoms or larger fragments. By itself, this capacity is of limited utility because the number of final structures can be overwhelming in many cases. Usually, the chemist has additional information (if only some general rules about chemical stability, which the program has no concept of) that can be used to limit the number of structural possibilities. For example, he

may know that because of a compound's stability, it cannot contain a peroxide linkage (O-O) and thus the programs need not consider such structures when there are two or more oxygens in the "building block" list.

In the past year CONGEN has reached the level of a practical production program which can aid chemists, both locally and at remote network sites, in solving the structures of drug-related compounds and natural products. The development of this program during the year has been strongly guided by the difficulties and new requirements which have appeared as it was applied to a wide variety of cases, and its efficiency and usefulness have increased dramatically. We report here the details of the modifications and additions we have made to CONGEN, and the effects they have had on its utility. Also, because of the rich repertoire of structure modification and testing functions available within CONGEN, we have found it to be an invaluable "laboratory" for the testing of new ideas, and we briefly describe two pilot projects which form the basis for future research. Discussion of applications of CONGEN to problems of biochemical interest is included in Part 3.

Program modifications

DEPTH-FIRST GENERATION. This modification has been both the most difficult and the most useful. The structure-generation algorithm which was originally part of CONGEN processed the "tree" of subgoals and subgoals-of-subgoals in a breadth first fashion. Although this was the most logically coherent and understandable encoding of the algorithm, it meant that a user would have to wait until the very end of a generation problem before he could see any of the results. This was particularly frustrating when a problem was submitted to CONGEN which was too big and/or time-consuming, because the user could never get any results at all. To alleviate this difficulty, we undertook a complete reorganization of the structure-generation algorithm so that it would proceed depth-first, giving results continuously as the computation progressed.

It is difficult to communicate the complexity of such a reprogramming without a major digression, but the flavor of the necessary changes is captured in the following example. At several points in the algorithm, there are what might be called "branching functions" whose purpose it is to solve some intermediate problem which has several alternate solutions. It is easiest to define such a function so that it computes the whole list of possibilities and returns the list to the caller. It is then the caller's responsibility to determine what is to be done with each possibility, and the branching function itself can be viewed as a separate module. This is a breadth first approach, and the difficulty is that the caller can make no progress until the branching function has constructed and returned all possibilities. The depth-first approach is to have the branching function itself be responsible for further

processing each time it creates a new result. To retain the modularity of the branching function, some mechanism is needed to allow the caller to "tell" it what this further processing consists of, and such a mechanism was instituted throughout the structure-generation algorithm.

We made use of the depth-first generation by instituting an interrupt mechanism in CONGEN whereby a user can examine the developing list of structures as they are created. This is a tremendous advantage both psychologically, because it gives the user a feeling that the program is "doing something", and operationally because it provides rapid feedback. A chemist can now often see quickly that a given case will create many more structures than expected, and the intermediate output can suggest forgotten constraints or superatoms. The following is an example of a terminal session in which the interrupt mechanism is used. The character control-S gives a "snapshot" of progress on the problem while control-I allows for the drawing of partial results. Both of these features are illustrated in the sample CONGEN session shown in Appendix A.

NEW CAPABILITIES FOR THE USER. There have been several additions to CONGEN which are visible to the user and which generally increase the flexibility and power of the program. These include

1) Making CONGEN aware of aromaticity, a chemical property of molecules which results from certain combinations of double bonds in rings. Aromaticity has a profound effect upon both the chemical reactivity and symmetry properties of molecules, and CONGEN can now be directed to detect aromaticity in its output structures, to compensate for the difference between the actual symmetry of an aromatic system and the symmetry which appears in the graph representing it, and to distinguish aromatic from non-aromatic atoms when it tests GOODLIST and BADLIST entries.

2) Giving the user the ability to type "?" to any prompt in the program, which results in a summary of the possible inputs. In some cases this summary is a list of possible commands, while in others it is a short explanatory message. A new interactive teletype-input routine was developed which makes it easy to include such help messages in the program, and which mimics the handy command-recognition and command-completion features of the TENEX operation system.

3) Including new specifications in the EDITSTRUC language for describing substructural features. The user can now declare a bond in a substructure to be an "anybond", which means that the atoms at the termini are connected but that the multiplicity of the connection is unspecified. This is especially handy when defining substructures containing aromatic portions because bond multiplicity is an indistinct concept in aromatic systems. Another new structural element which can be specified is a "linknode", a node which stands for a variable-length chain of

atoms of the given type rather than a single atom. The minimum and maximum lengths of such a chain can be specified as well. The linknode feature is useful for defining constraints on ring fusions and other constraints such as Bredt's rule which depend on path length. Other extensions have been made internal to CONGEN which will shortly be reflected in the user-level language of EDITSTRUC. These include numerical inequalities involving node properties (e.g., "the number of H's on atom 3 is greater than the number of H's on atom 5") or linknode lengths (e.g., "the sum of the lengths of linknodes 2 and 6 is greater than 5"), and greater control over the number of fittings found for a GOODLIST constraint (e.g., the ability to distinguish between "the number of N's in six-membered rings" and "the number of six-membered rings containing N").

4) Allowing greater flexibility in the selection of terminal type. This choice controls the output of structural drawings so they are best suited to the user's terminal. Several different types of character-oriented and graphics-display terminals are now supported.

5) Making CONGEN accessible from the GUEST login account at SUMEX. This involved preventing a GUEST user from reaching certain critical points in CONGEN which would allow greater system access than is normally authorized for guests. We can now offer trial access to CONGEN via the guest mechanism without worrying about SUMEX misuse.

6) Creating a BATCH command for CONGEN. This allows the user to submit time-consuming, compute-bound calculations to the batch-processing facility of SUMEX. The computation is then run automatically at off-hours when it will not overload the system resources. The user can now run CONGEN in its interactive mode to input all of his data and then submit the large tasks to BATCH for overnight processing.

7) Including a pruning function MSPRUNE which is used to test a list of candidate structures for consistency with a set of observed peaks from a mass spectrum. The candidates are typically generated by CONGEN using structural data from other sources. The user specifies the observed MS peaks (as elemental compositions or nominal masses or a combination of both) along with a set of constraints on the allowed cleavage processes. MSPRUNE retains only those candidates which can account for the observations via one of these allowed processes. The constraints speak of the number of bonds broken and the number of steps in a process, the proximity of pairs of cleaved bonds (i.e., whether or not two adjacent bonds can break in a given process), the multiplicity or aromaticity of each cleaved bond and the possible neutral transfers. MSPRUNE is the first CONGEN function which can aid directly in the interpretation of "raw" spectral data.

8) Internal CONGEN Developments. The basic algorithms used for structure generation in CONGEN are firmly rooted in

mathematical graph theory. During the past year, there has been significant refinement of several of these graph theoretical algorithms. The new algorithms have been coded in SAIL, an extended ALGOL type language; and a sophisticated executive has been developed to coordinate the various SAIL routines as well as to direct the communication and control between the SAIL component and LISP component of CONGEN.

The power and utility of CONGEN rests, to a great extent, on the fact that it can generate structures under user supplied constraints. The most powerful of the routines used in constrained structure generation is the fragment imbedder [37,48]. It is this routine which permits CONGEN to efficiently generate only those structures containing given polyatomic fragments (i.e., superatoms). The fragment imbedding program was completely rewritten so that it operates now in a "depth first" rather than "breadth first" style. This was done so that the user can request CONGEN to produce examples only of candidate structures in those cases where the total number of candidate structures is very large. This change also increases the efficiency of the fragment embedding process and has the advantage that if a CONGEN run must be interrupted, the user is left with at least some candidate structures rather than just intermediate results.

During the grant period, a very general substructure matching algorithm was developed and coded in SAIL. This algorithm accepts as input a structure and a "pattern" and returns the number of times the pattern distinctly occurs in the structure. Here a pattern is a partially specified substructure in which atom names, bond widths and hydrogen attachments all may assume a range of values. This routine is used by CONGEN for post checking of structures and classifying lists of structures.

An improved technique to determine the topological symmetry group of a structure was also developed and coded in SAIL. This routine is used in several parts of CONGEN, e.g., fragment imbedding. This new routine is, statistically, at least an order of magnitude faster than the old group finding routine.

The language LISP, although quite powerful, does not produce very efficient machine code. It was for this reason that several of the routines used by CONGEN were coded in SAIL. However, because of the widely variant data types, LISP and SAIL are not compatible languages. Hence, all of the SAIL programs reside in their own TENEX fork, and they communicate with the LISP fork via a shared memory page. The new CONGEN SAIL code executive program handles all interfork communication for the SAIL routines, and it allows one to make additions or modifications to the SAIL portion of CONGEN with relative ease. This ease of change is also aided by the fact that all the SAIL programs are written in highly modularized form.

Preliminary testing of the new CONGEN SAIL fork indicates

these modifications and additions will yield a significant increase in the overall efficiency of CONGEN, and hence will enable one to consider a broader range of chemical problems.

INTERNAL CONGEN IMPROVEMENTS - LISP. Because of the diverse assortment of chemical problems to which CONGEN has been applied, we have been able to exercise all parts of the program in a variety of contexts. As a result, we have been able to uncover a number of hidden inefficiencies in the LISP section of CONGEN, and although correcting these has not had a direct impact on the command structure of the program, we estimate that a decrease of over 50 in CPU time has been achieved for typical CONGEN cases. In some cases this decrease is as high as 90.

These improvements have been numerous, but one stands out as most significant. Several changes were made to the graph-matching routine which is responsible for testing the presence or absence of structural features in molecules or molecular fragments. The new routine uses list space (a key resource in the LISP programming system) much more parsimoniously, and it incorporates a new and very efficient representation of substructures which makes optimum use of the linked-list data representation in LISP. Also included were a number of heuristics which, although they do not alter the output of the graph matcher, do dramatically decrease the amount of time spent on typical tests. The highly efficient SAIL graph matcher, described above, will soon supplement the LISP version, though the latter will still be needed in some cases because of its greater flexibility.

Other inefficiencies were detected and fixed in the portion of CONGEN which builds tree-like molecules and molecular fragments, where it was discovered that a built-in assumption (that the most common monovalent atom would be hydrogen) was adversely affecting the running times of some CONGEN cases, and in the portion responsible for computing the symmetry groups of graphs.

PILOT PROJECTS. CONGEN provides an excellent environment for the testing of new ideas because it contains an extensive "library" of functions for the creation, manipulation and testing of topological representatives of molecular structure. Below we describe two pilot projects which were explored within this environment and which provide the basis for proposed future research topics.

We developed within CONGEN a program called XMECH [60] whose purpose it was to study the possible mechanisms of cyclizations and skeletal rearrangements of monoterpanes, terpanes and sesquiterpanes. The study of these compound classes is an important sub-field of natural-products chemistry, and simple carbonium-ion mechanisms, such as cyclizations to double bonds and 1,2-alkyl and/or 1,2-hydride shifts, are frequently invoked to rationalize interrelationships between various

skeletal types. Using XMECH we were able to explore various combinations of these basic mechanisms and to develop exhaustive lists of skeletal types, known and unknown, which should be accessible from known biogenetic precursors via this approach. Our results indicate that although such mechanistic rationalizations are widely used, the method is quite non-selective: If a sufficient number of mechanistic steps is included to account for even a modest fraction of known skeletons, a vastly larger number of skeletal types are obtained which have never been seen in nature. It seems clear that there are much subtler mechanistic considerations which account for the specificity of biogenetic pathways, and our work points out the danger of rationalizing that specificity with an overly simple model. XMECH has laid the groundwork for a much more general program, REACT, in which a user will be able to define chemical reactions and apply them to problems of mechanistic chemistry and structure elucidation.

A second pilot project is the program MDGGEN which embodies a new, general approach to the interpretation of a mass spectrum in terms of structural possibilities for an unknown. The method used in MDGGEN compliments the MSPRUNE function described above (section 7 of NEW CAPABILITIES FOR THE USER) because it uses MS data at the beginning of a problem rather than as a final filter on candidate structures. Whereas MSPRUNE is logically part of the TEST phase in the traditional DENDRAL scheme of PLAN-GENERATE-TEST, MDGGEN logically belongs in the PLAN phase. Conceptually, MDGGEN is related to the PLANNER program, except that MDGGEN analyzes MS data without relying upon class-specific fragmentation rules as does PLANNER. Using a very simple and general fragmentation theory, MDGGEN processes selected peaks from a mass spectrum and constructs possible ways of segmenting the overall composition of the molecule to account for those peaks. These segmented descriptions are graphs similar to topological chemical structures except that one node may stand not just for a single chemical atom, but a collection of atoms (a composition) representing a connected piece of the molecule. We call these mass-distribution graphs, or MDG's. The structure-generation facilities of CONGEN allow us to assemble the atoms within each node-composition in all unique ways, and to imbed these assemblies in all unique ways into the overall MDG structures. In this way, we arrive at chemical structures which account for the MS data according to the simple theory. MDGGEN is still in its infancy, with the practical limitations of computer time and storage requirements restricting it to small molecules (up to perhaps ten non-hydrogen atoms) and relatively few observed peaks (up to roughly seven or eight ion compositions). This early development, which could take place rapidly because of the existing facilities within CONGEN, has helped us to focus our attention on the critical advances which will be needed in creating a more flexible and generally useful program.

2.3 PLANNER

The DENDRAL PLANNER program [28,33] is designed to analyze the mass spectrum of a compound or of a mixture of related compounds. Because there is no *ab initio* way of relating a mass spectrum of a complex organic molecule to the structure of that molecule, PLANNER requires fragmentation rules for the class of compounds to which the unknown belongs. This is its major limitation.

Applications and limitations of PLANNER have been discussed extensively. [28,33] The program is very powerful in instances where mass spectrometry rules are strong (i.e., general, with few exceptions). In instances where rules are weak or nonexistent, additional work on known structures and spectra may yield useful rules to make PLANNER applicable (see INTSUM and RULEGEN, below). One unique feature of

PLANNER is its ability to analyze the spectra of mixtures in a systematic and thorough way. Thus, it can be applied to spectra obtained as mixtures when GC/MS data are unavailable or impossible to obtain.

The power of the PLANNER has been substantially increased by including the MOLION program (discussed below) as a subroutine for computing the list of plausible molecular ions. Since this subprogram does not depend on knowledge of the compound class, the PLANNER no longer needs to have class-specific rules for determining the mass and empirical formula of the unknown molecule.

The major use of the Planner in the past year has been as a means of testing new class-specific mass spectrometry rules proposed by the Meta-DENDRAL program described below. One measure of quality of a set of proposed rules is their ability to discriminate among isomers in the same class. For example, the monoketoandrostandane rules can be partly evaluated by their ability to assign the keto group to the correct substituent position, based on the mass spectrum of the compound. Since there are eleven possible positions, we are asking the rules to discriminate the correct structure from the other ten monoketoandrostandanes.

2.4 Meta-dendral Rule Formation Programs

When the mass spectrometry rules for a given class of compounds are not known, the INTSUM, RULEGEN and RULEMOD programs can help a chemist formulate those rules. Essentially, these programs categorize the plausible fragmentations for a class of compounds by looking at the mass spectra of several molecules in the class. All molecules are assumed to belong to one class whose skeletal structure must be specified. Also, the mass

spectra and the structures of all the molecules must be given to the program.

INTSUM collects evidence for all possible fragmentations (within user-specified constraints) and summarizes the results. For example, a user may be interested in all fragmentations involving one or two bonds, but not three; aromatic rings may be known to be unfragmented; and the user may be interested only in fragmentations resulting in an ion containing a heteroatom. Under these constraints, the program correlates all peaks in the mass spectra with all possible fragmentations. The summary of results shows the number of molecules in whose spectra there is evidence for each particular fragmentation, along with the total (and average) ion current associated with the fragmentation.

The INTSUM program [34] is in routine, production use to assist in interpretation of the mass spectra of new classes of molecules (see Part 3 for details).

The RULEGEN program attempts to explain the regularities found by INTSUM in terms of the underlying structural features around the bonds in question that seem to "drive" the fragmentations. For example,

INTSUM will notice significant fragmentation of the two different bonds alpha to the carbonyl group in aliphatic ketones. It is left to RULEGEN to discover that these are both instances of the same fundamental alpha-cleavage process that can be predicted any time a bond is alpha to a carbonyl group.

The RULEMOD program modifies and condenses the set of rules produced by INTSUM and RULEGEN together. It looks at the negative evidence associated with each candidate rule in order to select the best ones, then merges rules that seem to explain the same breaks (if possible). The program was substantially improved in several ways, as described in the next section.

2.4.1 Improvements Made to the Meta-DENDRAL Programs

2.4.1.1 INTSUM Improvements

Transfers of arbitrary neutral species can now be specified as part of the mass spectrometry processes, instead of transfers of hydrogen atoms alone. This capability increases the utility of the program in at least two ways: first, it allows a chemist to control the program better -- to produce the kinds of results that are more chemically meaningful -- and second, it allows the program to explore more complex processes within its space and time limitations. For example, carbon monoxide and water were listed as plausible neutral molecules to transfer in or out of fragments for the triketoandrostanes. Thus, the processes are listed with and without these transfers, just as chemists prefer,

instead of showing loss of CO as a set of two breaks around the keto group, or loss of H₂O as loss of oxygen (breaking the C=O bond) accompanied by loss of two hydrogens. What is more, the program can now produce these results without violating its chemical heuristics of (a) not breaking adjacent bonds, and (b) not breaking double bonds. This economy also pays off in increasing the complexity of the processes that can be considered. Because loss of CO, for example, is a result of a transfer instead of the result of breaking two bonds, the number of bonds broken in accompanying processes can be increased by two.

Another INTSUM improvement was to increase the options for initial data filtering. Thresholding is too simple for many problems, so we now provide an option to cluster peaks and select the n largest peaks from each cluster.

The format of the input data is also now less strict than before. We have written programs to read spectra in Aldermaston format. And we have merged CONGEN's Editstruc package into the INTSUM setup routines to allow a chemist to associate structures with spectra interactivity. This greatly decreases the chances of error in setting up the input data.

Several modifications were also made to the program to increase its efficiency, e.g., processing all intensities as integers (between 0 and 1000).

2.4.1.2 RULEGEN Improvements

The evaluation of prospective rules in RULEGEN guides the entire rule generation procedure. To tune this procedure, we modified the evaluation function in several ways and compared the resulting sets of rules. We were looking for an objective way of telling the program to keep rules general, but "not too general". The current evaluation function is substantially improved as a result.

Because the RULEGEN program searches such a large space of partial and complete rules, it requires large amounts of computer time (sometimes more than 60 cpu minutes). Thus, we have investigated several improvements for efficiency alone. In addition, we have made the program easier to set up and run in batch mode to reduce the chemist's personal time investment. And we have made the program easily restarted from any intermediate point -- to protect the chemist from machine failures.

2.4.1.3 RULEMOD Improvements

At the time of the last annual report RULEMOD was a new program still in its experimental stages. Since then we have added new subprograms and integrated the program with other programs to make it a useful and necessary part of Meta-DENDRAL.

Two new subprograms greatly improve RULEMOD's performance. (1) A program to add specifications to rules was completed. It looks for plausible ways of making a rule more specific in order to decrease the number of counterexamples to the rule. (2) A complementary program to make rules more general was also completed. The program tries to find ways to reduce the number of descriptors on nodes of subgraphs in order to increase the breadth of applicability of rules. Its major constraint is that it cannot make any change that would increase the number of counterexamples. Both of these subprograms make the final rules much closer to rules that chemists approve of.

The subprogram that merges rules was also improved. The program tries to merge pairs of rules into a more general form for economy and clarity of rules. Its major constraint is that no explanations are lost, i.e., all the data points explained by the initial pair of rules will still be explained after merging. Formerly we insisted that the more general form must cover all the same data points as the initial rules, but this was found to be too narrow a constraint. By giving the program a more global view of the entire set of rules, we can let the more general, merged form explain fewer data points than its component rules as long as other rules explain the remainder.

2.4.2 Search for New Applications of the Rule Formation Programs

In this year the Meta-DENDRAL programs have matured enough to let us consider extending them beyond mass spectrometry. The domain that we chose was ^{13}C NMR spectroscopy, for a variety of reasons.

^{13}C NMR has been characterized as the spectroscopic technique of the 1970's [68]. Our laboratories have been involved in experimental work on ^{13}C NMR spectra of amines, keto and hydroxy steroids [62-64]. In addition, we have carried out a preliminary investigation of a Heuristic DENDRAL approach to interpretation of ^{13}C spectra of amines [39].

There are several parallels between rule formation in mass spectrometry and ^{13}C NMR spectrometry. In both techniques the precise reasons for molecular fragmentation (in the former) or NMR absorption (in the latter) are poorly understood. In the absence of a detailed theory capable of accurate prediction of spectra, we seek empirical rules which can relate observed data to measurable structural parameters. Some of the structural parameters presumed relevant, e.g., atom type, bond multiplicities, are shared in both techniques. Some of the current Meta-DENDRAL structural manipulation functions can be used for either technique. An important difference is that the planning phase of Meta-DENDRAL (i.e., INTSUM) necessary in applications in mass spectrometry is not required for ^{13}C NMR because we will deal initially with spectra whose absorption

peaks (or "shifts" relative to any internal standard) are assigned to specific atoms in the known structures. Typically scientists have sought an explanation for the ^{13}C NMR shift of an atom in terms of the structural environment of the atom. Searching such structural environments is a problem which is amenable to solution by existing and proposed parts of the Meta-DENDRAL program.

As in applications to mass spectrometry [58] we will propose a set of factors which might affect ^{13}C NMR absorptions. With a description of these factors we will use the Meta-DENDRAL program to produce a set of rules which will reproduce and predict resonance shifts of individual ^{13}C atoms.

The current Meta-DENDRAL program represents a basic framework for studying ^{13}C NMR rule formation. We believe that the program will require little revision to accommodate the differences in data and rules. We have already considered some of the problems of changing the form of rules. The subgraphs in the situation parts of rules need to be generated "outward" from a specific ^{13}C atom instead of outward from a bond broken in the mass spectrometer. The action parts of rules need to take account of an explicit absorption range whereas for mass spectrometry the rules predict much more precise data points (mass positions). We have made a preliminary test of the program's extensibility in the context of alkanes.

For the alkane study we used only a topological model of molecular structure, not a geometric model. The rules that were formed from a test set predicted shifts for ^{13}C atoms in other alkanes (outside the test set) with accuracy within 1.5 ppm. The major modifications needed in the program to produce these preliminary results were the following:

(a) change RULEGEN to generate rules by expanding the subgraph environments outward from a central atom rather than from a central atom rather than from a central bond;

(b) change the form of rules to associate a range of shifts with each subgraph rather than a precise fragment mass;

(c) redefine RULEGEN's evaluation function for partial rules to take account of the desire to predict narrow ranges of shifts.

Other domains were considered, including finding rules to associate pharmacological activity with molecular structure and finding rules for other organic chemical analysis techniques. Of all that we considered, ^{13}C NMR appears to offer the most in terms of both feasibility and utility.

2.5 Results

2.5.1 Keto-androstanes

We have shown that the Meta-DENDRAL program is capable of rationalizing the mass spectral fragmentations of sets of molecules in terms of substructural features of the molecules. On known test cases, aliphatic amines and estrogenic steroids, the Meta-DENDRAL program rediscovered the well-characterized fragmentation processes reported in the literature. On the three classes of ketoandrostanes for which no general class rules have been reported, the mono-, di-, and triketoandrostanes, the program found general rules describing the mass spectrometric behavior of those classes. The general rules shown in Tables II, IV, and VI explain many of the significant ions for compounds in these classes while predicting few spurious ions. The program has discovered consistent fragmentation behavior in sets of molecules which have not appeared by manual examination to behave homogeneously in the mass spectrometer.

Programs with knowledge of the scientific domain can provide "smart" assistance to working scientists, as shown by the reasoned suggestions this program makes about extensions to mass spectrometry theory. We are aware that the program is not discovering a new framework for mass spectrometry theory; to the contrary, it comes close to capturing in a computer program all we could discern by observing human problem-solving behavior. It is intended to relieve chemists of the need to exercise their personal heuristics over and over again, and thus we believe it can aid chemists in suggesting more novel extensions to existing theory. It can be argued that the two-dimensional connectivity model of molecules used in this study is not the right model for mass spectrometry; that there are deeper rationalizations of a fragmentation process than subgraph environments. However, this model is commonly used by working chemists and once fragmentations based on this model are defined, chemists can readily provide the remaining "mechanistic" rationalizations or see that further experimental work with labeled compounds is necessary. (Other limitations of the method have been discussed at the end of the methods section.)

Recent statistical pattern recognition work addresses some of the points on rule formation and spectrum prediction raised in this paper. We have avoided blind statistical methods for three important reasons. 1) We wish to explore thousands of possible subgraphs with associated features, as we search for those which are in some way important. Current pattern recognition procedures are restricted to much smaller numbers of manually (or computer-assisted) selected features, adding additional bias to the procedure. 2) We want to know how certain rules were obtained by the program and why certain other rules were rejected or not detected. We can trace the reasoning steps of the Meta-DENDRAL program and determine chemically meaningful answers to

such questions in a way that is not possible with purely statistical programs. 3) We wish to constrain the rule formation activity in ways that are natural to a working chemist. For example, we may want the program to avoid fragmentations involving aromatic rings or two bonds to the same atom, or, as mentioned above, we may want to look at fragmentations accompanied by loss of CO or other neutral fragments.

Rules can be formulated to explain data in terms that are known to be meaningful to chemists; most importantly, the rule formation constraints are under the control of the chemist. Also we feel that this approach provides a high level of generality in describing fragmentation processes. Although the rules are developed in the context of a particular set of compounds, they are not tied to that set but can be applied in other contexts, or compared to rules developed from other sets of compounds in a search for common features of the rules. For these reasons, we believe that the Meta-DENDRAL program offers a powerful and useful complement to pattern recognition programs for finding relationships between structures and spectral data.

We are cautiously optimistic about the general applicability of this rule formatton method, although we have demonstrated its utility for only a small number of compound classes and only in the context of mass spectrometry.

2.6 Heuristic Programming Project Workshop

In the first week of January, 1976, about fifty representatives of local SUMEX-AIM projects convened at Stanford for four days to explore common interests. Six projects at various degrees of development were discussed during the conference. They included the DENDRAL and META-DENDRAL projects, the MYCIN project, the Automated-Mathematician project, the Xray-Crystallography project, and the MOLGEN project. Because of the interdisciplinary nature of each of these projects, the first day of the conference was reserved for tutorials and broad overviews. The domain-specific background information for each of the projects was presented and discussed so that more technical discussions could be given on the following days. In addition the scope and organization of each of the projects was presented focusing on the tasks that were being automated, how people perform these tasks, and why the automation was useful or interesting.

In the following days of the workshop, common themes in the management and design of large systems were explored. These included the modular representations of knowledge, gathering of large quantities of expert knowledge, and program interaction with experts in dealing with the knowledge base. Several of the projects were faced with the difficulties of representing diverse kinds of information and with utilizing information from diverse

sources in proceeding towards a computational goal. Parallel developments within several of the projects were explored, for example, in the representation of molecular structures and in the development of experimental plans in the MOLGEN and DENDRAL projects. The use of heuristic search in large, complex spaces was a basic theme to most of the projects. The use of modularized knowledge typically in the form of rules was explored for several of the projects with a view towards automatic acquisition, theory formation, and program explanation systems.

For each of the projects, one session was devoted to plans for future development. One of the interesting questions for these sessions was the effect of emerging technology on feasibility of new aspects of the projects. The potential uses of distributed computing and parallel processing in the various projects were explored, particularly in the context of the DENDRAL project.

Most of the participants felt that the conference gave them a better understanding of related projects. And because many members of the SUMEX-AIM staff actively participated, the workshop also provided all projects with information about system developments and plans. The discussions and sharing of ideas encouraged by this conference has continued through a series of weekly lunches open to this whole community.

3 PART 3: APPLICATIONS TO BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS

3.1 Introduction

In our grant proposal we discussed the application of the instrumentation and computer programs described above to the study of molecular structure problems in a variety of biomedical applications areas. This is our primary research area, and we discussed specific classes of problems and compounds for investigation. We also made it quite clear that our facilities would be made available to wider community of collaborators/users as our resources permitted. Both categories of application, i.e., within our own group, and with an outside group, are described in some detail below. Our last annual report described several steps taken to encourage a broad community of researchers to use our facilities. For example, we sent a questionnaire to members of the American Society for Mass Spectrometry, Committee

III on Computer Applications, and a follow-up letter to persons indicating a desire to know more about access to our programs. The same note has been sent to several other persons whom we know from personal contacts might be interested. Because of the nature of their investigations, many of these people receive NIH support. Several of our publications (e.g., [45-49,53-61]) mention the availability of our programs. In addition, through individual contacts and formal presentations at conferences we have been encouraging outside use of the programs.

The availability of SUMEX as a mechanism for resource sharing has made it possible for us to extend access to our programs to a number of people. Without SUMEX, this access would be impossible, and most of our programs (those which are not easily exportable) could be used only by ourselves.

3.2 Applications by Professor Djerassi's Research Group

Our existing grants, outlined below, mesh well with our instrumentation and program development under the present award. Under NIH Grant GM06840 we have been studying natural products from marine sources with major emphasis on terpenoids and sterols. For this work we have been dependent on the use of our 711 instrument for high resolution mass spectrometry which we require for the identification of all new compounds, many of which are present in only very small quantities. We were particularly anxious to have access to GC coupled with a high resolution mass spectrometer because we hope to be able to screen large numbers of marine animals for their sterol content using this technique. We are currently engaged in intensive efforts in analysis of mixtures of marine sterols involving our computer-based procedures. The program for the development of the computer operated and assisted system of marine sterol structure analysis has been planned to proceed in three stages:

- 1) Analysis of all literature published concerning marine sterols so that a complete listing of known sterol structures and organisms studied could be compiled.
- 2) Collection, evaluation, digitization and computer file construction for the mass spectra of all known marine sterols, followed by the institution of a computer operated file search sequence for direct analysis of marine sterol GC-MS data.
- 3) The application of the INTSUM, RULEGEN, and RULEMOD programs to the computer file of marine sterol spectra so that a series of fragmentation rules can be extracted for use in the generation of possible structures from mass spectral data for new marine sterols, that is, sterols whose mass spectra cannot be matched with any spectra contained in the computer search file.

We are presently completing the second stage and beginning the third. The following discussion will be a summary of the work that has been completed, and the work that is in progress or planned.

The literature concerning marine sterols is extremely extensive. Over a thousand reports concerning marine sterols can be found scattered throughout a multitude of journals dating back to the initial report by Henze in 1908. In spite of the occurrence of a number of good review works in the literature, we have found the compilation of all reported marine sterol structures and organisms studied to have been an imposing task, which we have now completed successfully. The search has also pointed up a number of entire phyla of marine invertebrates for which no sterol analysis have been reported, and has therefore pointed out perhaps the best candidates to which the developing automated analytical procedures should be applied. The search has also generated an extensive and very refined list of descriptions which are now used in a computer generated update of our bibliography every two weeks for this very active field. This laboratory has been involved in sterol work for some years and so our own samples and mass spectral files have made a significant contribution to the compilation of the complete mass spectral file of marine sterols.

Table I represents a listing of marine sterol spectra as well as a listing of purely synthetic sterol mass spectra (for use in evaluation of the INTSUM results) which have been contributed by this laboratory. These spectra are now part of completely functional computer files. We have requested and received samples of other marine sterols from researchers around the world who have reported their isolation. A large number of these sterols have now had mass spectra taken and the enlargement of our computer mass spectral file is proceeding rapidly.

The series of programs for processing raw GC-MS data and searching mass spectral files have recently been instituted on the chemistry PDP 11/45 computer. The series of programs which have potential application to processing our data are CLEANUP (a program for subtracting GC column bleed or background and noise from raw GC-MS data, and resolving spectra of overlapping elutants), MOLION (a program for generation of molecular ion candidates from mass spectral secondary losses), and SEARCH (a program for searching and comparing experimental mass spectra to the file of known marine sterol mass spectra). Several data management programs exist for displaying the results of the file search and other operations. Development of a program to utilize GC retention indices is progressing. The first experimental file search for an actual sample run will be possible within the next few weeks, but we have already used the SEARCH program to process and evaluate several duplicate marine sterol mass spectra from our files as listed in table I. Table II represents the results of this kind of experiment. Three separate (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL (trivial name "FUCOSTEROL") mass spectra

were compared to 25 marine sterol mass spectra in the computer files via the SEARCH program. The program was able to select each of the mass spectra from the main file with the inclusion of one thirty carbon sterol (24Z)-24-PROPYLIDENECHOLEST-5-2N-3BETA-OL which possesses a structure similar to FUCOSTEROL, the twenty-nine carbon sterol. This kind of study has shown that in principle the SEARCH program functions for marine sterol correlations, but requires some fine tuning to reduce this kind of error. The search strategy modifications should be complete within the next several weeks.

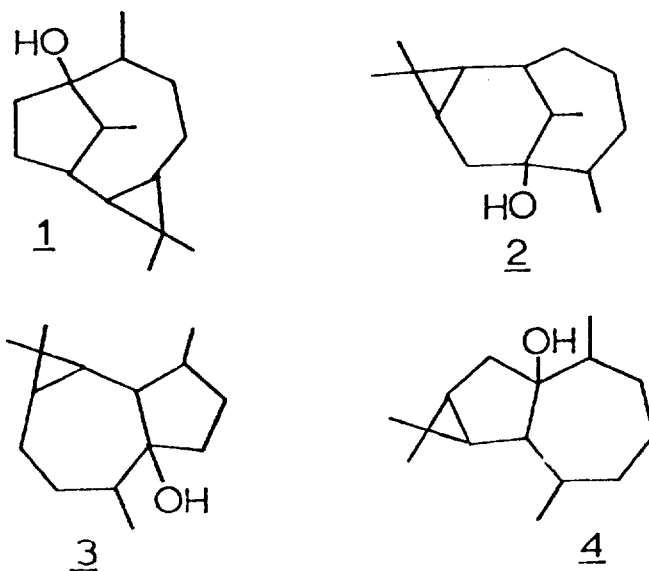
One other aspect of this work should be mentioned. We have found that for very complex marine sterol mixtures a single GC-MS run is sufficient to identify the major sterol components and a few minor components. Further separation procedures are required to analyze the remaining minor components. We have found many of the minor components to be of significant biosynthetic and ecological interest. We have spent a considerable effort perfecting rapid separations or enrichments of these minor sterol components so that GC-MS analysis can be run on them. We now have a procedure utilizing silica gel, alumina, silver nitrate impregnated alumina and silica gel, and high pressure reversed phase liquid chromatography which produces separations and/or enrichments so that GC-MS data can be obtained for every sterol of even a 30 component mixture. Perfecting these separations have required over six months. We have used the sterol extracts of two Gorgonians or soft corals, Pseudoplexaura Porosa and Plexaura Homomolla. Within these extracts we have discovered several new classes of marine sterols, including several twenty-two carbon sterols of unusual stereochemistry, a twenty-one carbon sterol, several new 5-BETA stanols, and a series of extremely interesting 19-nor-delta-5-sterols (publications in preparation). We feel certain that with the institution of the computer assisted procedures described herein, the time required for this kind of study (half a year) can be cut down to weeks.

Application of INTSUM to the marine sterol spectral files has just begun. One aspect of the INTSUM work which should be mentioned here is that in addition to the free 3-beta-hydroxy marine sterol files, a number of marine sterol derivatives (acetates, O-methyl ethers, trimethylsilyl ethers, and other derivatives) were compiled from the mass spectral library in this laboratory. INTSUM will be applied to these marine sterol derivative files in order to extract fragmentation rules. Comparison of the results for the free and derivatized sterols will point up the cases where some of the derivatives (which have superior GC properties) can be used with a minimum of loss of mass spectral information. We are confident that the file search system will be functioning before July. We already have marine extracts arriving from our collaborators in Brazil, and have offered the use of the system, once it is functioning, to researchers in Japan and Britain. We feel that the system will be of great benefit to the large number of researchers in the marine sterol field.

Another major area of interest in our chemical laboratories is the structural analysis of marine terpenoids using CONGEN in conjunction with a variety of spectroscopic data collected on these compounds. For the past year we have been involved in the application of CONGEN in the area of structural elucidation specifically related to marine natural products other than steroids. CONGEN's advantages in these studies lie chiefly in its ability to provide interactively the chemist with assurance that no plausible solutions have been overlooked, as well as an insightful measure of the progress of the problem, thereby suggesting clues to guide the course of the investigation.

(+)-Palustrol.

The utility of CONGEN has been demonstrated recently [57] in the identification of (+)-palustrol, a tricyclic sesquiterpene alcohol from the marine Xeniid *Cespitularia viridis*. Inferences derived from ¹H and ¹³C nmr spectra suggested molecular fragments whose assembly by CONGEN resulted in an initial set of 272 candidate structures. Examination of the set suggested appropriate nmr decoupling experiments resulting in the imposition of additional constraints which reduced the initial set of candidates to 88. Dehydration of the tertiary alcohol and spectral examination of the resulting olefins provided additional structural constraints which reduced the set further to 22. Recognition of an additional constraint after examining these possibilities eliminated two of the 22. Of the remaining 20 structures, only four (1 - 4) obey the isoprene rule, and of these four, 1 and 2 may be deleted because their dehydration would yield unsaturated analogs which violate Bredt's rule.



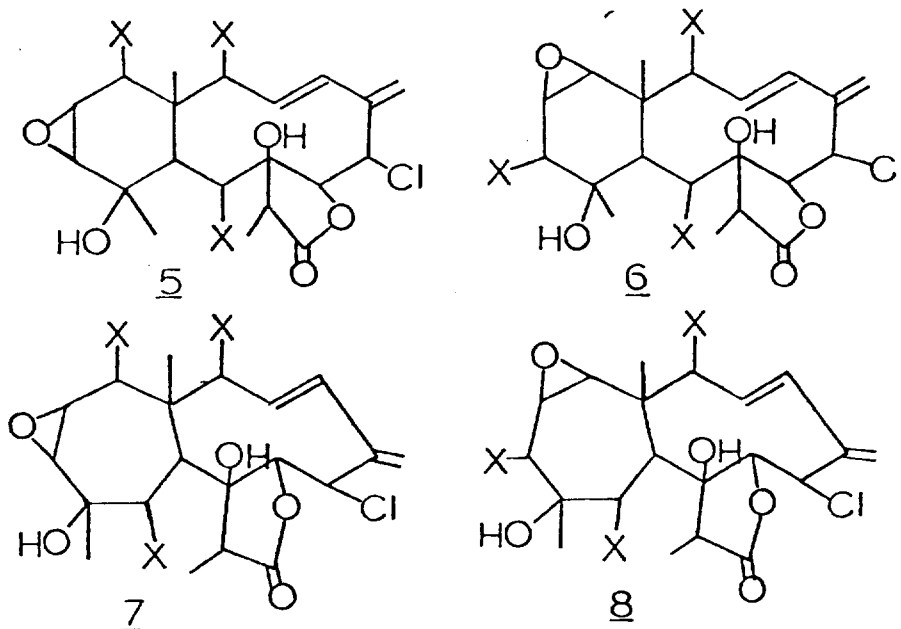
Examination of the literature revealed that structure 3 had been assigned to (-)-palustrol (L. Doleijs, V. Herout, and F. Sorm, CCCC, 26, 811 (1961)). The published infrared spectrum for

(-)-palustrol was identical in all respects to that of the unknown alcohol, thus establishing its structure. Our structure 3, however, displays the opposite rotation of polarized light.

Briareine D.

A recent study by Tursch and Bartholome (C. Bartholome, PhD. Thesis, University of Brussels, 1974) resulted in two alternative proposed structures for Briareine D, one of four chlorinated diterpene lactones isolated from the gorgonian *Briareum asbestinum*.

Rigorous examination of the structural inferences which led to the proposed structures yielded molecular fragments and constraints which were supplied to CONGEN for construction of structural candidates. The results confirmed the proposed structures, 5 and 6, and, more importantly, suggested two additional candidates (7,8) which had not been considered previously and could not be excluded on the basis of existing data.



(X = RCOO)

Work is currently in progress on the CONGEN-assisted structure elucidation of the aglycone portion of Lemnialioside, a diterpene glycoside from *Lemnalia digitata*, and a tricyclic sesquiterpene hydrocarbon from *Sinularia mayi*.

Further applications are summarized under headings of subsequent sections which refer to specific programs. Much of the effort in application of our programs to the mass spectral data implicitly assumes that the data are available. In fact, without the current and future instrumentation effort discussed in Part 1, these program applications would not be feasible.

3.2.1 CLEANUP

The spectral cleanup program, written for ourselves and our collaborators in the Dept. of Genetics, Stanford Hospital (see Local/Stanford Community, below) is now in routine use. A manuscript describing the method is now in press [61]. Several improvements have been made in the program to increase its capabilities for dealing with complex multiplets of overlapping GC peaks and to improve its efficiency. The resulting version of the program has been exported to several other laboratories which have expressed interest in our methods (see end of Part 3).

3.2.2 INTSUM

As a means of extending the rules of fragmentation in mass spectrometry, several classes of compounds are under study as we attempt to determine characteristic modes of fragmentation. The following is a brief description of each such class and the current status of our research:

1. **Pregnanes:** Pregnanes related to the progesterone skeleton have been analyzed in some detail in collaboration with Dr. S. Hammerum, (University of Copenhagen, Denmark). Two manuscripts describing this work have recently appeared [65,66].
2. **Androstanes:** Keto-substituted analogs of the skeleton of the important steroidal hydrocarbon, androstane, were being studied in collaboration with Dr. Roy Gritter (an IBM scientist who spent his sabbatical leave in our laboratory learning more about mass spectrometry). This study is important to our understanding of the mass spectral behavior of complex, polycyclic systems. It is providing a model for the use of Meta-DENDRAL programs. We have completed this study and a manuscript describing our method and results is now in press [58] in the Journal of the American Chemical Society.
3. **Macrolide Antibiotics :** We have finished the first stages of our analysis of the fragmentation of several members of these macrocyclic systems. We have solicited and obtained a small number of additional compounds to supplement our own limited number of samples. We are currently correlating the INTSUM results from closely related structures to identify systematic modes of fragmentation. We are designing experiments of deuterium labelling and metastable defocusing to help distinguish among alternative explanations by INTSUM for several prominent ions in the spectra of these compounds. Further efforts on this problem are hindered by lack of available standards.
4. **Insect Juvenile Hormones:** In collaboration with Dr. Loren Dunham, Zoecon Corp., we are investigating regularities in

the fragmentation behavior of the juvenile hormones. Previous work on the mass spectra of these compounds was carried out only at low resolving powers. We have obtained the high resolution mass spectral data for these compounds and have completed the INTSUM analysis of the data. Our findings have been described in a manuscript which will appear shortly [67] in Organic Mass Spectrometry. Our results will prove valuable for structural analysis and detection of these compounds and congeners.

- 5) Marine Sterols : The previous section summarizes our continuing efforts in marine sterol analysis, including the importance of INTSUM in these studies.

3.2.3 RULEGEN AND RULEMOD

As described above, RULEGEN and RULEMOD can be used to assist in discovery of mass spectrometry fragmentation rules which depend on substructural features of molecules. Thus, it can be used for classes of compounds where the fragmentation does not depend on the basic skeleton, but on local features expressed by common substructures. Our studies [58] on the performance of the program (see Meta-DENDRAL section) have involved analysis of spectra of previously well-characterized classes of compounds. We have analyzed spectra of aliphatic amines and estrogenic steroids in terms of fragmentation dependence on substructural features of these molecules. Excellent agreement with literature descriptions of fragmentation were obtained. We then proceeded with a study of the previously uncorrelated mono-, di- and triketoandrostanes. Our results [58] provide new insights into regularities of molecular fragmentation among members of the same group. The results also indicate little or no additivity of effects of keto substitution; spectra of diketoandrostanes are not superpositions of the respective monoketoandrostanes.

3.2.4 CONGEN

We are currently engaged in efforts to explore the utility of CONGEN to a variety of structure elucidation problems. The current areas of application are summarized below, together with progress to date.

- 1) Ion Structures: CONGEN has been used to construct possible ion structures under a variety of constraints in support of studies on the structures of ions in the mass spectrometer. These studies are crucial to a deeper understanding of molecular fragmentation. The programs results are used to ensure that no plausible alternatives have been overlooked during efforts to characterize the structures. We have recently published a detailed description of the use of CONGEN which illustrates the systematic approach available with the program [55].

- 2) Terpenoid Systems: We are using CONGEN to explore questions of the scope of terpenoid isomerism. We would like to determine some criteria which might allow us to say something about why only certain structural types are found in nature, to the exclusion of many possibilities which are very similar in structure. A manuscript describing our first results is now in press in Tetrahedron [60] and describes some aspects of the structural isomerism of mono- and sesquiterpenoid skeletons.
- 3) Scope of Structural Isomerism: We are investigating the philosophical and pedagogical aspects of the scope of structural isomerism. This investigation is important to our program design and strategy as we identify the ways persons consider and reject whole categories of structural possibilities. A manuscript describing this work has appeared in the Journal of Chemical Information and Computer Science [54].
- 4) Constraint Implementation: A detailed description of the kinds of constraints available to guide CONGEN in its exploration of structural possibilities has been presented [56]. This description also presents how constraints and efficient implementation of chemical "common sense" were derived from considerations of manual approaches to structural problems.
- 5) Marine Natural Products: The previous section described use of CONGEN in solving unknown structures in this area of application of our techniques.

3.3 Utilization of the Mass Spectrometry Resource

3.3.1 Applications of High Resolution Mass Spectrometry

A) Prof. Djerassi's Group

We have run about 75 samples to obtain high resolution mass spectra in support of DENDRAL research problems. These have included marine sterols (acquisition of reference spectra and verification of structures of new synthetic materials), macrolide antibiotics, ketoandrostanes and substituted pregnanes for Meta-DENDRAL studies of fragmentation processes.

B) Stanford Chemistry Department.

We have run a number of spectra for other researchers in the Department of Chemistry. Samples have included a number of diterpanes, alkaloids and unknown compounds from both chemical and enzymatic cyclization procedures.

C) Other Stanford Community

We have run spectra for a number of our collaborators in the Medical School. These have included samples from the Departments of Genetics, Psychiatry and Anaesthesia, representing structural analyses of metabolic products, drug purity and possible reaction products of an anesthetic, respectively.

D) U.S. and Foreign Collaborators.

Spectra have been obtained for Dr. Dunham, Zoecon Corp., of Juvenile hormones for INTSUM studies [67]; Dr. Gritter, now back at IBM, steroids for Meta-DENDRAL studies [58]; Dr. Fitch, Yale University, alkaloid metabolites; Dr. Tomer, Univ. of Brooklyn, spectra for fragmentation studies; Dr. Jaeger, Univ. of Wyoming, structure identification of crown ether components; Dr. Spangler, Univ. of Idaho, structure identification of sulfides for studies of remote sulfur-sulfur interaction in the mass spectrometer. High resolution spectra have been provided to Dr. Nakano, Venezuela, alkaloids, Drs. Mors and Gilbert, Brazil, steroids and alkaloids, Dr. Sultanbawa, Ceylon, triterpenes and alkaloids, and Dr. Orazi, Argentina, terpenoids.

3.3.2 Applications of GC/High Resolution Mass Spectrometry.

During the past year we have analyzed the following samples by GC/HRMS (these samples represent real applications and do not include the many samples of standard compounds which were analyzed during this time during development of the GC/HRMS system):

A) Prof. Djerassi's group - We have analyzed about 40 mixtures of marine natural products, primarily sterols, by GC/HRMS. Some samples were standard compounds necessary as reference materials but available only as mixtures. Some samples were mixtures of unknown compounds. Spectra were obtained primarily on underivatized sterols, occasionally from acetate derivatives.

B) Other Stanford collaborators - We have run GC/HRMS analyses of several mixtures of diterpenes and precursors, and enzymatic and chemical cyclization products of squalene epoxide analogs for Prof. van Tamelen, Dept. of Chemistry. We have analyzed ten urine fractions in conjunction with on-going work with Prof. Lederberg's group in the Dept. of Genetics. These have been primarily organic and amino acid fractions, derivatized as appropriate, and urinary polyamines analyzed as the trifluoroacetate derivatives.

3.3.3 Other Mass Spectral Studies

We have obtained a number of conventional mass spectra (low resolution) in cases where high resolution data were not required

or when the computer system was engaged in developmental work. For example, to build the library of low resolution mass spectral data of sterols, several LRMS of new compounds were obtained by GC/MS. Also, in collaboration with Dr. Mefflin, Dept. of Cardiology, we have investigated the use of specific ion monitoring as a tool for the identification of a new drug and its primary metabolite in human serum. Spectra of porphyrins have been obtained for Dr. Collman's group, Dept. of Chemistry, in studies of model systems for oxygen transport. We have obtained spectra of heart beat stimulants (digitoxigenins) for Prof. Kalman in Pharmacology, to verify identities of these compounds.

3.4 Applications of Programs by External Scientists

The DENDRAL project, one of the major users of the SUMEX-AIM computer facility, has formed a small community of regular, remote users. This "exodendral" community has continued to provide valuable contributions to program development, although the growth of this community has had to be slowed in response to increasing demands by other projects upon the SUMEX-AIM facility. As an example, for the months of September 1975 to February 1976, the number of CPU hours used by exodendral persons amounted to at least 8 percent of the CPU hours used by the DENDRAL project. There are currently four remote chemist-users whose groups regularly use CONGEN in their day to day work. Additionally, there are several remote users who use their accounts on an occasional basis, or who access SUMEX-AIM via the GUEST mechanism.

The SUMEX-AIM facility has grown markedly in number of projects over the past year. Due to this increase in system loading; the DENDRAL project, which had previously been able to offer trial usage of its programs to almost any chemist who expressed a need to use the programs, has found itself in the unfortunate position of having to carefully screen potential collaborators. Those chemists who have been granted access, have been requested to restrict their usage to off-prime time hours. CONGEN, the DENDRAL program which receives most of this usage, has evolved in a manner designed to try to remedy the system loading problem which can be created by the enthusiasm of its chemist-users. Since a typical, long GENERATE, PRUNE or IMBED within CONGEN can be very time consuming, as well as a voracious consumer of CPU cycles, a provision to permit a user to easily take advantage of SUMEX-AIM's off-hour batch processing has been implemented. A CONGEN user can now interactively set up his problem, and when ready to commence with a time consuming procedure, can, from within CONGEN, request automatic submission to BATCH, to be run late at night. The CONGEN users also benefit from this ability, in that they no longer must leave a terminal tied up during the sometimes hour-long compute times. This development then, can be viewed as responding to CONGEN users' needs as well as being an effort by the DENDRAL project to be conscientious in its resource-sharing responsibilities.

Following is a brief summary of the major users of CONGEN over the past year, as well as notes on chemists who contacted us about trial usage of the programs.

Dr. Clair Cheer, Professor of Chemistry, University of Rhode Island, Kingston, Rhode Island. Dr. Cheer is on sabbatical leave from the University of Rhode Island to the Stanford University Chemistry Department. He has, in recent work with Professor Djerassi's group, demonstrated the utility of CONGEN in the identification of (+)-Palustrol, a tricyclic sesquiterpene alcohol from the marine Xeniid *Cespitularia viridis* [57]. Dr. Cheer plans to continue his work with CONGEN once he returns to Rhode Island in December.

Dr. Jon Clardy, Professor of Chemistry, Iowa State University. Dr. Clardy read of CONGEN in an article appearing in the Journal of the American Chemical Society and contacted Professor Djerassi concerning the possibility of using the program from Iowa. He was offered GUEST access during the winter of 1975, but has not yet had an opportunity to evaluate the potentials of the program.

Dr. Douglas Dorman, Eli Lilly Corp., Indianapolis, Indiana. Dr. Dorman's research involves the identification and characterization of drug related compounds by chemical and spectroscopic methods. Using primarily the NMR and C13 NMR spectra of these various compounds, Dr. Dorman has found CONGEN to be a time-saving adjunct to his structure elucidation work (see letter in Appendix B).

Dr. H.M. Fales, National Heart and Lung Institute, Bethesda, Maryland. Dr. Fales, along with Doctors Sanford Markey and Peter Roller had a joint account set up for them in April of 1975. Most of the use of this account came during late summer at which time Dr. Fales experimented with the use of CONGEN for assistance in the elucidation of the structure of a novel quinolinone, known to be tumorigenic. Although the crystal structure had been solved at the time of his usage of CONGEN, Dr. Fales felt that the program produced an abundance of useful ideas. The main problem initially faced by Dr. Fales in using CONGEN was in getting a feel for problem size and the effects of various constraint types.

Professor Kenneth Gash, California State College at Dominguez Hills. Professor Gash is a professor of chemistry who is on temporary leave to Small College, the research branch of Dominguez Hills. Dr. Gash did some of the original work, in 1965, with Professor Morton Munk, on the structure elucidation program developed at Arizona State University. Dr. Gash has been reviewing some of the problems originally done with Munk's program and has been studying input, output and constraint capabilities found in CONGEN. He has generally concluded that if system response time was better, CONGEN provides an excellent tool for the chemist to use in structure elucidation problems.

Mr. Neil A. B. Gray, King's College, Cambridge, England. Mr. Gray, following a three week visit to the Stanford chemistry department, requested copies of all the current DENDRAL programs to be sent to him in England. He is a chemist who has been working in areas related to developments in various of the DENDRAL programs, and hopes to be able to benefit from work already done at Stanford. His current interest in intelligent constraint application during structure elucidation merges well with one of the directions in which CONGEN is tending to develop. Unfortunately, Mr. Gray does not have access to an ARPANET or TYMNET node to access SUMEX-AIM directly. Therefore, all collaboration has had to be carried on by mail.

Dr. Jerrold Karliner, Ciba Geigy Corporation, Ardsley, New York. Dr. Karliner and his research group at Ciba-Geigy have become regular users of CONGEN in their day-to-day operation of a research laboratory. Dr. Karliner is a completely self-taught user of CONGEN, and has served to encourage others to request permission to use this program. A letter from Dr. Karliner, describing his usage, is attached in Appendix B.

Dr. Milton Levenberg, Abbott Laboratories, Chicago, Illinois. Dr. Levenberg has been an occasional user of CONGEN as an adjunct to his work as head of a mass spectrometry laboratory. Primary usage has been to provide assurance that the proposal of a structure for a compound on the basis of chemical and spectroscopic evidence has not overlooked other plausible possibilities.

Dr. Gino Marco, Ciba Geigy Corporation, Greensboro, North Carolina. Dr. Marco heard about CONGEN during a company seminar presented by Dr. Karliner. After a brief trial use via the GUEST mechanism, Dr. Marco requested an account for use by his group of metabolic and organic chemists. Dr. Marco's research group studies unknown insect metabolites by micro-IR and micro-NMR methods, and attempts structure elucidation based on these forms of spectroscopic analysis. Testing the utility of the program before implementing it for day to day use, Dr. Marco discovered that CONGEN could greatly narrow the alternatives of complex metabolic conjugates which had to be considered in a typical elucidation problem. They have established a leased line to the nearest TYMNET node, and expect increased CONGEN usage in the future.

Dr. David Pensak, DuPont de Nemours and Company, Wilmington, Delaware. Indirectly requested information about CONGEN through a letter written by his immediate superior to Professor Lederberg. Dr. Pensak has been offered GUEST access, and has just begun a potential collaboration with a DENDRAL group which is studying model builders and their production of reliable geometries for certain types of molecules.

Professor Manfred Wolff, University of California at San Francisco. Dr. Wolff is chairman of the Department of

Pharmacological Chemistry, and inquired as to the possibilities of accessing SUMEX-AIM and appropriate programs for a faculty which is interested in many aspects of drug design and drug action, ranging from physical chemistry to purely biological studies. He has been encouraged to use GUEST access to explore CONGEN, although he has taken no action up to the present time.

We have cases where requests for GUEST access had to be denied due to system loading considerations. We made these decisions according to the extent to which the requested use would fit within the research guidelines of SUMEX/AIM and our own stated criteria from the 1973 proposal to NIH. In one case, for instance, the use was for an individual's report on potential educational uses of CONGEN.

The following two chemists have taken advantage of CONGEN by sending appropriate data and information to Stanford.

Professor L. Minale, Laboratorio per la Chimica di Molecole di Interesse Biologico del C.N.R., Napoli, Italy. Professor Minale has been collaborating with a member of the DENDRAL staff on the solution of the structure of the cyclic diether lipids from an unusual, very thermophilic bacterium (J. C. S. Chem. Comm. 543, 1974). Application of CONGEN resulted in five final candidate structures, several of which had not previously been considered. Work is currently underway to chemically differentiate the possibilities.

Professor Kogi Nakanishi, Department of Chemistry, Columbia University. Professor Nakanishi is one of the most active and productive persons engaged in structure elucidation activities. He has developed an active interest in CONGEN and is collaborating with us on several novel problems. One of these problems has involved the structure of the active component of defense secretions of an insect (termite). Other defense secretion components are under investigation as we explore structural alternatives based on current data.

3.5 Export of GC/MS Programs to Other Sites.

There has been a demand for the GC/LRMS programs developed for data resolution and cleanup [61], and in the past several months these programs have been distributed, upon request, to five different groups. Each group was supplied with a tape, in a format appropriate to the type of computer system they were using, containing the cleanup program, as well as extensive documentation describing the program, related file structure and the required dependent utility programs.

The following research groups have already received their copies of these programs:

Professor G. Eglinton, University of Bristol, England.

Dr. Tom Elwood, Department of Chemistry, Univ. of Utah.

Dr. J. Lawless, Ames Research Center, California.

Dr. Charles Sweeley, Dept. of Biochemistry, Michigan State University.

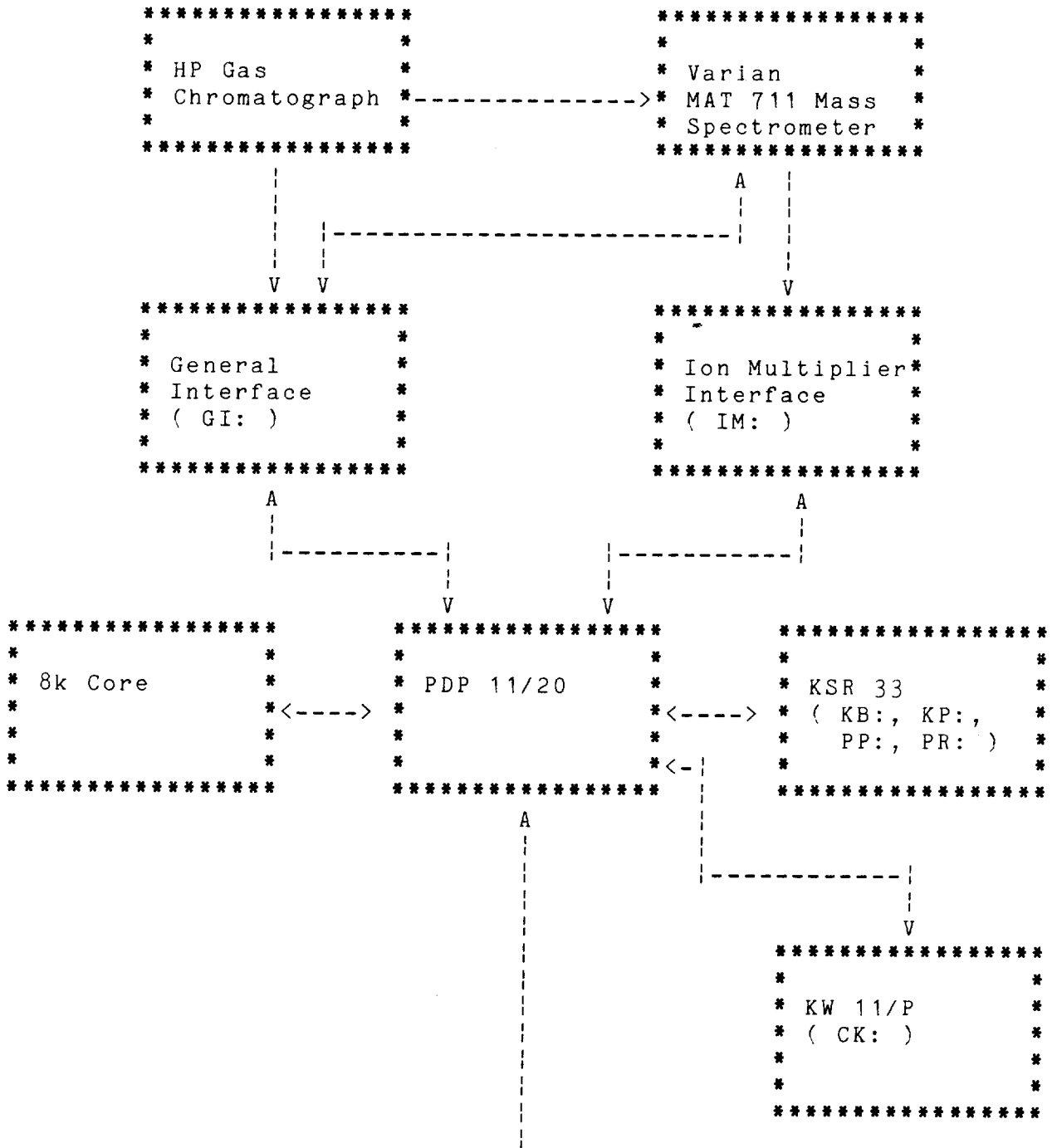
Mr. J. R. Wilcox, AEI Scientific Apparatus Inc., Elmsford, New York.

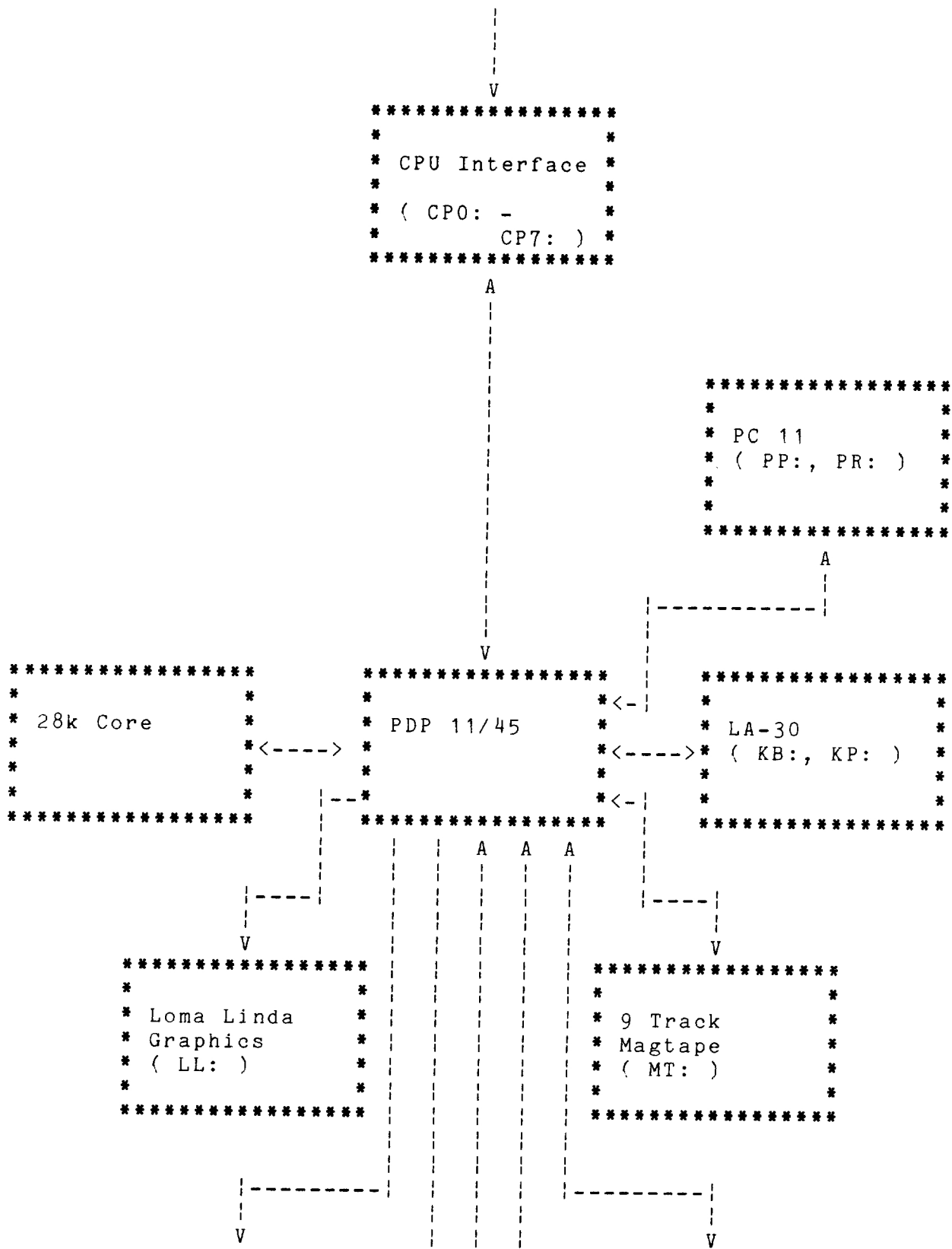
Dr. Philip Fishman, Department of Biochemistry, Washington University School of Medicine at St. Louis, Missouri has contacted us for further information concerning computer resolution of GC/MS elutants. He has been provided with the necessary information and invited to request the exportable versions of the programs.

4 BIBLIOGRAPHY

See Section II-D, SUMMARY OF PUBLICATIONS, for a listing of publications by this project.

Figure 1. Current Hardware Configuration





```
*****
*
* Line Printer *
* ( LP: ) *
*
*
*****
```

```
*****
*
* Disk Drives *
* ( DK0: - *
*      DK3: ) *
*
*
*****
```

V

V

```
*****
*
* Calcomp *
* Graphics *
* ( CC: ) *
*
*****
```

```
*****
*
* KW 11/L *
*
*
*****
```

V

```
*****
*
* TTY line to *
* PDP 10 *
*
*****
```

Appendix A. Interrupt Features in a Sample CONGEN Run

```
.
.
(program session has been in progress)
.
GENERATE
DO YOU WISH TO PRUNE WHILE GENERATING?(Y FOR YES):Y
SHALL I USE THE GLOBAL CONSTRAINT LIST?(Y FOR YES):Y
.
.
(computing)
.
.
S
17 structures have been generated so far
Shall I draw some?(Y for yes):Y
FROM:3
TO:3

3
  C
  =  P
C=C-C-C=H
  N
More?(Y for yes):N
Normal computation will now resume
.
.
27 STRUCTURES WERE GENERATED
.
(session continues)
.
IMBED
SUPERATOM NAME:PHN
DO YOU WISH TO PRUNE WHILE IMBEDDING?(Y FOR YES):N
.
.
(computing)
.
.
I
structure 8 of the original 19 is undergoing imbedding;
13 structures have been obtained
.
.
(computing continues)
.
.
32 STRUCTURES WERE OBTAINED
DRAWSOME
FROM:4
.
.
```


Appendix B. Letters from Collaborators

LILLY RESEARCH LABORATORIES

DIVISION OF ELI LILLY AND COMPANY • INDIANAPOLIS, INDIANA 46206 • TELEPHONE (317) 636-2211

May 6, 1976

Dr. Ray Carhart
Department of Computer Sciences
Stanford University
Stanford, California 94305

Dear Ray:

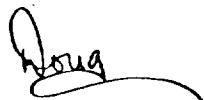
We are pleased to have an opportunity to offer our support and encouragement of your CONGEN effort.

We are using CONGEN to aid us in the solution of current structure elucidation problems. When used at an early stage, we find that the program is invaluable in suggesting the various compound types which are consistent with the extant data. In one case, for example, CONGEN generated a list of possible heteroaromatic ring systems. Such a list is particularly useful in the elucidation of the structures of pharmaceutical molecules, in that it allows us to correlate these possibilities with the physical and biological properties of a vast number of known pharmaceuticals.

In short, we have been quite pleased with our experience with CONGEN. Owing to possible security problems, we have not used the program as much as we would like. If the program were available in exportable form, I believe that usage would increase dramatically. Probably the full potential of the program will not be realized until this occurs. Even in its present form, however, CONGEN is a valuable contribution to chemistry.

Sincerely,

LILLY RESEARCH LABORATORIES



Douglas E. Dorman, Ph.D.
Physical Chemistry Research
Department MC525

CIBA-GEIGY Corporation
Ardsley, New York 10502
Telephone 914 478 3131

CIBA—GEIGY

April 19, 1976

Professor Carl Djerassi
Department of Chemistry
Stanford University
Stanford, California 94305

Dear Professor Djerassi:

Thank you for permitting us to use the CONGEN program for the past year. We have used this program only for the most demanding structure problems and therefore have not been a heavy user of the system. However, when utilized, we have found the CONGEN program to be a valuable and effective aid for structure elucidation problems.

We have used CONGEN to provide assurance that once a structure has been proposed on the basis of chemical and spectroscopic analysis that other, plausible structures have not been overlooked, thereby providing a greater confidence limit to our structural proposals. We have also utilized the program to determine possible structures when the analytical data does not provide a unique structure, and in this manner acquire additional chemical structures for consideration. Both approaches represent assets to structure determination work.

Clearly, the scope of CONGEN would be extended by making it available to other industrial chemists involved with determining structures of organic compounds. As you know, the industrial setting demands accurate results by the most rapid and economical means available. Exposing CONGEN to a wider range of industrial chemists would provide your staff with more experience with diverse structure problems and hence extend the versatility and utility of the programs. However, in order to attain more participation, the computer

Professor Djerassi

-2-

April 19, 1976

system would have to be made more readily accessible, especially during peak hours. We have been fortunate in this regard since the time zone difference between Stanford and Ardsley enables us to use CONGEN early in the morning when few others are using the SUMEX programs. However, when we attempt to use the program later in the day, the system becomes fairly slow, to the point where at times we find it more economical to simply log off and resume the following morning.

We consider CONGEN as another complementary structure elucidation method at CIBA-GEIGY and look forward to its continued use in the future.

Sincerely yours,



Jerrold Karliner Ph.D.
Group Leader
Analytical Research Department

Table I. Marine Sterol Standards

| | | 4-DEMETHYL MARINE STEROIDS | | | | | | Sept. 1975 | |
|-----------------|-----------------------|--|---|---|----------------|-----|-------|------------|--|
| | | | | | | | | | |
| | | | | | | | | | |
| | A-4 | A-3 ex M. Kikuyoshi | A-1 ex M. Boshara A-2 ex M. Boshara | A-15 ex L.J.G. and | B-2 11-9-73 | B-1 | D-20 | E-8 | |
| | A-18 11-9-73 | A-11 11-9-73 | A-5 11-9-73 | A-16 | | | E-1 | E-9 | |
| | A-19 ex L.J.G. and | A-12 ex L.J.G. and | A-6 | | | | E-2 * | | |
| | | | A-7 | | | | | | |
| | | A-13 | A-8 11-9-73 | A-17 | | | | | |
| | A-20 | A-14 ex M. Kikuyoshi | A-9 ex M. Kikuyoshi A-10 ex L.J.G. and | | | | | | |
| | | | B-3 11-9-73 | B-16 | | | E-3 | E-10 | |
| | B-20 | B-11 ex M. Kikuyoshi ex BA Kikuyoshi | B-4 | B-17 | | | | | |
| | C-1 | B-12 | B-5 | B-18 | | | E-4 | E-11 | |
| | C-2 11-9-73 | B-13 | B-6 11-9-73 | B-19 | | | | | |
| | C-3 | B-14 ex L.J.G. and ex BA Kikuyoshi | B-7 | | | | | | |
| | | | B-8 ex L.J.G. and | | | | | | |
| | | C-14 KALLAZAWA | B-9 | | | | | | |
| | C-4 ex L.J.G. and | B-15 ex L.J.G. and | B-10 11-9-73 | | | | E-5 | | |
| | D-8 11-9-73 | C-18 | C-5 11-9-73 | D-4 | | | E-6 | E-12 | |
| | D-9 | C-19 ex M. Kikuyoshi ex BA Kikuyoshi | C-6 11-9-73 | D-5 ex M. Kikuyoshi ex BA Kikuyoshi | | | | | |
| | | C-20 ex M. Kikuyoshi ex BA Kikuyoshi | C-7 11-9-73 | D-6 ex M. Kikuyoshi ex BA Kikuyoshi | | | E-7 | E-13 | |
| | | D-1 | C-8 11-9-73 | D-7 | | | | | |
| | D-10 | D-2 | C-9 11-9-73 | | | | | | |
| | | | C-10 ? | | | | | | |
| | D-11 | D-3 ex L.J.G. and | C-11 ex L.J.G. and | | | | | | |
| | | | C-12 ex L.J.G. and | | | | | | |
| | | C-13 | | | | | | | |
| | | C-14 | | | | | | | |
| | | C-15 | | | | | | | |
| | | C-16 | | | | | | | |
| | | C-17 | | | | | | | |
| | D-13 | D-12 + | | | | | | | |
| D-19 11-9-73 | D-18 | D-14 11-9-73 | | | | | | | |
| | | D-15 | | | | | | | |
| | | D-16 | | | | | | | |
| | | D-17 | | | | | | | |

indicates that at least a 1 mg sample is contained in the sample box.
 Shaded boxes indicate sterols which have not been found in marine sources.
 indicates C-24 epimers indistinguishable by GC-MS.
 indicates a grouping of sterol side chains of identical carbon number: 7,8,9, 10, and 11.
 * Note that this side chain contains 10 carbons rather than the 11 indicated by position within bracket.
 x Mass Spectrum obtained

TABLE II
 Evaluation of Several Fucosterol
 (Marine Algae Sterol) Mass Spectra from the
 Djerassi Mass Spectral Files

LIBRARY SEARCH REPORT FOR EXPERIMENT 21
 SEARCHING 25 SPECTRA IN MARINE ON DK2

| (1) SPEC # | 0 | RETIND | 0 | MOLION | CANDS: | 0 | 0 | 0 | 0 | 0 | CR | 0 |
|-------------|-----|--------|--------|--------|---|-------------|---|---|---|---|----|-----------------|
| RK | RAW | AVG | RETIND | MOL | CHEMICAL NAME | AMPLITUDES: | 0 | 0 | 0 | 0 | 0 | |
| 999 | 87 | 0 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | CR MASTER # SRC |
| 729 | 62 | 0 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | 0 MARINE 18 6 |
| 689 | 51 | 0 | 0 | 426 | (24Z)-24-PROPYLIDENECHOLEST-5-EN-3BETA-OL | | | | | | | 0 MARINE 21 6 |
| 578 | 44 | 0 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | 0 MARINE 19 6 |
| 544 | 67 | 0 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | 0 MARINE 20 6 |

| (2) SPEC # | 0 | RETIND | 0 | MOLION | CANDS: | 0 | 0 | 0 | 0 | 0 | CR | 0 |
|-------------|-----|--------|--------|--------|--|-------------|---|---|---|---|----|-----------------|
| RK | RAW | AVG | RETIND | MOL | CHEMICAL NAME | AMPLITUDES: | 0 | 0 | 0 | 0 | 0 | |
| 999 | 76 | 0 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | CR MASTER # SRC |
| 680 | 51 | 0 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | 0 MARINE 19 6 |
| 574 | 50 | 999 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | 0 MARINE 21 6 |
| | | | | | | | | | | | | 0 MARINE 18 6 |

| (3) SPEC # | 0 | RETIND | 0 | MOLION | CANDS: | 0 | 0 | 0 | 0 | 0 | CR | 0 |
|-------------|-----|--------|--------|--------|--|-------------|---|---|---|---|----|-----------------|
| RK | RAW | AVG | RETIND | MOL | CHEMICAL NAME | AMPLITUDES: | 0 | 0 | 0 | 0 | 0 | |
| 999 | 123 | 0 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | CR MASTER # SRC |
| 781 | 61 | 999 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | 0 MARINE 20 6 |
| 694 | 59 | 0 | 0 | 412 | (24E)-STIGMASTA-5,24(28)-DIEN-3BETA-OL | | | | | | | 0 MARINE 18 6 |
| | | | | | | | | | | | | 0 MARINE 21 6 |

II.D. SUMMARY OF PUBLICATIONS

DENDRAL PUBLICATIONS

- (1) J. Lederberg, "DENDRAL-64 - A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs", (technical reports to NASA, also available from the author and summarized in (12)). (1a) Part I. Notational algorithm for tree structures (1964) CR.57029 (1b) Part II. Topology of cyclic graphs (1965) CR.68898 (1c) Part III. Complete chemical graphs; embedding rings in trees (1969)
- (2) J. Lederberg, "Computation of Molecular Formulas for Mass Spectrometry", Holden-Day, Inc. (1964).
- (3) J. Lederberg, "Topological Mapping of Organic Molecules", Proc. Nat. Acad. Sci., 53:1, January 1965, pp. 134-139.
- (4) J. Lederberg, "Systematics of organic molecules, graph topology and Hamilton circuits. A general outline of the DENDRAL system." NASA CR-48899 (1965)
- (5) J. Lederberg, "Hamilton Circuits of Convex Trivalent Polyhedra (up to 18 vertices), Am. Math. Monthly, May 1967.
- (6) G. L. Sutherland, "DENDRAL - A Computer Program for Generating and Filtering Chemical Structures", Stanford Artificial Intelligence Project Memo No. 49, February 1967.
- (7) J. Lederberg and E. A. Feigenbaum, "Mechanization of Inductive Inference in Organic Chemistry", in B. Kleinmuntz (ed) Formal Representations for Human Judgment, (Wiley, 1968) (also Stanford Artificial Intelligence Project Memo No. 54, August 1967).
- (8) J. Lederberg, "Online computation of molecular formulas from mass number." NASA CR-94977 (1968)
- (9) E. A. Feigenbaum and B. G. Buchanan, "Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry", in Proceedings, Hawaii International Conference on System Sciences, B. K. Kinariwala and F. F. Kuo (eds), University of Hawaii Press, 1968.
- (10) B. G. Buchanan, G. L. Sutherland, and E. A. Feigenbaum, "Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry". In Machine Intelligence 4 (B. Meltzer and D. Michie, eds) Edinburgh University Press (1969), (also Stanford Artificial Intelligence Project Memo No. 62, July 1968).
- (11) E. A. Feigenbaum, "Artificial Intelligence: Themes in the Second Decade". In Final Supplement to Proceedings of the

IFIP68 International Congress, Edinburgh, August 1968 (also Stanford Artificial Intelligence Project Memo No. 67, August 1968).

- (12) J. Lederberg, "Topology of Molecules", in The Mathematical Sciences - A Collection of Essays, (ed.) Committee on Support of Research in the Mathematical Sciences (COSRIMS), National Academy of Sciences - National Research Council, M.I.T. Press, (1969), pp. 37-51.
- (13) G. Sutherland, "Heuristic DENDRAL: A Family of LISP Programs", Stanford Artificial Intelligence Project Memo No. 80, March 1969.
- (14) J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference I. The Number of Possible Organic Compounds: Acyclic Structures Containing C, H, O and N". Journal of the American Chemical Society, 91.
- (15) A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference II. Interpretation of Low Resolution Mass Spectra of Ketones". Journal of the American Chemical Society, 91:11 (May 21, 1969).
- (16) B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, "Toward an Understanding of Information Processes of Scientific Inference in the Context of Organic Chemistry", in Machine Intelligence 5, (B. Meltzer and D. Michie, eds) Edinburgh University Press (1970), (also Stanford Artificial Intelligence Project Memo No. 99, September 1969).
- (17) J. Lederberg, G. L. Sutherland, B. G. Buchanan, and E. A. Feigenbaum, "A Heuristic Program for Solving a Scientific Inference Problem: Summary of Motivation and Implementation", Stanford Artificial Intelligence Project Memo No. 104, November 1969.
- (18) C. W. Churchman and B. G. Buchanan, "On the Design of Inductive Systems: Some Philosophical Problems". British Journal for the Philosophy of Science, 20 (1969), pp. 311-323.
- (19) G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference III. Aliphatic Ethers Diagnosed by Their Low Resolution Mass Spectra and NMR Data". Journal of the American Chemical Society, 91.
- (20) A. Buchs, A. M. Duffield, G. Schroll, C. Djerassi, A. B.

- Delfino, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Applications of Artificial Intelligence For Chemical Inference. IV. Saturated Amines Diagnosed by Their Low Resolution Mass Spectra and Nuclear Magnetic Resonance Spectra", *Journal of the American Chemical Society*, 92, 6831 (1970).
- (21) Y.M. Sheikh, A. Buchs, A.B. Delfino, G. Schroll, A.M. Duffield, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference V. An Approach to the Computer Generation of Cyclic Structures. Differentiation Between All the Possible Isomeric Ketones of Composition C₆H₁₀O", *Organic Mass Spectrometry*, 4, 493 (1970).
- (22) A. Buchs, A.B. Delfino, A.M. Duffield, C. Djerassi, B.G. Buchanan, E.A. Feigenbaum and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference VI. Approach to a General Method of Interpreting Low Resolution Mass Spectra with a Computer", *Helvetica Chimica Acta*, 53, 1394 (1970).
- (23) E.A. Feigenbaum, B.G. Buchanan, and J. Lederberg, "On Generality and Problem Solving: A Case Study Using the DENDRAL Program". In *Machine Intelligence 6* (B. Meltzer and D. Michie, eds.) Edinburgh University Press (1971). (Also Stanford Artificial Intelligence Project Memo No. 131.)
- (24) A. Buchs, A.B. Delfino, C. Djerassi, A.M. Duffield, B.G. Buchanan, E.A. Feigenbaum, J. Lederberg, G. Schroll, and G.L. Sutherland, "The Application of Artificial Intelligence in the Interpretation of Low-Resolution Mass Spectra", *Advances in Mass Spectrometry*, 5, 314 (1971),
- (25) B.G. Buchanan and J. Lederberg, "The Heuristic DENDRAL Program for Explaining Empirical Data". In proceedings of the IFIP Congress 71, Ljubljana, Yugoslavia (1971). (Also Stanford Artificial Intelligence Project Memo No. 141.)
- (26) B.G. Buchanan, E.A. Feigenbaum, and J. Lederberg, "A Heuristic Programming Study of Theory Formation in Science." In proceedings of the Second International Joint Conference on Artificial Intelligence, Imperial College, London (September, 1971). (Also Stanford Artificial Intelligence Project Memo No. 145.)
- (27) Buchanan, B. G., Duffield, A.M., Robertson, A.V., "An Application of Artificial Intelligence to the Interpretation of Mass Spectra", *Mass Spectrometry Techniques and Appliances*, G. W. A. Milne, Ed., John Wiley & Sons, Inc., 1971, p. 121.
- (28) D.H. Smith, B.G. Buchanan, R.S. Engelmores, A.M. Duffield, A.

- Yeo, E.A. Feigenbaum, J. Lederberg, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference VIII. An Approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids", Journal of the American Chemical Society, 94, 5962 (1972).
- (29) B.G. Buchanan, E.A. Feigenbaum, and N.S. Sridharan, "Heuristic Theory Formation: Data Interpretation and Rule Formation". In Machine Intelligence 7, Edinburgh University Press (1972).
- (30) J. Lederberg, "Rapid Calculation of Molecular Formulas from Mass Values". Journal of Chemical Education, 49, 613 (1972).
- (31) H. Brown, L. Masinter, and L. Hjelmeland, "Constructive Graph Labeling Using Double Cosets". Discrete mathematics, 7, 1 (1974). (Also Computer Science Memo 318, 1972).
- (32) B. G. Buchanan, Review of Hubert Dreyfus' "What Computers Can't Do: A Critique of Artificial Reason", Computing Reviews (January, 1973). (Also Stanford Artificial Intelligence Project Memo No. 181)
- (33) D. H. Smith, B. G. Buchanan, R. S. Engelmores, H. Aldercreutz and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference IX. Analysis of Mixtures Without Prior Separation as Illustrated for Estrogens". Journal of the American Chemical Society 95, 6078 (1973).
- (34) D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference X. Intsum. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids". Tetrahedron, 29, 3117 (1973).
- (35) B. G. Buchanan and N. S. Sridharan, "Rule Formation on Non-Homogeneous Classes of Objects". In proceedings of the Third International Joint Conference on Artificial Intelligence (Stanford, California, August, 1973). (Also Stanford Artificial Intelligence Project Memo No. 215.)
- (36) D. Michie and B.G. Buchanan, "Current Status of the Heuristic DENDRAL Program for Applying Artificial Intelligence to the Interpretation of Mass Spectra", in "Computers for Spectroscopy," R.A.G. Carrington, Ed., Adam Hilger, London, 1973. Also: University of Edinburgh, School of Artificial Intelligence, Experimental Programming Report No. 32 (1973).
- (37) H. Brown and L. Masinter, "An Algorithm for the Construction of the Graphs of Organic Molecules", Discrete Mathematics,

8, 227 (1974). (Also Stanford Computer Science Dept. Memo STAN-CS-73-361, May, 1973)

- (38) D.H. Smith, L.M. Masinter and N.S. Sridharan, "Heuristic DENDRAL: Analysis of Molecular Structure," in "Computer Representation and Manipulation of Chemical Information," W.T. Wipke, S. Heller, R. Feldmann and E. Hyde, Eds., John Wiley and Sons, Inc., 1974, p. 287.
- (39) R. Carhart and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference XI: The Analysis of C13 NMR Data for Structure Elucidation of Acyclic Amines", Journal of the Chemical Society (Perkin II), 1753 (1973).
- (40) L. Masinter, N.S. Sridharan, R. Carhart and D.H. Smith, "Application of Artificial Intelligence for Chemical Inference XII: Exhaustive Generation of Cyclic and Acyclic Isomers". Journal of the American Chemical Society, 96, 7702 (1974). (Also Stanford Artificial Intelligence Project Memo No. 216.)
- (41) L. Masinter, N.S. Sridharan, R. Carhart and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XIII. Labeling of Objects having Symmetry". Journal of the American Chemical Society, 96, 7714 (1974).
- (42) N.S. Sridharan, Computer Generation of Vertex Graphs, Stanford CS Memo STAN-CS-73-381, July, 1973.
- (43) N.S. Sridharan, et.al., A Heuristic Program to Discover Syntheses for Complex Organic Molecules, Stanford CS Memo STAN-CS-73-370, June, 1973. (Also Stanford Artificial Intelligence Project Memo No. 205.)
- (44) N.S. Sridharan, Search Strategies for the Task of Organic Chemical Synthesis, Stanford CS Memo STAN-CS-73-391, October, 1973. (Also Stanford Artificial Intelligence Project Memo No. 217.)
- (45) R. G. Dromey, B. G. Buchanan, J. Lederberg and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIV. A General Method for Predicting Molecular Ions in Mass Spectra". Journal of Organic Chemistry, 40, 770 (1975).
- (46) D. H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XV. Constructive Graph Labelling Applied to Chemical Problems. Chlorinated Hydrocarbons". Analytical Chemistry, 47, 1176 (1975).
- (47) R. E. Carhart, D. H. Smith, H. Brown and N. S. Sridharan, "Applications of Artificial Intelligence for Chemical Inference. XVI. Computer Generation of Vertex Graphs and Ring Systems". Journal of Chemical Information and Computer Science, 15, 124 (1975).

- (48) R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XVII. An Approach to Computer-Assisted Elucidation of Molecular Structure". *Journal of the American Chemical Society*, 97, 5755 (1975).
- (49) B. G. Buchanan, "Scientific Theory Formation by Computer." In *Proceedings of NATO Advanced Study Institute on Computer Oriented Learning Processes*, 1974, Bonas, France.
- (50) E. A. Feigenbaum, "Computer Applications: Introductory Remarks," in "Proceedings of Federation of American Societies for Experimental Biology," 33, 2331 (1974).
- (51) R. Davis and J. King, "Overview of Production Systems" To appear in *Machine Representation of Knowledge*, *Proceedings of the NATO ASI Conference*, July, 1975. (Also Stanford Artificial Intelligence Project Memo No. ***)
- (52) B. G. Buchanan, "Applications of Artificial Intelligence to Scientific Reasoning." In *Proceedings of Second USA-Japan Computer Conference*, American Federation of Information Processing Societies Press, August, 1975.
- (53) R. E. Carhart, S. M. Johnson, D. H. Smith, B. G. Buchanan, R. G. Dromey, J. Lederberg, "Networking and a Collaborative Research Community: A Case Study Using the DENDRAL Program," in "Computer Networking and Chemistry", P. Lykos, Ed., American Chemical Society, Washington, D.C., 1975, p. 192.
- (54) D. H. Smith, "The Scope of Structural Isomerism" (Paper XVIII in our series of AI Applications in Chemistry). *Journal of Chemical Information and Computer Sciences*, 15, 203 (1975).
- (55) D. H. Smith, J. P. Konopelski and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIX. Computer Generation of Ion Structures." *Organic Mass Spectrometry*, 11, 86 (1976).
- (56) R. E. Carhart and D. H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XX. 'Intelligenceint' Use of Constraints in Computer-Assisted Structure Elucidation," *Computers in Chemistry*, in press.
- (57) C. Cheer, D. H. Smith, C. Djerassi, B. Tursch, J. C. Braekman, and D. Dalozze, "Applications of Artificial Intelligence for Chemical Inference. XXI. Chemical Studies of Marine Invertebrates. XVII. The Computer-Assisted Identification of [+]-Palustrol in the Marine Organism *Cespitularia* sp., aff. *Subvirdis*," *Tetrahedron*, in press.

- (58) B. G. Buchanan, D. H. Smith, W. C. White, R. Gritter, E. A. Feigenbaum, J. Lederberg and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XXII. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program." *Journal of the American Chemical Society*, in press.
- (59) T. R. Varkony, R. E. Carhart, and D. H. Smith, "Computer-Assisted Structure Elucidation. Modelling Chemical Reaction Sequences Used in Molecular Structure Problems" (Paper XXIII in our series of A.I. Applications in Chemistry), in "Computer-Assisted Organic Synthesis", W. T. Wipke, Ed., American Chemical Society, Washington, D.C., 1976, in press.
- (60) D. H. Smith and R. E. Carhart, "Structural Isomerism of Mono- and Sesquiterpenoid Skeletons," (Paper XXIV in our series of A.I. Applications in Chemistry), *Tetrahedron*, in press.
- (61) R. G. Dromey, M. J. Stefik, T. Rindfleisch, and A. M. Duffield, "Extraction of Mass Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography/Mass Spectrometry Data," *Analytical Chemistry*, in press.
- (62) H. Eggert and C. Djerassi, "The Carbon-13 Magnetic Resonance Spectra of Acyclic Aliphatic Amines," *Journal of American Chemical Society*, 95, 3710 (1973).
- (63) H. Eggert and C. Djerassi, "Carbon-13 Nuclear Magnetic Resonance Spectra of Keto Steroids," *Journal of Organic Chemistry*, 38, 3788 (1973).
- (64) H. Eggert, C. VanAntwerp, N. Bhacca and C. Djerassi, "Carbon-13 Nuclear Magnetic Resonance Spectra of Hydroxy Steroids," *Journal of Organic Chemistry*, 41, 71 (1976).
- (65) S. Hammerum and C. Djerassi. "Mass Spectrometry in Structural and Stereochemical Problems CCXLV. The Electron Impact Induced Fragmentation Reactions of 17-oxygenated Progesterones." *Steroids*, 25, 817 (1975).
- (66) S. Hammerum and C. Djerassi. "Mass Spectrometry in Structural and Stereochemical Problems CCXLIV. The Influence of Substituents and Stereochemistry on the Mass Spectral Fragmentation of Progesterone." *Tetrahedron*, 31, 2391 (1975).
- (67) L. L. Dunham, C. A. Henrick, D. H. Smith, and C. Djerassi, "Mass Spectrometry in Structural and Stereochemical Problems. CCXLVI. Electron Impact Induced Fragmentation of Juvenile Hormone Analogs," *Org. Mass Spectrom.*, in press.

(68) C. Djerassi, Foreword to "¹³C NMR-Spectroscopy," by E. Breitmayer and W. Voelter, Verlag Chemie GmbH, Weinheim/Bergstr., 1974.