SECTION I

Form Approved
O.M.B. 68-R0249

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 - Computer in Chemistry (RR-00612 renewal)

2. PRINCIPAL INVESTIGATOR

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

2B. TITLE OF POSITION
Professor of Chemistry

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

2D. DEGREE
Ph.D.

2E. SOCIAL SECURITY NO.
[redacted]

2F. TELEPHONE DATA — Area Code: 415 — TELEPHONE NUMBER AND EXTENSION: 497-2783

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

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

3. DATES OF ENTIRE PROPOSED PROJECT PERIOD *(This application)*
FROM: May 1, 1980 — THROUGH: April 30, 1985

4. TOTAL DIRECT COSTS REQUESTED FOR PERIOD IN ITEM 3
$1,611,719

5. DIRECT COSTS REQUESTED FOR FIRST 12-MONTH PERIOD
$511,400

6. PERFORMANCE SITE(S) *(See Instructions)*
Stanford University
Stanford, CA 94305

7. Research Involving Human Subjects *(See Instructions)*
A. [X] NO B. [ ] YES Approved: _____ Date
C. [ ] YES – Pending Review

8. Inventions *(Renewal Applicants Only - See Instructions)*
A. [X] 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, CA 94305
IRS No. 94-1156365
Congressional District 12

10. NAME, TITLE, AND TELEPHONE NUMBER OF OFFICIAL(S) SIGNING FOR APPLICANT ORGANIZATION(S)
Larry J. Lollar
Sponsored Projects Officer

Telephone Number (s) 415-497-2883

11. TYPE OF ORGANIZATION *(Check applicable item)*
[ ] FEDERAL [ ] STATE [ ] LOCAL [X] 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
Assoc. Vice President-Controller
Stanford University
Stanford, CA 94305
Telephone Number 415-497-2251

13. IDENTIFY ORGANIZATIONAL COMPONENT TO RECEIVE CREDIT FOR INSTITUTIONAL GRANT PURPOSES *(See Instructions)*

14. ENTITY NUMBER (Formerly PHS Account Number)
IRS No. 94-1156365

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
	B. SIGNATURE(S) OF PERSON(S) NAMED IN ITEM 10	DATE

SECTION 1

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE

RESEARCH OBJECTIVES

LEAVE BLANK

PROJECT NUMBER

NAME AND ADDRESS OF APPLICANT ORGANIZATION

Stanford University, Stanford, CA 94305

NAME, SOCIAL SECURITY NUMBER, OFFICIAL TITLE, AND DEPARTMENT OF ALL PROFESSIONAL PERSONNEL ENGAGED ON PROJECT, BEGINNING WITH PRINCIPAL INVESTIGATOR

Carl Djerassi	[redacted]	Professor of Chemistry	Department of Chemistry
Dennis H. Smith		Senior Research Associate	Department of Chemistry
Bruce G. Buchanan		Adjunct Professor	Department of Computer Science
James G. Nourse	[redacted]	Research Associate	Department of Chemistry
Neil A. B. Gray	--	Research Associate	Department of Computer Science

TITLE OF PROJECT

Resource-Related Research - Computers in Chemistry

USE THIS SPACE TO ABSTRACT YOUR PROPOSED RESEARCH. OUTLINE OBJECTIVES AND METHODS. UNDERSCORE THE KEY WORDS (NOT TO EXCEED 10) IN YOUR ABSTRACT.

Our proposed research concerns computer-assisted structure elucidation of organic compounds of biological importance. We propose to make a quantum jump in both performance of existing and new programs and their availability to a nationwide community of biomedical scientists. We will build more powerful programs (basing our efforts on the solid foundation of programs developed under earlier grant support by: a) assembling a Semi-Automated Structure Elucidation System (SASES) which will act as a computer-based "laboratory" for carrying out experiments involving computer representation and manipulation of chemical structures, including structure elucidation, spectral data interpretation and prediction and conformational analysis to establish relationships between three-dimensional structures and their biological and chemical properties (structure/property relationships); b) developing at the heart of SASES the GENOA program, a method for structure Generation with Overlapping Atoms, which will include as a component our existing CONGEN program; c) developing a method for constrained generation of molecular conformations; and d) extending our topological structure representations to include both configurational and conformational stereochemistry and infusing proper treatment of stereochemistry throughout our computer programs. We will increase the availability of our programs to the outside community through resource sharing in the following ways: a) through a dedicated computer system which will exploit the proposed relationship of our work to the SUMEX-AIM computer resource; b) by continuing to develop exportable versions of our programs; and, c) by holding intensive workshops to introduce other research groups to our techniques.

LEAVE BLANK

BUDGET ESTIMATES FOR ALL YEARS OF SUPPORT REQUESTED FROM PUBLIC HEALTH SERVICE
DIRECT COSTS ONLY (Omit Cents)

DESCRIPTION		1ST PERIOD (SAME AS DETAILED BUDGET)	ADDITIONAL YEARS SUPPORT REQUESTED *(This application only)*					
			2ND YEAR	3RD YEAR	4TH YEAR	5TH YEAR	6TH YEAR	7TH YEAR
PERSONNEL COSTS		190,586	204,704	220,459	229,843	247,511		
CONSULTANT COSTS *(Include fees, travel, etc.)*								
EQUIPMENT		265,674	11,460	-	-	-		
SUPPLIES		2,600	2,782	2,977	3,185	3,408		
TRAVEL	DOMESTIC	1,500	1,500	1,500	1,500	1,500		
	FOREIGN							
PATIENT COSTS								
ALTERATIONS AND RENOVATIONS								
OTHER EXPENSES		51,040	37,836	40,483	43,320	46,351		
TOTAL DIRECT COSTS		511,400	258,282	265,419	277,848	298,770		
TOTAL FOR ENTIRE PROPOSED PROJECT PERIOD *(Enter on Page 1, Item 4)* ⟶						$ 1,611,719		

REMARKS: *Justify all costs for the first year for which the need may not be obvious. For future years, justify equipment costs, as well as any significant increases in any other category. If a recurring annual increase in personnel costs is requested, give percentage. (Use continuation page if needed.)*

Salary increases, and increases in other categories figured at 7% per year.

Staff benefits determined according to the following:

5/1/80	-	8/31/80	21.6%
9/1/80	-	8/31/81	22.4%
9/1/81	-	8/31/82	23.2%
9/1/82	-	8/31/83	24.0%
9/1/83	-	8/31/84	24.8%
9/1/84	-	4/30/85	25.6%

(see attached)

Budget Remarks

Personnel

Professor Djerassi will continue in his role as principal investigator, providing overall scientific direction to the project, Dr. Smith acting as co-investigator. Drs. Smith, Nourse, Gray and Buchanan will be responsible for the design, implementation and application of the programs which we propose to develop, including representing the scientific interface to the outside community of collaborators.

Two programmers, Mr. Terry and a person to be appointed, will provide scientific programming support for the project including the following responsibilities. Mr. Terry will have primary responsibility for maintenance and documentation of application programs, with the assistance of other members of the scientific and technical staff. He will share responsibility for design and programming of new algorithms with the scientific staff. Together with the second programmer, he will provide technical support for collaborators and be responsible for developing versions of CONGEN and newer programs for other computer systems. The second programmer, to be named, will have primary responsibility for the dedicated computer system (the DEC VAX) requested in this proposal, including interfaces to the SUMEX PDP-10 system, and all other necessary system support functions including obtaining and implementing new languages, text editors and so forth.

The pre- and postdoctoral fellows supported by the grant will be involved in applications of the programs to chemical problems in our own laboratories or in research involving development of new computer techniques in collaboration with the professional staff.

The secretarial responsibilities (105 percent of a full time person) are distributed among three persons, reflecting our best estimates of actual time spent on this grant and other grants, to myself or Prof. Feigenbaum, which support the remainder of their salaries (Ms. Learned-Driscoll is supported for the remainder of her time by the Chemistry Department).

Equipment

A Dedicated Computer System - the DEC VAX-11/780

We have requested funds in the first year of the proposal to purchase a dedicated computer system for our project. The reasons for this approach are discussed in detail in Section A.2.c, Relationship to the SUMEX-AIM Resource, and in Section C.5.a. Briefly, the implementation of this machine as an adjunct to the SUMEX resource is a research study in itself, namely, how best to provide access to computational resources in a computing environment where heavy demands are being made on both development and application of software tools for solving biomedical problems. We intend to make the dedicated system available for production use of our applications programs, assuring high quality interactive service to our collaborators. We cannot get such additional capacity for this work from SUMEX which is already overcrowded and committed to new program development rather than production use. We will exploit our close relationship with SUMEX for the development of new programs, for continued close contact with the AI community centered there, as a gateway to communication facilities allowing remote terminal

access and for sharing of peripheral equipment and operations support so far as is possible. At the end of this section we comment on our situation were the dedicated machine not approved as part of this proposal.

In considering the type of machine which could provide the needed, additional capacity, we have focussed our attention on the qualities we would require in a machine, letting these considerations guide our choice. We require high quality, interactive service to support our programs in a reasonable way. An interactive program must provide rapid response to the scientist using it. This means we need a system with a good time-sharing operating system. Our programs tend to be larger than can fit comfortably on mini-computers, and certainly the more complex programs which we propose to develop will only grow larger. This requirement suggests a system which possesses a large address space, large enough that space is not a limiting factor. The combination of time sharing and large address space suggests a virtual memory machine with demand paging. Finally, we want a machine that will maximize long-term compatibilities with the SUMEX AI community and other biochemical and chemical computing resources and laboratory systems.

We have considered several alternative computers which might meet the above qualities and would provide the needed computational power to support our community. We have rejected all 16-bit (or fewer) mini-computers on the basis of address space limitations which are too restrictive for the size of our programs. Although it is possible in principle to fit our new programs such as the exportable version of CONGEN into such a small machine using extensive overlays, the programming effort would be prohibitive.

One plausible alternative is to purchase a computer compatible with the existing DEC-10 or DEC-20 systems at SUMEX and Rutgers on which our new software such as CONGEN will run without change. There are only two reasonable alternatives here, the DEC-20 family or the Foonly F-2 (a DEC-10 compatible machine being developed by a Bay-area company). The higher end of the DEC line is prohibitively expensive (e.g., the 2040 or 2050), leaving the DEC-2020 as the only reasonable choice. These systems have the advantage of running the highly developed interactive monitors (TENEX or TOPS-20) now used at SUMEX and other AI resources. This would maximize near-term compatibility and minimize operational costs. The 2020 (and the F-2) are computationally quite slow, however, particularly for arithmetic processing since they do not have a floating point processor. We expect our new research will make significant demands in this area.

There are several other negative factors about selecting either the 2020 or the F-2. The current price-performance index of the DEC-2020 is quite poor; no price advantage per user is achieved over the much more complex (and expensive) DEC-20's. Foonly is a very small company whose future is insecure. Only a few laboratories to which we might export software and collaborate on program developments and applications operate DEC-10 or -20 compatible machines. Finally, this class of machines is rapidly becoming obsolete. DEC appears to be directing their future developments toward VAX-like machines.

In parallel with this, the ARPANET AI community has been investigating long-term alternatives for INTERLISP support given the address space limitations of the DEC-10. A number of alternatives are under consideration including a "personal" LISP machine being developed at MIT, a PRIME system being considered by BB and N (Bolt Beranek and Newman Inc.) and the DEC-VAX. It appears that a consensus is developing around VAX given its attractive architecture, good interactive operating systems (UNIX or perhaps VMS) and vendor support. The DEC-VAX also appears to be

increasingly popular in scientific laboratories, including labs of several of our workshop attendees (see Annual Report, Appendix I). As another example, the National Resource for Computation in Chemistry will soon take delivery of a VAX system, meaning that additional, chemistry related software will soon become available on the VAX. (We note that similarly configured VAX's and 2020's have similar prices; the VAX will deliver considerably more computational power, however).

In summary, purchase of a 2020 or similar machine would be an expedient, short-term option to pursue in terms of our compatibility with ourselves. We feel strongly, however, that this would be short sighted. We must look further into the future in our proposal to make the best estimate possible on the trends in computer developments followed by both manufacturers and users. We are already seeing shifts in DEC, the AI community and, most importantly, computational chemists with whom we interact, toward VAX and similar machines. It is becoming clear that VAX or its equivalent is the machine to obtain, especially as a medium for providing high quality, interactive service to our collaborators, whether by network access to Stanford or programs exported to other sites. We have, therefore, tentatively decided to purchase a VAX as the dedicated machine to support applications of our programs. The configuration of the VAX as we propose it would include:

1) Basic system, including 512 KByte memory, one RP06 disc drive, one TE16 tape drive, VMS operating system, LA120 operator console, DZ-11A for eight terminal lines ($185,000);

2) 1 MByte additional memory ($35,000);

3) Either 1 Mbyte additional memory ($35,000) or one additional RP06 disc drive ($34,000) ($35,000 figured in the total appearing on the budget)

4) FORTRAN compiler ($3,300);

5) Floating point accelerator ($9,900).

Item (2) is strongly recommended by DEC in order to support more than two or three concurrent users. The choice in item (3) remains to be made. The tradeoff here is whether or not support for a larger complement of users (8 - 12) is best handled by a separate disc system for paging, or by additional memory. Further discussions with other VAX users will be required to reach a final conclusion. The final budget figure ($254,214) was obtained by summing the above figures, taking an 11 percent educational discount and adding 6-1/2 percent California State sales tax.

We are currently evaluating the various operating systems available for the VAX, including VMS from DEC, UNIX from Bell Laboratories or the Interactive Systems Corp. UNIX version which couples both VMS and UNIX in one package. Choice of one of the non-DEC operating systems will add an additional increment to our budget. In addition, new software from DEC, such as new compilers, will represent additional costs. Given the price of the FORTRAN compiler, we can estimate other languages to cost about the same. We have requested funds in the budget to cover sortware purchases, with the greatest expense in year one.

We have also budgeted, as a one-time cost, our best estimates of the expense required to implement the SUMEX/VAX link (see Section C.5). This figure, $15,000, will cover communications interfaces, circuit boards, cabling and integrated circuits.

As a final note, the rapidly changing situation with respect to new machines and cheaper prices for existing machines may allow us to reduce these estimates in the coming months. We are making every effort to find more cost-effective ways to meet our basic goals for the dedicated computer.

We believe that a DEC maintenance contract to support the hardware is the most cost-effective way to ensure maximum computer availability for our outside collaborators. Based on the hardware cited above and DEC's standard contract prices, the cost for year one (nine-month basis, three months covered under warranty) is $13,450, including a maintenance contract on the existing GT40.

If this proposal is approved and funded without the proposed computer, we will be forced to provide access to our programs, for both development and applications, to SUMEX alone, i.e., perpetuate the current situation. Not only will an important experiment in resource access and resource sharing be lost, but our collaborators (see Section F) will become increasingly impatient at the slow response time of SUMEX. We know from past experience that if response time in a highly interactive program becomes too long, scientists will simply find more productive uses for their time and will cease to use the computer. It also must be recognized that, because SUMEX is funded separately from our grant, we must also try to ensure long-term computational support for our research. The machine configuration requested is, alone, insufficient to meet our needs in the absence of SUMEX, but would do so with augmentation (additional peripheral equipment, memory, network interface, additional terminal ports and perhaps additional personnel).

Graphics Terminals

The goals of our proposed research are heavily involved with stereochemical representations of molecular structure. The requirements for visualizing these structures plus our desires to seek improvements in the interactive capabilities of our programs (most of our collaborators have requested some form of graphic input and output) have led us to propose a low-cost alternative to expensive graphics systems (Evans and Sutherland, Vector-General) or even expensive graphics terminals like the DEC GT-40. We are currently examining alternative display systems which represent new technology and provide graphics capabilities together in some instances with capabilities for input of graphic (in our case structural) information.

We have examined several display systems so far, including those manufactured by Grinnel Systems (Santa Clara, CA), DeAnza Systems (Santa Clara, CA) and Tektronix. At this time, the Tektronix 4025 appears to have a better set of desirable features, including capacity for pseudo-rotation of structures based on sequential display of stored images, programmable functions for special graphics such as automatic construction of ring systems, and cursor control for structure input. The budget figure requested includes the basic terminal plus additional display and graphics memory (to a total of 32K and 64K bytes, respectively) to allow storage of sufficient multiple images for pseudo-rotation (to add three-dimensional characteristics to the visualized structures). We will continue to search for other low-cost graphics terminal which would meet our needs together with the needs of our collaborators.

A second terminal has been requested in the second year of the budget. By that time we should be heavily involved in representation and manipulation of three-dimensional representations of structure and would require additional graphics terminal support for program development and application.

SECTION II – PRIVILEGED COMMUNICATION

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	TITLE	BIRTHDATE *(Mo., Day, Yr.)*
Carl Djerassi	Professor of Chemistry	10/29/23
PLACE OF BIRTH *(City, State, Country)*	**PRESENT NATIONALITY** *(If non-U.S. citizen, indicate kind of visa and expiration date)*	**SEX**
Vienna, Austria	U. S. Citizen	☒ Male ☐ 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 Foundation Patent Award ('71) and Chem. Pioneer Award ('73) of Am. Inst.Chem.; Hon. Member and Centenary Lecturer, Chem. Soc. (London); Member of National Academy of Sciences, American

MAJOR RESEARCH INTEREST	ROLE IN PROPOSED PROJECT

RESEARCH SUPPORT *(See instructions)* See page 10

HONORS (continued)

Academy of Arts and Sciences, Royal Swedish Academy of Sciences, Leopoldina, Bulgarian Academy of Sciences. Honorary D. Sc. Kenyon College, Nat. Univ. of Mexico, Federal Univ. of Rio de Janeiro, Worcester Polytechnic Inst., Wayne State, Columbia, Uppsala, Coe College, University of Geneva. Wolf Prize in Chemistry (1978), National Inventors Hall of Fame (1978).

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

Miscellaneous

Chairman of AAAS Gordon Res. Conf. on Steroids and Nat. Prod. ('52-'54); Member Amer. Pugwash Comm. ('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 Streng ('75-76); Comm. on National Medal of Science; NAS Institute of Medicine (Member of Membership Comm. ar Comm. on International Health).

Publications

Author or co-author of six books (four dealing with organic mass spectrometry(and over 930 scientific publications. The most recent publications (since January 1978) are listed here, with papers Nos. 927, 925, 924, 915, 907 and 906 being particularly relevant to this application.

☆U.S. GOVERNMENT PRINTING OFFICE: 1977—241-161:3024

RESEARCH SUPPORT: CARL DJERASSI

Since essentially all of my research is supported by the NIH, it is important for the reviewers to understand what each grant actually covers in terms of personnel, equipment, etc., and how the present renewal application of RR-00612 fits into this overall picture. Therefore, I am going into somewhat more detail than is generally required for this section.

GM-06840 (20-23): "Marine Chemistry with Special Emphasis on Steroids" 5/1/78 - 4/30/82. Current annual budget $133,042. This grant supports the bulk (equivalent to three predoctorate and five postdoctorate fellows) of my collaborators who are working on the isolation, structure elucidation, biosynthesis and possible biological function of marine steroids.

GM-20276 (05): "Magnetic Circular Dichroism of Cytochromes P-450" 7/1/78 - 6/30/79. Current budget $38,201. Aside from equipment maintenance and liquid helium, this grant supports partial salaries of two postdoctorate research associates. A renewal application, entitled "MCD of Porphyrins and Artificial Heme Proteins" has been submitted.

RR-00612 (09-10): "Resource Related Research - Computers in Chemistry" 5/1/78 - 4/30/80. Current budget $144,051 (5/1/79 - 11/30/79). The current application is for renewal of this project. The renewal represents a significant change in emphasis from the current proposal in that the SUMEX-AIM computer resource is the resource to which our research will be related in the future, rather than the mass spectrometry laboratory. For this reason, the following two (partially overlapping) applications were submitted (and are pending) for support of personnel and supplies and purchase of new equipment to maintain and upgrade the mass spectrometry laboratory. Because of this shift in emphasis, the current proposal requests no funds for support of the laboratory and does not overlap in any way with the following two proposals.

AM-04257: "Mass Spectrometry in Organic and Biochemistry" 12/1/76 - 11/30/79. Current annual budget $74,800. This grant currently supports one mass spectrometer, a senior technician to operate the instrument, one postdoctoral fellow and three predoctoral fellows, who are engaged in isolation and synthetic work related to interpretation of fragmentation patterns of known and unknown molecular structures.

A renewal application has been submitted for this grant and is currently pending. This renewal is designed to support the mass spectrometer, personnel and supplies currently supported under RR-00612. In addition funds were requested to upgrade the mass spectrometer facilities. No action has as yet been taken on this application.

Pending Application: A NIGMS Shared Instrumentation Resource proposal entitled "A Shared Mass Spectrometry Resource" has been submitted for the period 9/1/79 to 8/30/82. The first year budget is $145,046, including $130,850 of capital equipment. This proposal requests support for equipment similar to that requested in the renewal of AM-04257, with some differences reflecting the resource sharing aspects of the NIGMS program. The only personnel for whom support is requested under this proposal include 10 percent of Dr. Smith and 10 percent of a senior technicial. No action has yet been taken on this application. If AM-04257 is approved then only $18,500 is requested from NIGMS for a chemical ionization accessory.

RESEARCH SUPPORT: CARL DJERASSI (continued)

Pending Application: An application entitled "Circular Dichroism of Cyclic Ketones - Conformational Isotope Effects" with an annual budget of $55,782 for the period 4/1/79 to 3/31/81 has been submitted to the National Science Foundation. In terms of personnel, it covers the salaries of one postdoctorate and two predoctorate fellows for synthetic work on chiral ketones, whose chirality is only due to ^{13}C or deuterium. No action has as yet been taken on this application.

RECENT PUBLICATIONS

900. Magnetic Circular Dichroism Studies LII. Magnetic Circular Dichroism of Purified Forms of Rabbit Liver Cytochromes P-450 and P-420. Biochemistry 17, 33 (1978), by J. H. Dawson, J. R. Trudell, R. E. Linder, G. Barth, E. Bunnenberg and C. Djerassi.

901. Optical Rotatory Dispersion Studies CXXII. Synthesis and Circular Dichroism of (3S)-Deuteriocyclohexanone. Tetrahedron Letters, 535 (1978), by C. Djerassi, C. L. VanAntwerp and P. Sundararaman.

902. Hommes Affamés, Parasites Affamés. Prospective et Santé, No. 4, 87-91 (1978), by Carl Djerassi.

903. Marine Natural Products. Synthesis of Four Naturally Occurring 20β-H Cholanic Acid Derivatives. J. Org. Chem. 43, 1442 (1978), by D. J. Vanderah and C. Djerassi.

904. One Step Conversion of Aldehydes to Esters. Tetrahedron Letters, 1627 (1978), by P. Sundararaman, E. C. Walker and C. Djerassi.

905. Terpenoids LXXV. $\Delta^{9(12)}$-Capnellene, A New Sesquiterpene Hydrocarbon from the Soft Coral Capnella Imbricata. Tetrahedron Letters, 1671 (1978), by E. Ayanoglu, T. Gebreyesus, C. M. Beechan, C. Djerassi and M. Kaisin

906. Applications of Artificial Intelligence for Chemical Inference XXVII. Computer-Assisted Structure Manipulation. Studies in the Biosynthesis of Natural Products. Tetrahedron 34, 841 (1978), by T. H. Varkony, D. H Smith and C. Djerassi

907. Recent Advances in the Mass Spectrometry of Steroids. Pure and Appl. Chem. 50, 171 (1978), by C. Djerassi.

908. Optical Rotatory Dispersion Studies CXXIII. Experimental Evidence for Preference of Axial Deuterium over Axial Hydrogen.
J. Amer. Chem. Soc. 100, 3965 (1978)
by S.-F. Lee, G. Barth, K. Kieslich and C. Djerassi

909. Optical Rotatory Dispersion Studies CXXIV. Synthesis and Circular Dichroism of 3(S)^a- and 3(R)^e-Deuterio-4(R)-t-Butylcyclohexanone and 2(R)^a- and 2(S)^e-Deuterio-4(R)-Isopropylcyclohexanone.
Tetrahedron Letters, 2457 (1978)
by P. Sundararaman and C. Djerassi

910. Minor and Trace Sterols in Marine Invertebrates IV. Identification of Sterols with Short Side Chains in the Sponge Damiriana hawaiiana.
Helv. Chim. Acta 61, 1470 (1978)
by C. Delseth, R.M.K. Carlson, C. Djerassi, T. Erdman and P.J. Scheuer

911. "Coping with Interdependence: The Population Problem", in Proceedings of the Symposium on Energy, Food, Population and World Interdependence, Committee on Chemistry and Public Affairs, American Chemical Society, Washington, D.C., 1978, pp. 40-43; 53-55. By Carl Djerassi.

912. Minor and Trace Sterols in Marine Invertebrates V. Isolation, Structure Elucidation and Synthesis of 3β-Hydroxy-26,27-Bisnorcholest-5-en-24-one from the Sponge Psammaplysilla purpurea.
Steroids 31, 815 (1978)
by E. Ayanoglu, C. Djerassi, T. R. Erdman and P.J. Scheuer

913. Terpenoids LXXIV. The Sesquiterpenes from the Soft Coral Sinularia mayi.
Tetrahedron 34, 2503 (1978)
by C. M. Beechan, C. Djerassi and H. Eggert

914. 24-Ethyl-$\Delta^{5,24(28),28}$-cholestatrien-3β-ol - a Naturally Occurring Allenic Marine Sterol.
J. Amer.Chem. Soc. 100, 5574 (1978)
by N. Theobald, J.N. Shoolery, C. Djerassi, T.R. Erdman and P.J. Scheuer

915. Applications of Artificial Intelligence for Chemical Inference XXVIII. Computer-Assisted Simulation of Chemical Reaction Sequences. Applications to Problems of Structure Elucidation.
J. Chem. Inf. Comput. Sci. 18, 168 (1978)
by T.H. Varkony, R.E. Carhart, D.H. Smith and C. Djerassi

916. Birth Control after 1984 Revisited.
Bull. Amer. Acad. of Arts and Sciences 32, No. 1, (1978)
by C. Djerassi

917. Determination of the Absolute Configuration of Stelliferasterol and Strongylosterol - Two Marine Sterols with "Extended" Side Chains.
Tetrahedron Letters, 4369 (1978)
by N. Theobald and C. Djerassi

918. Minor and Trace Sterols in Marine Invertebrates IX. *Verongulasterol - a Marine Sterol with a Novel Side Chain Alkylation Pattern.*
Tetrahedron Letters, 4373 (1978)
by W.C.M.C. Kokke, W. H. Fenical, C. S. Pak and C. Djerassi

919. Optical Rotatory Dispersion Studies CXXVI. Synthesis and Chiroptical Properties of Cyclohexanones with Chirality Solely Due to Isotopic Substitution: $^{12}CH_3$ vs $^{13}CH_3$ and CH_3 vs CD_3
Tetrahedron Letters, 4377 (1978)
by C. S. Pak and C. Djerassi

920. Isolation and Structure of 26,27-Cycloaplysterol (Petrosterol): A Cyclopropane-Containing Marine Sterol.
Tetrahedron Letters, 4379 (1978)
by B. N. Ravi, W.C.M.C. Kokke, C. Delseth and C. Djerassi

921. Minor and Trace Sterols in Marine Invertebrates VIII. Isolation, Structure Elucidation and Partial Synthesis of Two Novel C_{30} Marine Sterols - Stelliferasterol and Isostelliferasterol.
J. Amer. Chem. Soc., 100, 7677 (1978)
by N. Theobald, R.J. Wells and C. Djerassi

922. Optical Rotatory Dispersion Studies CXXV. Independent Evidence for Preference of Axial Deuterium vs. Axial Hydrogen through Variable Temperature Circular Dichroism Spectra of (4S)-2,2-dimethyl-4-deuteriocyclohexanone and (3S)-3-deuterio-4,4-dimethylcyclohexanone.
J. Amer. Chem. Soc., 100, 8010 (1978)
by S.F. Lee, G. Barth and C. Djerassi

923. Partial Reduction of Aquomethemoglobin on a Sephadex G-25 Column as Detected by Magnetic Circular Dichroism Spectroscopy and Revised Extinction Coefficients for Aquomethemoglobin.
Analytical Biochemistry, 90, 474 (1978)
by R. E. Linder, R. Records, G. Barth, E. Bunnenberg, C. Djerassi, B. E. Hedlund, A. Rosenberg, E. S. Benson, L. Seamans and A. Moscowitz

924. Minor and Trace Sterols in Marine Invertebrates VI. Occurrence and Possible Origins of Sterols Possessing Unusually Short Hydrocarbon Side Chains
Bioorganic Chemistry, 7, 453 (1978)
by R.M.K. Carlson, S. Popov, I. Massey, C. Delseth, E. Ayanoglu, T.H. Varkony and C. Djerassi

925. A Novel Role of Computers in the Natural Products Field.
Naturwissenschaften, 66, 9 (1979)
by C. Djerassi, D. H. Smith and T. H. Varkony

926. The Synthesis of Demethylgorgosterol.
Tetrahedron Letters, 767 (1979)
by R. D. Walkup, G. D. Anderson and C. Djerassi

927. Applications of Artificial Intelligence for Chemical Inference XXIX. Exhaustive Generation of Stereoisomers for Structure Elucidation.
J. Amer. Chem. Soc., 101, 1216 (1979)
by J. G. Nourse, R. E. Carhart, D. H. Smith and C. Djerassi

928. Minor and Trace Sterols in Marine Invertebrates XI. 5α-24-Norcholestan-3β-ol and (24Z)-Stigmasta-5,7,24(28)-trien-3β-ol, Two New Marine Sterols from the Pacific Sponges Terpios Zeteki and Dysidea Herbacea.
Helv. Chim. Acta, 62, 101 (1979)
by C. Delseth, L. Tolela, P.J. Scheuer, R.J. Wells and C. Djerassi

SECTION II – PRIVILEGED COMMUNICATION

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	TITLE	BIRTHDATE (Mo., Day, Yr.)
Dennis H. Smith	Research Associate	11-12-42

PLACE OF BIRTH (City, State, Country)	PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date)	SEX
New York	USA	☒ Male ☐ Female

EDUCATION *(Begin with baccalaureate training and include postdoctoral)*

INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Massachusetts Institute of Technology Cambridge, MA	S.B.	1964	Chemistry
University of California Berkeley, CA	Ph.D.	1967	Chemistry

HONORS

Alfred P. Sloan Foundation Scholarship, NASA Predoctoral Traineeship, Phi Lambda Upsilon, Sigma Xi, Editorial Board of Journal of Chemical Information and Computer Science.

MAJOR RESEARCH INTEREST	ROLE IN PROPOSED PROJECT
Mass Spectrometry and Computer Applications in Chemistry	Senior 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 California at Berkeley, Berkeley, CA
1965-1967	NASA Pre-Doctoral Traineeship, University of California at Berkeley, Berkeley, CA

Publications: See attached list.

(32) 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," J. Chem. Inf. Comp. Sci., 15, 124 (1975); Erratum, Ibid., 16, 125 (1976).

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

(34) 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," J. Amer. Chem. Soc., 97, 5755 (1975).

(35) D.H. Smith, "The Scope of Structural Isomerism," J. Chem. Inf. Comp. Sci., 15, 203 (1975).

(36) D.H. Smith, J.P. Konopelski, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIX. Computer Generation of Ion Structures," Org. Mass Spectrom., 11, 86 (1976).

(37) 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).

(38) 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, 1, 79 (1976).

(39) C. Cheer, D.H. Smith, C. Djerassi, B. Tursch, J.C. Braekman, and D. Daloze, "Applications of Artifical 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, 32, 1807 (1976).

(40) 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," J. Amer. Chem. Soc., 98, 6168 (1976).

(41) D.H. Smith and R. E. Carhart, "Structural Isomerism of Mono- and Sesquiterpenoid Skeletons," Tetrahedron, 32, 2513 (1976).

(42) 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., 11, 1120 (1976).

(43) T.H. 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 and J. Howe, Eds., American Chemical Society, Washington, D.C., 1977, p. 188.

(44) "Computer-Assisted Structure Elucidation," D.H. Smith, Ed., American Chemical Society, Washington, D.C., 1977.

(45) R.E. Carhart, T.H. Varkony, and D.H. Smith, "Computer Assistance for the Structural Chemist," in "Computer-Assisted Structure Elucidation," D.H. Smith, Ed., American Chemical Society, Washington, D.C., 1977, p. 126.

(46) D.H. Smith, M. Achenbach, W.J. Yeager, P.J. Anderson, W.L. Fitch, and T.C. Rindfleisch, "Quantitative Comparison of Combined Gas Chromatographic/Mass Spectrometric Profiles of Complex Mixtures," Anal. Chem., 49, 1623 (1977).

(47) B.G. Buchanan and D.H. Smith, "Computer Assisted Chemical Reasoning," in "Computers in Chemical Education and Research," E.V. Ludena, N.H. Sabelli, and A.C. Wahl, Eds., Plenum Press, New York, N.Y., 1977, p. 401.

(48) D.H. Smith and R.E. Carhart, "Structure Elucidation Based on Computer Analysis of High and Low Resolution Mass Spectral Data," in "High Performance Mass Spectrometry: Chemical Aplications," M.L. Gross, Ed., American Chemical Society, 1978, p. 325.

(49) T.H. Varkony, D.H. Smith, and C. Djerassi, "Computer-Assisted Structure Manipulation: Studies in the Biosynthesis of Natural Products," Tetrahedron, 34, 841 (1978).

(50) D.H. Smith and P.C. Jurs, "Prediction of 13C NMR Chemical Shifts," J. Am. Chem. Soc., 100, 3316 (1978).

(51) T.H. Varkony, R.E. Carhart, D.H. Smith, and C. Djerassi, "Computer-Assisted Simulation of Chemical Reaction Sequences. Applications to Problems of Structure Elucidation," J. Chem. Inf. Comp. Sci., 18, 168 (1978).

(52) D.H. Smith, T.C. Rindfleisch, and W.J. Yeager, "Exchange of Comments: Analysis of Complex Volatile Mixtures by a Combined Gas Chromatography-Mass Spectrometry System," Anal. Chem., 50, 1585 (1978).

(53) W.L. Fitch, P.J. Anderson, and D.H. Smith, "Isolation, Identification and Quantitation of Urinary Organic Acids," J. Chrom., 162, 249 (1979).

(54) W.L. Fitch, E.T. Everhart, and D.H. Smith, "Characterization of Carbon Black Adsorbates and Artifacts Formed During Extraction," Anal. Chem., 50, 2122 (1978).

(55) W.L. Fitch and D.H. Smith, "Analysis of Adsorption Properties and Adsorbed Species on Commercial Polymeric Carbons," Environ. Sci. Tech., 13, 341 (1979).

(56) J.G. Nourse, R.E. Carhart, D.H. Smith, and C. Djerassi, "Exhaustive Generation of Stereoisomers for Structure Elucidation," J. Am. Chem. Soc., 101, 1216 (1979).

(57) C. Djerassi, D.H. Smith, and T.H. Varkony, "A Novel Role of Computers in the Natural Products Field," Naturwiss., 66, 9 (1979).

(58) N.A.B. Gray, D.H. Smith, T.H. Varkony, R.E. Carhart, and B.G. Buchanan, "Use of a Computer to Identify Unknown Compounds. The Automation of Scientific Inference," Chapter 7 in "Biomedical Applications of Mass Spectrometry," G.R. Waller, Ed., in press.

(59) T.C. Rindfleisch and D.H. Smith, in Chapter 3 of "Biomedical Applications of Mass Spectrometry," G.R. Waller, Ed., in press.

(60) T.H. Varkony, Y. Shiloach, and D.H. Smith, "Computer-Assisted Examination of Chemical Compounds for Structural Similarities," J. Chem. Inf. Comp. Sci., 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	TITLE	BIRTHDATE (Mo., Day, Yr.)
Buchanan, Bruce G.	Adjunct Professor of Computer Science	7/7/40
PLACE OF BIRTH (City, State, Country)	**PRESENT NATIONALITY (If non-U.S. citizen, indicate kind of visa and expiration date)**	**SEX**
St. Louis, Missouri	U.S.	XX Male ☐ Female

EDUCATION *(Begin with baccalaureate training and include postdoctoral)*

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

HONORS
Recipient of National Institutes of Health Career Development Award (1971-1976).
continued on attached sheet

MAJOR RESEARCH INTEREST	ROLE IN PROPOSED PROJECT
Artificial Intelligence	Principal Investigator

RESEARCH SUPPORT *(See instructions)*

NIH 5R24 RR 00612-09 Resource Related Research: Computers and Chemistry 5/1/77 to 9/30/79 Time presently committed: 20%

Dept. of Defense MDA 903-77-C-0322; Heuristic Programming 8/1/77 to 9/30/79. Time presently committed: 45%

National Science Foundation MCS 7702712 Knowledge-Based Intelligent Systems 6/1/77 to 5/30/79. Time presently committed: 10%

National Science Foundation MCS 78-02777; MOLGEN: A Computer Science Application to Molecular Genetics . Time presently committed: 25%

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 - present	Adjunct Professor; Computer Science Department, Stanford University, Stanford, CA 94305.
1972-1976	Research Computer Scientist, Computer Science Department, Stanford University.
1966-1971	Research Associate, Artificial Intelligence Project, Stanford University.

Selected Publications - See attached

*U.S. GOVERNMENT PRINTING OFFICE: 1977—241-161:3024

Recent Honors

Editorial Board, Artificial Intelligence: An International Journal.
Chairman of IJCAI-79 Program Committee (International Joint Conference on Artificial Intelligence, Tokyo, 1979).
Invited Speaker, Workshop on The Logic of Discovery & Diagnosis in Medicine (Pittsburgh, October 1978).
Invited Speaker, Douglass College Seminars for Faculty, (Rutgers University, 1978).

Selected Publications

Bruce G. Buchanan, G.L. Sutherland, E.A. Feigenbaum, "Toward an Understanding of Information Processes of Scientific Inference in the Context of Organic Chemistry." in B. Meltzer and D. Michie (eds.), Machine Intelligence, 5, Edinburgh: Edinburgh University Press, 1970.

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 Am. Chem. Society, 91, 2973, 1969.

C.W. Churchman and B.G. Buchanan, "On the Design of the Inductive Systems: Some Philosophical Problems," British Journal for the Philosophy of Science, 20, 311, 1969.

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

D. Michie and B.G. Buchanan, "The Scientist's Apprentice," in R.A.G. Carrington (ed.), Computers for Spectroscopy, London: Adam Hilger, 1974.

E.H. Shortliffe and B.G. Buchanan, "A Model of Inexact Reasoning in Medicine," Mathematical Biosciences, 23, 351,1975.

Selected Publications (continued)

E.H. Shortliffe, R. Davis, S.C. Axline, B.G. Buchanan, C.C. Green, S.N. Cohen, "Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System," Computers and Biomedical Research, 8, 303, 1975.

Randall Davis, Bruce Buchanan, Edward Shortliffe, "Production Rules as a Representation of a Knowledge-Based Consultation Program," Artificial Intelligence, 8,1,1977.

B.G. Buchanan, D.H. Smith, W.C. White, R.J. Gritter, E.A. Feigenbaum, J. Lederberg, and Carl Djerassi, "Application 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, 98, 6168, 1976.

B.G. Buchanan, R. Davis, V. Yu and S. Cohen, "Rule Based Medical Decision Making by Computer," Proceedings of MEDINFO77, Toronto, 1977.

Bruce G. Buchanan and Tom Mitchell, "Model-Directed Learning of Production Rules," in D.A. Waterman and F. Hayes-Roth (eds.), Pattern Directed Inference Systems, New York: Academic Press, 1978.

Randall Davis and Bruce G. Buchanan, "Meta-Level Knowledge: Overview and Applications," Proceedings of the Fifth IJCAI,1,920, August 1977.

Victor L. Yu, Bruce G. Buchanan, Edward H. Shortliffe, Sharon M. Wraith, Randall Davis, A. Carlisle Scott, Stanley N. Cohen, "Evaluating the Performance of a Computer-Based Consultant." Forthcoming, Computer Programs in Biomedicine, Amsterdam.

Bruce G. Buchanan, Tom M. Mitchell, Reid G. Smith and C. Richard Johnson, Dr., "Models of Learning Systems," in J. Belzer (ed.), Encyclopedia of Computer Sciences and Technology, New York: Marcel Dekker, Inc., 1978, Vol 11.

Bruce G. Buchanan and Edward A. Feigenbaum, "DENDRAL and Meta-DENDRAL: Their Applications Dimension," to appear in Artificial Intelligence, Fall 1978.

SECTION II – PRIVILEGED COMMUNICATION

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	TITLE	BIRTHDATE *(Mo., Day, Yr.)*
James G. Nourse	Research Associate	December 14, 1947

PLACE OF BIRTH *(City, State, Country)*	PRESENT NATIONALITY *(If non-U.S. citizen, indicate kind of visa and expiration date)*	SEX
Buffalo, NY	U.S. citizen	☒ Male ☐ Female

EDUCATION *(Begin with baccalaureate training and include postdoctoral)*

INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Columbia University, New York, NY	B.S.	1969	Chemical Engineering
California Inst. of Tech., Pasadena, CA	Ph.D.	1974	Organic Chemistry
UCLA, Los Angeles, CA	Post Doc	1973-74	Organic Chemistry
Princeton University, Princeton, NJ	Post Doc	1974-76	Organic Chemistry

HONORS

Sigma Xi, Invited Lecturer 1978 Conference on the Permutation and its Application to Chemistry and Physics, Bielefeld, Germany

MAJOR RESEARCH INTEREST	ROLE IN PROPOSED PROJECT
Stereochemistry, Computer Applications	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.)*

Research Associate, Department of Chemistry, Stanford University, 5/79 -
Research Affiliate, Department of Computer Science, Stanford University, 8/76 to 4/79
Postdoctoral Research Associate, Department of Chemistry, Princeton University, 2/74 to 7/76
Postdoctoral Research Associate, Department of Chemistry, UCLA, 10/73 to 1/74

☆U.S. GOVERNMENT PRINTING OFFICE 1977—241-161:3024

Publications- James G. Nourse

M. G. Hutchings, J. G. Nourse, and K. Mislow, "A First Approach to the Stereochemical Analysis of Tetraarylmethanes", Tetrahedron, 30, 1525, 1974.

J. G. Nourse, "An Algebraic Description of Stereochemical Correspondence", Proceedings of the National Academy of Sciences, 72, 2285, 1975.

J. G. Nourse and K. Mislow, "Dynamic Stereochemistry of Tetraarylmethanes and Cognate Systems. The Role of the Permutation Subgroup Lattice", Journal of the American Chemical Society, 97, 4571, 1975.

J. G. Nourse and J. D. Roberts, "Nuclear Magnetic Resonance Spectroscopy. Carbon-13 Spectra of Some Macrolide Antibiotics and Derivatives. Substituent and Conformational Effects", Journal of the American Chemical Society, 97, 4584, 1975.

J. G. Nourse, "Pseudochirality", Journal of the American Chemical Society, 97, 4594, 1975.

J. G. Nourse, "Generalized Stereoisomerization Modes", Journal of the American Chemical Society, 99, 2063, 1977.

J. G. Nourse, "The Configuration Symmetry Group and its Application to Stereoisomer Generation, Specification, and Enumeration", Journal of the American Chemical Society, 101, 1210, 1979.

J. G. Nourse, P. E. Carhart, D. H. Smith, and C. Djerassi, "Exhaustive Generation of Stereoisomers for Structure Elucidation", Journal of the American Chemical Society, 101, 1216, 1979.

J. G. Nourse, "Application of the Permutation Group in Dynamic Stereochemistry" to appear in the Proceedings of the Conference on the Permutation Group and its Application to Chemistry and Physics, held at Bielefeld, Germany, 9-12 July 1978.

J. G. Nourse, "Application of the Permutation Group to Stereoisomer Generation for Computer Assisted Structure Elucidation", to appear in the Proceedings of the Conference on the Permutation Group and its Application to Chemistry and Physics, held at Bielefled, Germany, 9-12 July, 1978.

SECTION II – PRIVILEGED COMMUNICATION

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	TITLE	BIRTHDATE *(Mo., Day, Yr.)*
Neil A. B. Gray	Research Associate	September 5, 1947

PLACE OF BIRTH *(City, State, Country)*	PRESENT NATIONALITY *(If non-U.S. citizen, indicate kind of visa and expiration date)*	SEX
Edinburgh, Scotland, UK	British (currently on J-1 visa, applying for H-1)	☒ Male ☐ Female

EDUCATION *(Begin with baccalaureate training and include postdoctoral)*

INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	SCIENTIFIC FIELD
Imperial College, London	B. S.	1968	Chemistry
University of Cambridge	M. S.	1970	Theoretical Chemistry
	M. S.	1971	Computer Science
	Ph.D.	1977	Computer Science

HONORS

MAJOR RESEARCH INTEREST	ROLE IN PROPOSED PROJECT
Applications of Computers to Chemistry	

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.)*

Science Research Council/NATO research fellow. Development of programs for structure generation and analysis including new approaches to the processing of high and low resolution mass spectral data and to handling magnetic resonance data. Stanford University, 1977-1979.

Junior Research Fellowship at King's College. Work on various computer-assisted approaches to chemical structure elucidation including both rule-based and pattern-recognition methods. Part of this work being incorporated into a Ph.D. thesis for the Computer Science Department. University of Cambridge, 1973-1977.

Research Assistant in Professor Eglinton's Organic Geochemistry Unit in the School of Chemistry on an OSTI-grant (Office for Scientific and Technical Information). Design and implementation of file-search and interpretive schemes for processing low-resolution mass spectral data, work on data-communications software and creation of small, dedicated data-acquisition system on PDP-8 mini-computers. University of Bristol, 1971-1973.

M.Sc. thesis on limitations of CNDO-type semi-empirical quantum mechanical calculations. University of Cambridge, 1968-1970.

Publications:

N.A.B. Gray and A.J. Stone.
"Justifiability of the ZDO Approximation in Terms of a Power Series Expansion."
Theoret. Chim. Acta, 18, 389 (1970).

D. Field, N.A.B.Gray and P.F. Knewstubb.
"Computational Study of the Reactions between CH4 and CH4+."
J.Chem.Soc., Faraday II, 68, 852 (1972).

D.H. Smith, N.A.B. Gray, C.T. Pillinger, B.J. Kimble and G. Eglinton.
"Geochemical and environmental applications of a compound classifier based on computer analysis of low resolution mass spectra."
Advances in Organic Geochemisty 1971.
H.R.V. Gaertner and H. Wehner (Eds).
Pergammon (1972).

T.O. Gronneberg, N.A.B. Gray and G. Eglinton.
"Computer based search and retrieval system for rapid mass spectral screening of samples."
Anal. Chem., 47, 415 (1975).

N.A.B. Gray and T.O. Gronneberg.
"Programs for spectrum classification and screening of gas-chromatographic/mass-spectrometric data on a laboratory mini-computer."
Anal.Chem., 47, 419 (1975).

N.A.B. Gray.
"A program for generating empirical spectrum classification schemes."
Org. Mass Spectrom., 10, 507 (1975).

N.A.B. Gray, J.A. Zoro, T.O. Gronneberg, S.J. Gaskell, J.N. Cardoso and G. Eglinton.
"Automatic classification of mass spectra by a laboratory computer system."
Analyt. Letters, 8, 461 (1975).

N.A.B. Gray.
"Structural Interpretation of spectra."
Anal. Chem., 47, 2426 (1975).

N.A.B. Gray.
"Similarity measures for binary coded mass spectral data."
Anal. Chem., 48, 1420 (1976).

N.A.B. Gray.
"Constraints on learning machine classification methods."
Anal. Chem., 48, 2426 (1976).

N.A.B. Gray, D.H. Smith, T.H. Varkony, R.E. Carhart
and B.G. Buchanan.
"Use of a Computer to Identify Unknown Compounds:
The Automation of Scientific Inference."
Biochemical Applications of Mass Spectrometry.
G.R. Waller (Ed).
Interscience (1979 ?).

Table of Contents

Section / Subsection			Page
List of Figures			27
A.	INTRODUCTION.		28
	A.1	Objective.	28
	A.2	Background.	29
	A.3	Rationale.	37
B.	SPECIFIC AIMS.		40
C.	METHODS OF PROCEDURE.		42
	C.1	SASES — A Semi-Automatic Structure Elucidation System.	42
	C.2	The GENOA Program: Structure Generation with Overlapping Atoms	44
	C.3	Development of Automated Approaches to the Exploitation of Spectroscopic Data.	51
	C.4	Conformation Generator for CONGEN.	62
	C.5	Resource Sharing.	70
D.	SIGNIFICANCE		77
E.	FACILITIES AVAILABLE.		79
F.	COLLABORATIVE ARRANGEMENTS.		81
G.	PRINCIPAL INVESTIGATOR ASSURANCE		87
H.	REFERENCES.		88

List of Figures

1. Block Diagram of the SASES system. 43

2. Current Status of GENOA and its Interface to the CONGEN Program. 45

3. Illustration of the Use of GENOA in the Step-By-Step Specification of Overlapping Structural Information for the Structure of Palustrol [17A]. . . . 48

4. Meta-DENDRAL C-13 Spectrum Interpretation Rule 52

5. Isoterpinolene —— example structure for C-13 interpretation functions . . 53

6. Spectral data input for Isoterpinolene 54

7. Adjacency matrix derived for Isoterpinolene using Alpha-neighbors . . . 55

8. Adjacency matrix derived for Isoterpinolene using Beta-neighbors 56

9. Proposed Link of CONGEN to existing coordinate-based methods 63

10. Access to DENDRAL programs on SUMEX 71

11. SUMEX hardware 80

Djerassi, Carl
[redacted]

RESEARCH PLAN

A INTRODUCTION.

A.1 Objective.

The overall objective of our research is to develop and apply computational techniques to the procedures of structural analysis of known and unknown organic compounds based on structural information obtained from physical and chemical methods and to place these techniques in the hands of a wide community of collaborators to help them solve questions of structure of important biomolecules. These techniques are embodied in interactive computer programs which place structural analysis under the complete control of the scientist working on his or her own structural problem. Thus, we stress the word assisted in computer-assisted structure elucidation or analysis.

Our principal objective in this proposal is to extend our existing techniques for computer assistance in the representation and manipulation of chemical structures along two complementary, interdigitated lines. We propose to put together a comprehensive, interactive system to assist scientists in all phases of structural analysis (SASES, or Semi-Automated Structure Elucidation System) from data interpretation through structure generation to data prediction. This system will act as a computer-based laboratory in which complex structural questions can be posed and answered quickly, thereby conserving time and sample. In a complementary development we will extend our techniques from the current emphasis on topological, or constitutional, representations of structure to detailed treatment of conformational and configurational stereochemical aspects of structure.

We will make our programs available to a community of collaborators through network access to a dedicated machine requested in this proposal, exportable versions of our software and workshops held to introduce others to our computational methods.

By meeting our objectives we will fill in the "missing link" in computer assistance in structural analysis. Our capabilities for structural analysis based on the three-dimensional nature of molecules is an absolute necessity for relating structural characteristics of molecules to their observed biological, chemical or spectroscopic behavior. These capabilities will represent a quantum leap beyond our current techniques and open new vistas in applications of our programs, both of which will attract new applications among a broad community of structural chemists and biochemists who will have access to our techniques.

A.2 Background.

There are three convergent aspects to the background of this renewal proposal which are important to discuss in some detail. This discussion will serve to set the context of our proposed research in the framework of: a) our past work under this grant; b) the work of other research groups on problems related to past and proposed research; and 3) the relationship of our research effort to the SUMEX-AIM computer resource to which our research will be related.

A.2.a Background of the DENDRAL Project.

The problem of exploiting modern computer hardware and novel computer science techniques for the purpose of structure determination of organic compounds was initially addressed in the mid-1960's under the collaborative direction of Professors Edward Feigenbaum in Computer Science and Joshua Lederberg in the Department of Genetics. Among the results of this collaboration was an algorithm called DENDRAL (for DENDritic ALgorithm) and a computer program, given the same name, based on this algorithm. The program was capable of constructing or "generating" all structural isomers of molecular formulas corresponding to acyclic organic compounds [1,2]. (Although this program now exists only in greatly modified form as one of the kernel techniques in the CONGEN program [3], the term DENDRAL has remained as an informal name for the project.)

It was obvious from the outset that such a program must be constrained to produce only plausible structural isomers when applied to generating candidate isomers for an unknown structure. Although constraints can, and do, come from a variety of sources including chemical and physical methods and chemical intuition, initial efforts were on use of mass spectral data as a source of constraints, and collaboration with Professor Djerassi's group in the Department of Chemistry was begun, this group acting as a source of expertise in mass spectrometry. The results of this expanded collaboration were embodied in the Heuristic DENDRAL program, in which structural candidates were generated under heuristics, or rules, relating observed mass spectral data to plausible structural constraints, to yield a much smaller set of plausible structures than the total allowed by the molecular formula alone. Initial success with acyclic ketones [4] led eventually to a more general program for saturated, monofunctional compounds [5].

At about this time, financial support from the NIH was sought and funding for further work was begun in 1971. The project continued to be guided by the same general philosophy (which continues in only slightly modified form into the present proposal), namely, structure determination via constrained generation of isomers, each isomer representing a candidate structure for an unknown compound. However, in 1971 this work was pursued on two related, but somewhat divergent, courses which are only now beginning to converge again. One course involved advances in structure generation. Work on molecules of sufficient complexity to be of interest in biomedical research requires the capability for generating cyclic structures. One early approach [6] proved only a temporary stopgap and it was not until 1974 that the general problem was finally solved [7] and proven mathematically [8,9,10].

During this period interpretation of mass spectral data obtained from complex molecules was pursued successfully for compounds in specific chemical classes [11] for which detailed rules of mass spectral fragmentation were available in the literature or could be obtained by automated methods (the Meta-DENDRAL program [12,13]). This work utilized only a very primitive scheme for structure generation of representatives of a given class because the general cyclic structure generator was unavailable.

At the end of this period we possessed a complete structure generator, without constraints, and knowledge of how to use mass spectral data to provide constraints. It would have been possible at that point to integrate the interpretive program with the structure generator. This was not done because of the reliance, in most structure determination problems involving complex, polyfunctional molecules, on data from several physical and chemical methods besides mass spectrometry. Mass spectral data by themselves are insufficient to solve structural problems in the absence of knowledge of the chemical class to which the unknown belongs and detailed knowledge of the modes of mass spectral fragmentation for that class. However, at that time, with the limited computational tools available, it was not possible to develop a system for complex molecules which was capable of constraining the structure generator based on interpretation of diverse spectroscopic or chemical data. In addition, we did not feel such an automated approach, which largely excluded the chemist from the computations, was the proper way to utilize the computer in structural problems. For these reasons we developed mechanisms for constraining the structure generator with various substructural constraints [3,14], independent of the source of such substructural information. The resulting program (called CONGEN, for CONstrained structure GENeration) was made highly interactive and designed specifically to stand by itself as an aid in structure determination [3], performing only the structure generation aspects of the problem. At the same time, work on data interpretation, especially automated rule formation, continued separately, both for mass spectral [13] and ^{13}CMR [15,16] data. During this period CONGEN was utilized to solve a variety of structural problems in our own laboratories [17] and initial experiments with network availability of CONGEN via the SUMEX resource were carried out [18].

More recently, our efforts have concentrated on increasing the power of CONGEN in a number of ways and increasing its availability. We added the capability for computer simulation of chemical reaction sequences as a separate program (called REACT [19]), interfaced to CONGEN via structure files, and applied REACT to several problems involving reaction mechanisms [19], simulation of biochemical reactions [20] and applications to structure elucidation problems [21]. We have taken preliminary steps to provide assistance to the structural chemist after CONGEN has been used to propose structures, including automated examination of large numbers of structures to constrain or rank-order the structures based on combinations of desired and undesired structural features, and to assist in planning new experiments to differentiate among the candidates [22]. We have also been developing methods for prediction of mass spectra and ranking candidate structures based on extent of agreement between predicted and observed spectra [23]. We have made some progress in interpretation of ^{13}CMR data [16]. In addition we now have the capability for general treatment of configurational stereoisomerism in the STEREO program [24,25] as the first step in developing stereochemical extensions to CONGEN.

To increase accessibility, we have recently completed an exportable version of the CONGEN program, held an intensive series of workshops here at Stanford on the use of the new version and have begun exporting CONGEN to several sites around the country, including both academic and industrial laboratories whose work is heavily concerned with structure elucidation of compounds of biological importance. We are in the process of arranging for wider availability through network access here at SUMEX, and exploring the utility of access through the National Resource for Computation in Chemistry (NRCC) and the NIH/EPA Chemical Information System. These recent developments are described in detail in the accompanying Annual Report, Appendix I.

Achieving the goals of our last proposal period has left us in the following position. We have a very efficient, interactive program, CONGEN, for suggesting structural isomers as candidates for an unknown structure. We have ancillary

programs (e.g., REACT, SURVEY, EXAMINE, MSANALYZE, LOOK) to assist chemists in the evaluation of the candidates. We have studied automatic rule formation (the Meta-DENDRAL program) to determine relationships among known structures and associated spectral data. We have used such relationships to interpret data in mass spectrometry and, in much less general ways, ^{13}CMR spectroscopy. We have demonstrated the utility of spectral prediction and ranking as a method for evaluating large number of candidate structures, again in the area of mass spectrometry. We have developed the capability for generation of stereoisomers in the STEREO program as a first step toward consideration of three dimensional aspects of molecular structure.

We are now prepared to make a concerted effort toward representation and manipulation of structures in three dimensions. In many ways this is the logical next step in development of techniques for computer-assisted structure elucidation. By removing the current limitation, in all of the programs summarized above, of treatment of only structural isomers, we can consider a much more detailed and comprehensive approach to the use of computers in structure elucidation. Such an approach could utilize computational procedures throughout the entire procedure of structural assignment, including: 1) interpretation of diverse spectroscopic data to obtain substructural information; 2) postulating structural candidates based on the inferred substructures; and 3) evaluating those candidates by prediction of spectroscopic properties. The key to success of this approach is the capability for representing configurational and conformational stereochemistry. For example, data from many spectroscopic techniques are sensitive to molecular stereochemistry and such sensitivity must be taken into account to analyze adequately or to predict spectral data. The Specific Aims, Section B of the Research Plan, represent a logical set of steps to achieve these new objectives.

A.2.b Background of Proposed Research.

The preceding discussion has reflected only the development of our own research over the past few years. Our past efforts and our new proposals are obviously influenced by work of others in the field. This section is meant to provide a brief history of the use of computer techniques in structure elucidation (excluding of course, X-ray crystallography) to illustrate how our work is related to that of other groups. Although this review is not exhaustive, the important milestones are mentioned and references within the important papers referenced in this section can be used as a more detailed guide to relevant literature.

A.2.b.i Data interpretation and Prediction.

Early work on use of computers specifically for structure elucidation involved library search techniques, whereby a match of observed data, e.g., a spectrum, is sought in collection of spectra of known compounds. Mass spectrometry received (and still receives) the most attention due to the relative specificity of mass spectral fragmentations patterns, the sensitivity of the technique and the lack of direct relationships between functionalities in a compound and its fragmentation pattern (relationships provided much more explicitly by other methods such as NMR). Over the years several groups have have worked intensively on mass spectral search techniques; these efforts have been reviewed recently [26,27]. Sophisticated systems such as that developed by McLafferty and co-workers [28] now attempt to correlate several features of fragmentation patterns with structure to obtain substructural information from a library. Interest in such techniques has stimulated development of interactive systems to allow computer network access to large libraries on central computers from most areas of the US and some parts of

Europe [29]. More recently, collections of spectral data from other techniques have been utilized in library search systems, notably IR [30] and ^{13}CMR [31]. Other groups have developed computer-based systems for simultaneous search of several spectroscopic data bases (MS, IR, NMR) to attempt to derive structural information from close matches [32].

In our own work involving mass spectral analyses [33] we have made extensive use of library search techniques. Our goal has always been to identify using such techniques as many unknowns (in a complex mixture analyzed by GC/MS, for example) as possible before turning to more sophisticated procedures to aid in the identification of the remaining unknowns. This two-stage approach will be followed in the future wherever possible. In our own experience existing libraries, even the extensive collections of mass spectral data, are of limited utility in the natural products field where our computer techniques, especially CONGEN, have found the most application.

Computer approaches to the interpretation of spectroscopic data, in particular mass spectra data, were introduced even before computerized library search techniques. Initially, high resolution mass spectral data acquired on a specific compound class, the peptides, were analyzed by special-purpose programs [34]. Note that such a precise definition of compound class largely removes the requirement for a complete and irredundant structure generator. Although similar methods were proposed for other classes such as ketones [35], even the next step to low resolution spectra of aliphatic molecules required a structure generator for any generality [2,4,5] (obviously, special purpose programs cannot be written for every compound class!). Subsequently, we proposed a more general method for interpretation of high resolution mass spectral data [11] which could be applied in principle to any compound class, but could only be used under the assumption that the class of the unknown structure was known. Other programs have been proposed for analysis of mass spectral data, again in rather narrowly defined chemical contexts [36]. These efforts have not led to a general interpretive program for mass spectra primarily because of the difficulty in establishing the fragmentation rules, needed by such programs, for complex, polyfunctional molecules.

More recently, computational techniques have been applied to interpretation of other types of spectral data, where the goal of the interpretive procedure is to determine plausible substructures present in an unknown molecule. Methods have been developed for partial interpretation of IR [37] and ^{13}CMR [38,39] data and in two systems, use of data from a variety of spectroscopic techniques as an aid in substructural assignment [40,41,42] and in one of the systems eventual automatic generation of candidate structures [41,42]. These methods are currently little more than automated look-up procedures, where positions of observed resonances or absorptions are compared to tables relating position to substructure. In a method which consider data from several techniques [41,42] there is no systematic procedure to verify or assign plausibilities to the inferences from one technique by examination of relevant data from other, complementary techniques.

Pattern recognition procedures have been applied to diverse spectroscopic data in attempts to determine substructural information from mass [see reviews, references 26, 27] IR [43] and ^{13}CMR [38,44] spectral data. These approaches have been the subject of recent reviews [26,27,45]. In most pattern recognition work in this area, a program is "trained" to determine spectroscopic signatures which permit yes or no answers to specific questions about the presence of selected functionalities. Known structures and associated spectra are utilized in the training. These methods have been successful in areas where table look-up procedures also meet with some success, i.e., where the signatures for a functionality display little variation among a set of diverse structures. Such is

not the case for either mass or ^{13}CMR data and, although such programs can be trained to answer questions correctly among the training sets of structures, there have been no reported instances, to our knowledge, where the techniques have been applied successfully to prospective analysis of spectral data in terms of structure for actual unknowns of biological importance.

Our work has involved primarily symbolic, rather than numeric, calculations, based on strong models, e.g., of a spectroscopic technique like mass spectrometry, in order to relate structural features to observed data. Because of this emphasis we have not utilized statistical procedures. We still feel that, particularly in spectral interpretation, rule-based systems which are based on strong models and which can take advantage of the wealth of chemical information which has been collected in the past, stands a better chance of success in determining structural features from spectral data.

We also note that pattern recognition approaches have been used in prediction of spectral data, especially mass spectral data [46]. This approach is clearly an alternative to our past [11,13,23] and proposed developments in spectral prediction; no attempt has been made so far to compare the results of the two techniques because the applications have been so far been quite different.

Pattern recognition procedures have sometimes taken into account geometrical descriptions of molecular structure [47] based on manually selected molecular features deemed important or interesting in a particular investigation. In our past [24,25] and proposed studies relating to stereochemical descriptions of structure we have taken a more general approach describing configurations and conformations, again because in our applications there exist strong models relating observed data to structure, e.g., vicinal coupling constants in ^{1}HMR.

A.2.b.ii Topological Structure Generation.

In conjunction with the above interpretive procedures, systematic methods for structure generation have been developed. Given computer programs which can propose partial structures, it is only logical to consider how these partial structures can be assembled into complete structures. The first computer program for exhaustive, irredundant structure generation dealt with acyclic structural isomers, with or without multiple bonds [1,2]. Subsequently, groups headed by Sasaki [41], Munk [48a] and our own group [7] proposed methods for construction of structures including those with ring systems. Since that time, these groups, including more recent efforts of DuBois [49] and Gribov [50], have concentrated on increasing the repertoire of constraints and chemical knowledge of the programs to improve performance [3,14,42,48b]. Of these systems, the work of Sasaki and co-workers on the CHEMICS program [41,42] is the most ambitious in that an attempt is made to automate the entire procedure of structure elucidation from spectral interpretation through to proposal of final structures.

The work of Munk and co-workers, and to a lesser extent that of Gribov and DuBois, is most closely related to our past goals. In fact, we have on several occasions cooperated with Munk's group on both conceptual matters and specific approaches. For example, his group employs our teletype structure drawing program as one method for output of structures. We have adopted some of their ideas on user interaction in our CONGEN program. Both CONGEN and their program, CASE, are designed for the same task. There are, however, substantial differences in computational procedures and user interaction; which approaches are preferable is arguable. We feel our methods rest on a firmer, mathematically proven, foundation. However, past comparisons of results of solved structural problems have so far been

in agreement. Our proposals to extend our efforts into spectral interpretation, structure generation and spectrum prediction based on stereochemical representations of structure, and our proposed capabilities for structure generation given overlapping structural units, discussed subsequently, represent fundamental departures from the similarities among the research efforts mentioned above.

The work of Sasaki and co-workers bears most closely on our own proposals for more automated systems for structure elucidation. We feel their approach is limited in several ways which our proposed research is designed to overcome: 1) the concept of total automation has so far been demonstrated successfully only for relatively simple structures [41,42], e.g., where ^{1}HMR spectra are essentially first-order. We feel that the chemist, with his or her specific knowledge about what is probably a much more complicated unknown, must remain for the near future, at least, an integral part of a system for structure elucidation which may nevertheless be highly automated but which still retains its interactive nature in order to guide and control the procedures; 2) there is no treatment of stereochemistry in their approach, a necessity, we feel, for a more useful program; and 3) structures must be built in CHEMICS from a library of small structural fragments, in part because CHEMICS cannot perform structure generation with overlapping substructures. This leads to considerable inefficiency for larger structures. Our new methods for structure generation are designed to overcome this limitation.

A.2.b.iii Conformational Structure Generation.

The problems of the generation and/or enumeration of the conformations possible for a given chemical structure have been approached in several ways. These are approaches which do not specifically compute energies for the conformations. An energy calculation is often done after the geometric coordinates of the conformation have been determined. Representatives of these approaches will be divided into three classes:

1) Studies related to properties of polymers

2) Studies related to conformational analysis of single cyclic structures

3) Exhaustive computerized methods using coordinate representations.

Approaches to conformation enumeration and/or generation related to the study of polymeric properties are often concerned with the study of linear acyclic structures. Among the properties studied are cyclization equilibria which necessitates consideration of the possible cyclic or near-cyclic structures of a linear chain. R. P. Smith [51] did early calculations which enumerated the possible conformations of a linear molecule. These were all the possible conformations which could be imbedded in a diamond lattice of carbon atoms. This is a simple geometric model since there are only four possible bond directions on a diamond lattice. He also enumerated the possible cyclic structures up to ring size 18, also on a diamond lattice. A large amount of work has been done by P. J. Flory on the statistical properties of chain molecules [52]. The method used assumes a discrete number of rotameric states for a single bond, usually three (trans, gauche(+), gauche(-)). The method is used to compute average end-to-end distances of chains among other properties. These end to end distances can be used to study cyclization equilibria. In these studies conformations are rarely explicitly constructed, but are instead studied as statistical averages because of their enormous number. Computations of the number of conformations of linear chains which can close to form rings have been done by Semlyen [53]. This is a geometrical method which explicitly considers bond

lengths and angles. In general these studies primarily enumerate conformations rather than generate them explicitly since the number of conformations, or the average number, with a particular property are the principle interest. Conformations are only rarely constructed, and then by Monte Carlo methods rather than exhaustively.

Approaches to conformation enumeration and/or generation related to conformational analysis are generally concerned with monocyclic structures. To determine the conformations of cyclic molecules it is necessary to know the theoretical possibilities. Hendrickson [54] was one of the first to deal with this problem by determining the possible symmetrical conformations of cyclic hydrocarbons. This was done by labelling the bonds of the ring with the possible signs of the dihedral angle for that bond. There were conditions on allowed sequences of signs which were considered energetically unfeasible and conditions which maintained symmetry as no unsymmetrical conformations were considered. These conformations were "processed" by doing empirical energy calculations. This work provided a useful enumeration and classification of these symmetrical cyclic conformations. More recently, Scheraga [55] has approached the problem of determining the possible conformations of cyclic peptides with emphasis on symmetrical structures. This was a geometrical approach with postulated cycle closure conditions which was used eventually to examine the energy space for each molecule considered. Dale [56] has done a systematic manual construction of possible conformations for cyclic molecules for ring sizes 9-16. There was no comment on the completeness of the resulting list. Conformations were generated by labelling bonds as either gauche or trans subject to a number of heuristic rules on allowed sequences of signs. An interesting representation of the resulting conformations was given which involved considering the ring as a polygon. The processing of the resulting conformations was again an energy calculation. Saunders [57] has attempted to generate all possible diamond lattice conformations of rings up to 24 atoms. This approach relies on the restricted number of bond directions possible on a diamond lattice and several heuristic rules for disallowed situations. Only even numbered rings larger than 4 atoms can be fit onto a diamond lattice.

Another approach has been taken by Strauss [58], which generalizes work of Pitzer [59] on the pseudorotation of a 5-membered ring. A reference plane is chosen and the z-coordinates of atoms with respect to that plane are considered. The possible conformations can be classified by their symmetry properties with respect to the symmetry group of the ring. This method is restricted to "convex" conformations, so is not exhaustive, but has the advantage of providing a convenient representation of pseudorotation processes. A somewhat similar approach has been taken by Cremer [60] to describe puckering coordinates of a ring. These can be used to construct possible conformations. A similar approach has been used by Altona [61] to analyze the conformations of some five-membered ring sugars.

Computer programs have been written to generate exhaustively possible conformations which use coordinate representations and vary torsional angles by increments. One such program has been recently described by Dirkx [62] for cyclic molecules. Different programs were used to treat different molecules as a general program was not thought reasonable. When cyclic peptides were considered, only a very few torsional states were allowed to reduce the magnitude of the problem. No problems with symmetry were considered. A similar idea has been used by Murakami [63] for acyclic conformations. This program has been applied to the side chain conformations of prostaglandins allowing three rotational states per side chain bond. A more general program has been described by Marshall and Barry [64,65]. This program is coordinate based and generates conformations by varying some coordinates (which are input to the program and may or may not be torsional angles) by very small increments. The conformations are checked for various constraints

while they are being generated. Symmetrical structures are not explicitly considered. The program is used in part for determining drug receptor sites.

Much can be done in structure elucidation with knowledge of conformational properties of molecules. In particular, it is necessary to get information about atomic coordinates and possible conformations. The problem of generating the possible conformations (based on discrete values for some internal coordinates such as torsional angles) for an organic chemical structure of given constitution and configuration has only been addressed for several special cases and a general solution for any possible structure is lacking. Such a general solution would have to consider acyclic, cyclic, and polycyclic structures of any possible symmetry. We propose (Section C.4) to develop such a solution.

A.2.c Relationship to the SUMEX-AIM Resource.

The pursuit of our research goals over the past few years and continuing through this proposal has involved a slow but steady shift away from initial reliance on mass spectral data as a source of constraints for our structure generation procedures. We now are proposing to embark on much more general approaches to computer applications in structural analysis including, but not limited to, analysis of mass spectral data. This shift in emphasis has prompted important changes in the nature of the budget in the current proposal and in the nature of the resource to which our proposal is related. In the past, the resource to which our research was related was the mass spectrometry laboratory in the Department of Chemistry. We will continue to analyze mass spectral data provided by the mass spectrometry laboratory as part of our work on spectral prediction. However, funds to support the laboratory personnel and operations and to maintain the instrumentation are now being requested from other sources (see Research Support). Now, however, our computational approaches are becoming much more general and we plan wide dissemination of the programs resulting from our work. These more general approaches to aids for the structural biochemist will yield computer programs with much wider applicability than, for example, the existing CONGEN program. We expect that this will create a significant increase in requests for access to our programs, placing heavy emphasis on our relationship with SUMEX to provide this access.

For the above reasons, we identify the Stanford University Medical EXperimental computer facility for research in Artificial Intelligence in Medicine, SUMEX-AIM, as the resource to which our proposed research is related. (A plan for resource management is presented in Section C.5, Resource Sharing). The SUMEX-AIM resource has provided the computational basis for our past program developments and for initial exposure of the scientific community to these programs. The resource is, however, funded completely separately from our own research; we are only one of a nationwide community of users of the SUMEX-AIM facility. In a sense, then, relating our new research to SUMEX formalizes a relationship which already exists. However, such a formalization seems much more relevant now than in the past because of our broader emphasis on software tools and new capabilities for sharing the results of our research. The relationship which we propose (and discuss more fully in Section C, Methods of Procedure) is one which goes far beyond mere consumption of cycles on the SUMEX machine. It has been the goal of the SUMEX project [18] to provide a computational resource for research in symbolic computational procedures applied to health-related problems. As such research matures, it produces results, among which are computer programs, of potential utility to a broad community of scientists. A second goal of SUMEX has been to promote dissemination of useful results to that community, in part by providing network access to programs running on the SUMEX-AIM facility during their development phases. SUMEX does not, however,

have the capacity to support extensive operational use of such programs. It was expected from the beginning that user projects would develop alternative computing resources as operational demands for their programs grew. Such a state has been reached for the CONGEN program and future developments to yield more generally useful programs will simply magnify the problem.

We have proposed, therefore, under the new relationship between SUMEX-AIM and our project, to participate as before in the SUMEX-AIM community in sharing methods and results with other groups during development of new programs. In addition, we plan to purchase a small machine which will allow us to provide more extensive operational access to our existing and developing programs, and to provide a test environment for adapting our programs to a more realistic laboratory computing environment than the special-purpose SUMEX resource. This facility will derive substantial benefit from its relationship with SUMEX including sharing of network gateways, some peripheral equipment and operational support. On the other side of the coin, SUMEX benefits by moving a substantial part of the DENDRAL production load to a more cost-effective system, thereby freeing the SUMEX resource for new program development. Also, working with the proposed new machine (DEC-VAX, see Budget Remarks) will be advantageous as a model for SUMEX's future development. Collaborators who wish to use existing programs for specific problems would access SUMEX via the network as before, but now would be routed to the new machine. New developments, such as those we propose, would be carried out on SUMEX itself, taking advantage of the much more extensive repertoire of peripheral devices, languages, debugging tools and text editors, i.e., precisely the tasks for which that system was designed.

Our proposed relationship to SUMEX-AIM has important implications beyond the practical considerations mentioned above. There is a significant research component to our proposal to make the dedicated computer as integral part of the resource sharing aspects of our relationship to SUMEX. If our proposal is approved and funded, the DENDRAL project would be the first of the SUMEX-AIM projects to have developed sufficient maturity to request and obtain additional computer facilities to support production use of its programs in real-world, biomedical applications. In a sense, then, we will be acting in a pathfinding role for the rest of the SUMEX-AIM community as other projects reach maturity and seek additional computing resources to support the needs of their collaborators. Our approaches to interfacing with SUMEX with the dedicated machine, implementing new software, regulating access to divert development and applications to the appropriate machine are all experiments which we are willing to undertake together with SUMEX, knowing that we will be providing direction to future efforts along similar lines. We will also be in a pathfinding role for a large segment of the biochemical community involved in computing, as we move to a machine which will be much more widely available in Department and laboratory environments than DEC-10's and -20's. There are currently no widely available computing resources which provide access to symbolic, problem solving programs operating in an interactive environment. We would be able to fulfill that need to the extent that applications have direct, biomedical relevance, to the limits of our available computing resources.

A.3 Rationale.

We have initiated this proposal at what we feel is a particularly opportune time in the development of computer aids to structure elucidation. We are beginning to push our techniques for spectral interpretation, structure generation (e.g., CONGEN) and spectral prediction to their limits within the confines of topological representations of molecular structure. Even so, these techniques are perceived to

be of significant utility in the scientific community as evidenced by our workshops, the demand for the exportable version of CONGEN and the number of persons requesting collaborative or guest access to our programs at Stanford. In order to proceed further in providing to the community programs which are more generally applicable to biological structure problems and more easily accessible we must address squarely the limitations inherent in existing approaches and search for ways to solve them. The major objectives of our proposal are addressed to these issues in the following ways.

None of the techniques for computer-assisted structure elucidation of unknown molecular structures described in the previous section, including our own, make full use of stereochemical information. As existing programs were being developed this limitation was less important. The first step in many structure determinations is to establish the constitution of the structure, or the topological structure, and that is what CONGEN, for example, was designed to accomplish. However, most spectroscopic behavior and certainly most biological activities of molecules are due to their three-dimensional nature. For example, in a recent program for prediction of the number of resonances observed in ^{13}CMR spectra [39] the topological symmetry group of a molecule is used in prediction. However, in reality it is the symmetry group of the stereoisomer that must be used. This group reflects the usually lower symmetry of molecules possessing chiral centers and which generally exist in fewer than the total possible number of conformations. This will increase the number of carbon resonances observed over that predicted by the topological symmetry group alone. More generally, few of the techniques described in the background section can be used in accurate prediction of structure/property relationships, whether the properties be spectral resonances or biological activities.

A structure is not, in fact, considered to be established until its configuration, at least, has been determined. Its conformational behavior may then be important to determine its spectroscopic or biological behavior. For these reasons we emphasize in this proposal development of stereochemical extensions to CONGEN, existing related programs and the proposed new programs GENOA and SASES, including machine representations and manipulations of configuration and conformation and constrained generators for both aspects of stereochemistry.

None of the existing techniques for computer-assisted structure elucidation of unknown molecules, excepting very recent developments in our own laboratory, are capable of structure generation based on inferred partial structures which may overlap to any extent. Such a capability is a critical element in a computer-based system, such as we propose, for automated inference of substructures and subsequent structure generation based on what is frequently highly redundant structural information including many overlapping part structures. Important elements of our proposal are concerned with further developments of such a capability for structure generation (the GENOA program).

Given the above tools for structure representation and generation, we can consider, and have proposed, new interpretive and predictive techniques for relating spectroscopic data (or other properties) to molecular structure. The capability for representation of stereochemistry is required for any comprehensive treatment of: 1) interpretation of spectroscopic data; 2) prediction of spectroscopic data; 3) induction of rules (Meta-DENDRAL-like rule formation [13,15,16]) relating known molecular structures to observed chemical or biological properties. These elements, taken together, will yield a general system for computer aided structural analysis (the SASES system) with potential for applications far beyond the specific task of structure elucidation.

Parallel to our program development we will embark on a concerted effort to extend to the scientific community access to our programs, and critical parts of our proposal are devoted to methods for promoting this resource sharing. Our rationale for this effort is that the techniques must be readily accessible in order to be used, and that development of useful programs such as we propose can only be accomplished by an extended period of testing and refinement based on results obtained in analysis of a variety of structural problems, analyzed by those scientists actively involved in solutions to those problems.

To this end, we have proposed to purchase a dedicated computer system for applications of our programs. Our current collaborators (Section F) will obtain much better computational support by accessing the proposed system than the can obtain currently via SUMEX, which is very heavily loaded. Programs developed on the new machine will enjoy wider exportability. In addition, by offering better service, through both network access and export, we can attract new applications and make firm guarantees of computational support with fast response time for interactive programs, something we cannot do at the present time.

The overall rationale for all our developments is that structure determination of unknown structures and relationship of known structures to observed data are complex and time-consuming tasks. We know from our past experience that computer programs can complement the chemist's knowledge and reasoning power, thereby acting as valuable assistants. If we meet our present objectives, we feel strongly that our programs will become essential tools in the repertoire of techniques available to the structural chemist or biochemist.

B SPECIFIC AIMS.

The specific objectives for the requested five year period of support include the following:

1) Develop SASES (Semi-Automated Structure Elucidation System) as a general system for computer aided structural analysis, utilizing stereochemical structural representations as the fundamental structural description. SASES will represent a computer-based "laboratory" for detailed exploration of structural questions on the computer. It will have as key components the following:

a) Capabilities for interpretation of spectral data which, together with inferences from chemical or other data, would be used for determination of (possibly overlapping) substructures;

b) The GENOA (structure Generation with Overlapping Atoms) program which will have the capability of exhaustive generation of (topological and stereochemical) structural candidates and include as an essential component the existing CONGEN program;

c) Capabilities for prediction of spectral (and biological) properties to rank-order candidates on the basis of agreement between predicted and observed properties.

2) Develop the GENOA program and integrate it with CONGEN. GENOA will represent the heart of SASES for exploration of structures of unknown compounds, or configurations or conformations of known compounds. GENOA will be a completely general method for construction of structural candidates for an unknown based on redundant, overlapping substructural information, and it will include capabilities for generation of topological and stereochemical isomers.

3) Develop automated approaches to both interpretation and prediction of spectroscopic data, including but not limited to the following spectroscopic techniques:

a) carbon-13 magnetic resonance (^{13}CMR);

b) proton magnetic resonance (^{1}HMR);

c) infrared spectroscopy (IR);

d) mass spectrometry (MS)

e) chiroptical methods including circular dichroism (CD), magnetic circular dichroism (MCD).

The interpretive procedures will yield substructural information, including stereochemical features, which can be used to construct structural candidates using GENOA. The predictive procedures will be designed to provide approximate but rapid predictions of expected spectroscopic behavior of large numbers of structural candidates, including various conformers of particular structures. Such procedures can be used to rank-order candidates and/or conformers. The predictive procedures will also be designed to provide more detailed predictions of structure/property relationships for known or candidate structures in specific biological applications.

4) Develop a constrained generator of stereoisomers, including:

a) design and implement a complete and irredundant generator of possible conformations for a given known, or a candidate for an unknown, structure;

b) provide constraints for the conformation generator so that proposed structures for a known or unknown compound possess only those features allowed by: i) intrinsic structural features such as ring closure and dynamics of the chemical structure; and ii) data sensitive to molecular conformations (e.g., MCD, NMR);

c) integrate the stereochemical developments with the GENOA program as a final, comprehensive solution to the structure generation problem and allow for interface of the program with other methods dependent on atomic coordinates.

5) Promote applications of these new techniques to structural problems of a community of collaborators, including improved methods for structure elucidation and potential new biomedical applications, through resource sharing involving the following methods of access to our facilities and personnel;

a) nationwide computer network access, via the SUMEX computer resource and the dedicated machine requested in the first year of the current proposal;

b) exportable versions of programs to specific sites and via the National Resource for Computation in Chemistry and the NIH/EPA Chemical Information System;

c) workshops at Stanford to provide collaborators with access to existing and new developments in computer-assisted structure elucidation in an environment where complex questions of utility and application can be answered directly by our own scientific staff;

d) interface to a commercially available graphics terminal for structural input and output, at as low a cost as possible, so that chemists can draw or visualize structures more simply and intuitively than with our current, teletype-oriented interfaces.

C METHODS OF PROCEDURE.

C.1 SASES — A Semi-Automatic Structure Elucidation System.

A long term aim of our research is the development of a system which performs automated data analysis, structure generation, spectral prediction and ranking of candidate structures. In this section we will outline the construction of the system ("SASES"), and in the following three sections we will discuss the proposed development of the component parts.

There are two important themes underlying the system. The first is that, in our opinion, the task of structure elucidation is sufficiently complex that the chemist must remain an integral part of the system. The name for the system, SASES, for Semi-Automated Structure Elucidation System, reflects our judgment that at several places throughout the system the chemist must be able to examine the status of the computations and express his/her own judgments on how best to proceed. This will be an interactive system and not a black box with data input and structure output. The chemist will be able to exercise the various aspects of structure determination (data interpretation, constrained structure generation, spectral prediction, structure ranking) throughout the process. The second theme is the reliance on stereochemical representations of structures throughout SASES. This capability will make component parts of SASES applicable to several other problems, such as structure/property relationships, besides the central task of structure elucidation.

The novelty of this proposed system for structure elucidation, and the features which set it apart from other current systems (including our present system) are the incorporation of stereochemical information throughout and the ability to make use of any and all redundant or overlapping partial structural information. The full incorporation of stereochemical information (both configurational and conformational) will allow the use and fuller interpretation of a wider body of spectral and chemical data. The ability to use any redundant or overlapping data will simplify the use of the system and provide a much "smarter" chemist's assistant.

The SASES system is described in block diagram form in Figure 1 The solid lines connecting the chemist to various programs or output from SASES imply the capability for interacting with the procedures. The dashed lines simply indicate that both "INTERP" and "PREDICTOR" modules can access the same data.

The "INTERP" module, shown in Figure 1, represents a program which will combine a number of existing and proposed functions for inferring structural information from diverse data (See Section C.3). The chemist will interact with the interpretive procedures in at least two ways. He will be able to add substructural inferences to the growing list based on his own judgements, or data from other techniques (e.g. chemical behavior, UV spectroscopy etc). In addition, he will be capable of examining (perhaps in question/answer mode with the program) the growing list of substructures thereby noting extensions to them or new experiments (e.g. proton decoupling) to be performed. The output of "INTERP" will be in the form of a file of computer representations of substructures, containing both topological and stereochemical information (possibly with associated plausibility ratings).

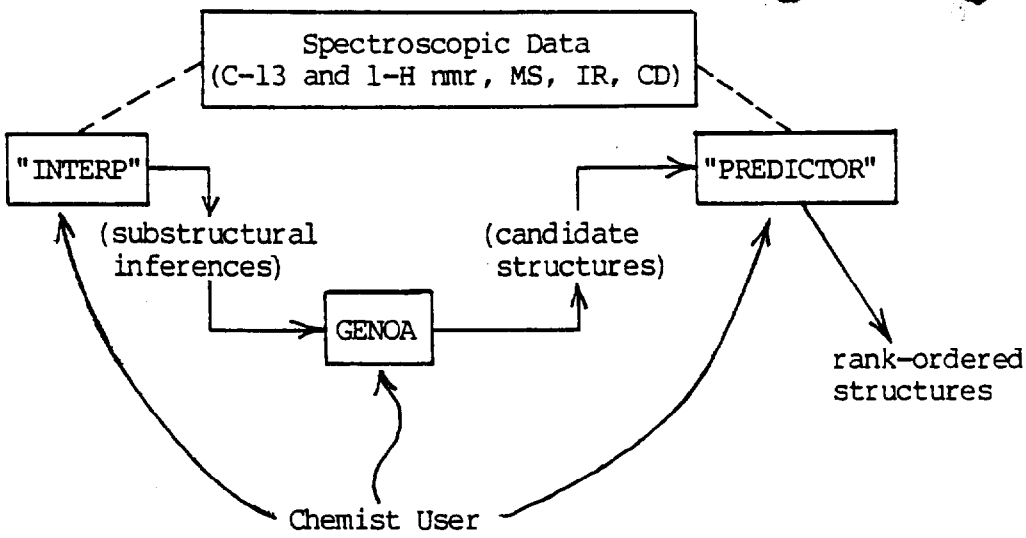


Figure 1. Block Diagram of the SASES system.

The GENOA program (for GENeration of structures with Overlapping Atoms) will include:

1) The current CONGEN program for generation of constitutional (i.e. bond connectivity) and configurational stereoisomers;

2) The current preliminary version of GENOA;

3) The proposed conformation generator (Section C.4);

4) The proposed developments to GENOA (Section C.2).

GENOA will build structures incrementally, with the chemist able to exert as much (or as little) control as desired. In particular, the chemist will be able to add additional constraints as required as the individual structures take shape in the stepwise generation procedure. The chemist will also have control over the extent to which stereochemistry is considered. For example, early in structure generation problems, when the number of candidates is large, topological generation may be sufficient to guide the next step with details of stereochemistry incorporated as detailed knowledge about the structure becomes available (the number of topological representatives thereby decreasing).

The output of GENOA will be a file of structures which can be examined by the chemist, and more importantly, saved away for further processing at a later time. Such facilities are extremely important at intermediate stages of a complex structure problem. Analysis can proceed utilizing the set of structures which were plausible based on previous data, thereby avoiding time-consuming recomputation of the problem. These interactive facilities are an integral part of the existing CONGEN and GENOA programs and will certainly be included in all subsequent developments.

The "PREDICTOR" functions will communicate with GENOA through the structure file, and will be capable of sharing observed data with the interpretive procedures (Figure 1). These functions will be used to predict spectral properties for GENOA's candidate structures. As discussed in Section C.3, the "PREDICTOR" functions are

intended to utilize more precise models of spectroscopic behavior, that relate to total molecular conformation, as opposed to the simpler models used in "INTERP", which relate primarily to local, topological structure. The "PREDICTOR" functions will also rank order structures based on agreement between observed and computed spectroscopic data. The output will be a rank-ordered structure file, which can be used subsequently in GENOA for further pruning, use of the EXAMINE/SURVEY functions and so forth.

The structure of SASES will be designed to be completely modular in that each of the three major portions can be run independently. In this way, we can provide a computer-based "chemical laboratory" in which "experiments" can be performed on the computer to help save time and valuable sample. Links between the modules will be files of structural information, further contributing to the modularity. This modularity will increase the utility of SASES in a number of ways. For example, it will be desirable to run the "INTERP" module on existing data as a stand-alone early in a structure elucidation problem in order to develop ideas about what additional information would be required before attempting to construct structures. The GENOA module will continue to find extensive use as a stand-alone structure generator for many problems where structural inferences have been determined by other methods. In addition, the combined structure generator and predictor will be used for several studies on structures whose topology is known but whose stereochemistry remains imprecisely defined. In Section F, Collaborative Arrangements, we discuss two collaborative projects where our techniques will find application to known structures in attempts to relate observed properties (NMR couplings, biological activity) to computed conformers.

C.2 The GENOA Program: Structure Generation with Overlapping Atoms

As discussed in Section A.2, one of the significant limitations of current approaches to structure generation is the requirement for disjoint substructures. In other words, substructures and atoms input to a structure generator such as CONGEN must not overlap; the total number of atoms of each type in the collection of substructures and remaining atoms must agree with the molecular formula. One of the aims of the previous proposal was to develop an approach which would remove this limitation. This work was characterized as "constructive substructure search" or "constraints interpretation/translation." Together, these names are suggestive of a procedure whereby inferred substructures are built from previously inferred components (atoms and substructures), considering all possible overlaps. Such a procedure removes the requirement for non-overlapping components and allows the chemist to input what are usually-redundant structural data, obtained from various spectroscopic techniques, in a completely natural way. This makes the program much easier to use, an important consideration for any program to be used by a wide community of persons.

Our initial work on such a program led to a version based on the INTERLISP version of CONGEN (hereafter referred to as old CONGEN, or OCONGEN). This program [66] accomplished its designed goal, of taking "GOODLIST," or desired, substructures and incorporating them into the structure generation problem in all possible ways. The program was, however, quite limited in that it had no mechanism for elimination of undesired structural features, nor was it interfaced to OCONGEN for final construction of structures. It was also very inefficient.

This approach was set aside for several months as we emphasized the development and initial export of the new CONGEN program. Although we believed that an approach to structure GENeration with Overlapping Atoms (hereafter referred to as

GENOA) was an ultimately desirable goal, frankly it was not until the series of workshops late last year (see attached Annual Report, Appendix I) on the use of the new CONGEN program that we realized both the practical and conceptual potential of GENOA. Workshop participants were spending more time to analyze and decompose their existing substructural information into non-overlapping substructures than they used in actually solving the problem with CONGEN. The statement, made to workshop participants, that it is difficult to handle, completely and irredundantly, problems expressed with overlapping substructures was accepted but hardly understood. After all, they are forced to consider structural information in that way when solving the problems manually; a decent computer program ought to do the same.

As a result, during the past three months we have assembled, in the BCPL language utilizing portions of CONGEN and extensive new code, a much more efficient version of GENOA based in part on the earlier INTERLISP concepts but improved by experience and by removing some of the limitations. The current status of GENOA is summarized in Figure 2.

Briefly, GENOA obtains the molecular formula for an unknown compound, and the number (may be an integer, a range or zero) and name of inferred substructures, one at a time. For each new substructure, GENOA builds the requested number and ensures that the required number of all previous substructures is met. Utility functions allow definition of substructures, and visualizing and saving all intermediate results. As an example of use of GENOA with overlapping substructural information, we present in Figure 3 one of the many possible ways to supply substructural information for the structure of palustrol [17a], together with examples of intermediate problems.

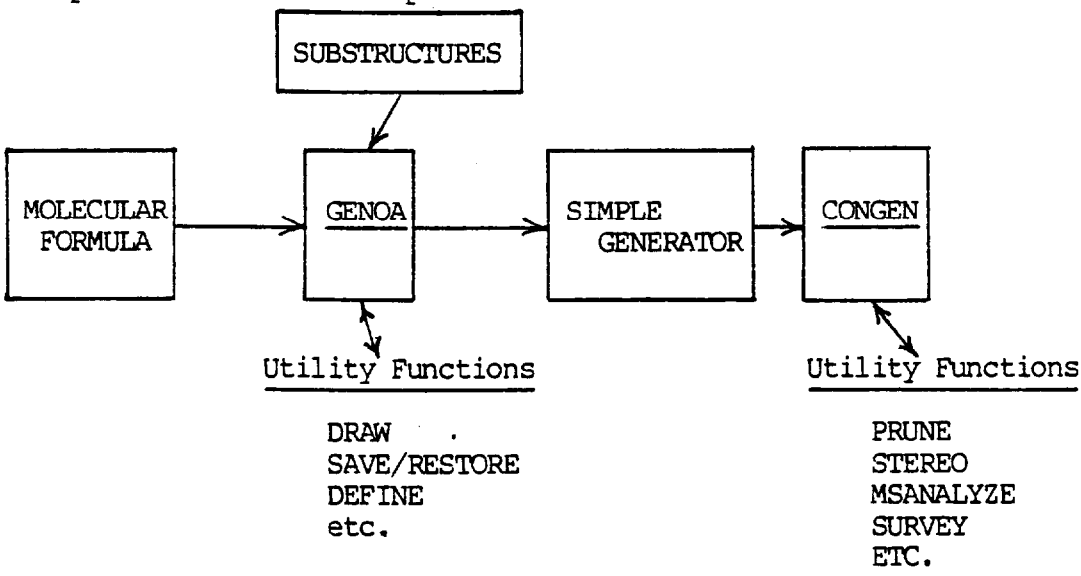


Figure 2. Current Status of GENOA and its Interface to the CONGEN Program.

This example (Figure 3) is shown to illustrate several points about the use of GENOA in a structural problem. One can, as was done in this problem using CONGEN [17a], wait until many of the data are gathered, determine non-overlapping substructures and use constraints to test for those substructures which might overlap, at the end of the problem. Using GENOA, however, yields greater efficiency through use of data and inferred substructures as they are gathered and applied to the problem. From the very beginning of the problem one can examine the implications of the next piece of information and, interactively, remove structures with undesirable features early in the problem. Because only those substructures

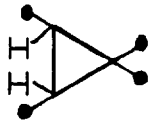
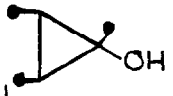
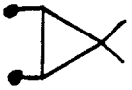
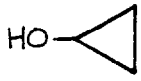
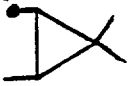
are constructed which have been supplied to GENOA, i.e., no attempt is made to generate complete structures until the user wishes, problems remain very small in terms of numbers of possibilities which must be considered at each step, even when very little information is known about the structure. To illustrate these points, consider the steps in Figure 3 (The sequence of steps parallels approximately the order in which data were collected on the structure).

After definition of the molecular formula (Step 1), GENOA was given the results of ^{13}CMR analysis summarized in Step 2, thereby specifying the number of carbons of each degree (the alcohol functionality had been previously determined). The result is a single problem for GENOA, which looks exactly like the input data. Step 3 specified the presence of two tertiary methyl groups from ^{1}HMR. Three problems result Figure 3, each representing one way of obtaining two such groups. Examination of these three problems at the computer terminal and consideration of data obtained from use of shift reagent allowed Step 4, specifying no methyls attached to the carbinol carbon, thereby removing two of the previous problems. The cyclopropyl ring with two, vicinal hydrogens, was specified in Step 5. There are two ways to construct that substructure from the previous problem, as shown. ^{13}CMR data allowed rejection of the case in which the carbinol carbon is in the three-membered ring, Step 6. In Step 7, two secondary methyls, from proton NMR, are specified, resulting in three problems, one of which is removed in Step 8, no isopropyl groups. In Step 9, the environment of the cyclopropyl ring is specified in more detail, and, finally, in Step 10, more detail based on decoupling experiments is provided, resulting in two problems. Note that at each step there was no concern over what atoms might overlap, and at each step examination of the results yielded additional constraints which had been overlooked up to that point. The final generation of structures, Step 11, yields 81 complete structures from the two problems. At any point throughout the procedure, the problem can be saved and then recalled at a later time when additional data are available. Thus, GENOA's analysis of a problem can work hand in hand with laboratory experimentation.

GENOA's method of construction of overlapping substructures is completely general in that any substructure of any size (not, of course, exceeding the molecular formula) can be specified. For example, if prior to Step 11, Figure 3, one wished to construct only those structures which obeyed the head-to-tail isoprene rule, one could define the fifteen carbon substructure representing that linkage and supply it to GENOA, which would then construct alternatives based on the problems already specified through Step 10. Subsequent generation would then be only of structures obeying that form of the isoprene rule. Note that the program determines not only how the required substructures can be built, but also makes structural inferences concerning the implications of each statement. For example, the characteristics of this problem force two of the methyl groups to be geminal and on the cyclopropyl ring, even though no explicit statement of that partial structure was made.

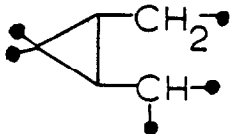
Djerassi, Carl

[redacted]

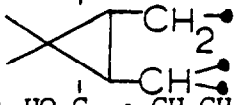
STEP	RESULTS AND EXAMPLES
1. DEFINE MOLECULAR FORMULA	$C_{15}H_{26}O$
2. (1) HO-C- (1) -C- (5) -CH- (4) $-CH_2-$ (4) CH_3-	1 PROBLEM, EXACTLY AS INPUT DATA
3. (2) CH_3-C-	3 PROBLEMS: 1) $(CH_3)_2-C-$; HO-C- ; 5 X -CH-; 4 X $-CH_2-$; 2 X CH_3- . 2) CH_3-C- ; CH_3-C-OH ; 5 X -CH-; 4 X $-CH_2-$; 2 X CH_3- . 3) $(CH_3)_2-C-OH$; -C- ; 5 X -CH-; 4 X $-CH_2-$; 2 X CH_3- .
4. (Ø) CH_3-C-OH	1 PROBLEM, #1, PREVIOUS STEP.
5. (1) 	2 PROBLEMS: 1) $(CH_3)_2-C-$; 3 X -CH ; 4 X $-CH_2-$; 2 X CH_3- ;  2) HO-C- ; 3 X -CH-; 4 X $-CH_2-$; 2 X CH_3- ; 
6. (Ø) 	1 PROBLEM, #2, PREVIOUS STEP.
7. (2) CH_3-CH-	3 PROBLEMS: 1) HO-C- ; CH_3CH-; 2 X -CH-; 4 X $-CH_2-$;  2) HO-C- ; $(CH_3)_2-CH$; 2 X -CH-; 4 X $-CH_2-$;  3) HO-C- ; 2 X CH_3CH-; -CH- ; 4 X $-CH_2-$; 

8. (Ø) $(CH_3)_2\text{-}\overset{|}{C}H\text{-}$ — 2 PROBLEMS, #1,#3, PREVIOUS STEP.

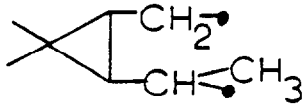
9. (1) — 2 PROBLEMS;



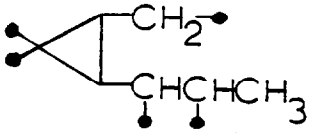
1) $HO\text{-}\overset{|}{C}\text{-}$; 2 X $CH_3\overset{|}{C}H\text{-}$; 3 X-$CH_2$- ;



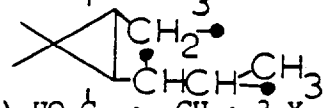
2) $HO\text{-}\overset{|}{C}\text{-}$; $CH_3\overset{|}{C}H\text{-}$; $\text{-}\overset{|}{C}H\text{-}$; 3 X -$CH_2$- ;



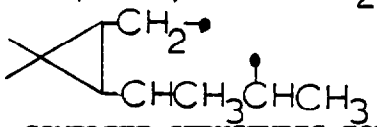
10. (1) — 2 PROBLEMS:



1) $HO\text{-}\overset{|}{C}\text{-}$; $CH_3\overset{|}{C}H\text{-}$; 3 X -$CH_2$- ;



2) $HO\text{-}\overset{|}{C}\text{-}$; $\text{-}\overset{|}{C}H\text{-}$; 3 X -$CH_2$- ;



11. GENERATE STRUCTURES:

(Ø) C=C; (1) 

(Ø) C≡C

81 COMPLETE STRUCTURES FOR FURTHER EXAMINATION, FOR EXAMPLE:

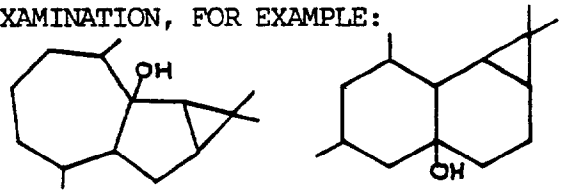


Figure 3. Illustration of the Use of GENOA in the Step-By-Step Specification of Overlapping Structural Information for the Structure of Palustrol [17A].

After the last known substructure is specified, a simplified structure generator, not the one utilized in CONGEN, is used currently to build complete structures. The generator is quite inefficient and creates many duplicate structures which must be removed (automatically). Control is then passed directly to the CONGEN program where all the currently available utilities for further processing, e.g., STEREO, MSANALYZE, may be used to prune or explore further the structural candidates.

Our proposals to extend the current version of GENOA include the following steps, which we feel represent a logical course to pursue in the context of our other goals. We will expend this effort because GENOA will be the heart of the proposed SASES program. We will:

a) evaluate the existing program with selected collaborators to determine strengths and weaknesses on actual problems;

b) extend the current version, applicable only to topological representations of structure, to make a complete, stand-alone version of GENOA integrated completely with CONGEN including:

i) greater user control over allowed overlaps of substructures;

ii) integrate the output of GENOA with CONGEN's efficient, interactive method for structure generation.

c) commence work on a version of GENOA which incorporates stereochemical constraints during generation;

Each of these proposals is now explained in more detail.

C.2.a Evaluate Current Version of GENOA.

Even though the current version has significant limitations, it possesses all the desirable interactive characteristics of CONGEN and has so far proven robust (error-free). In order to guide further development we will begin evaluation of GENOA by allowing selected collaborators to make use of the program in parallel with CONGEN. This will give us important measures of utility in real applications and allow us to correct obvious deficiencies as we proceed with further developments, and to produce a better program for wider distribution. Collaborators such as Lynn in Nakanishi's group at Columbia and Dreiding in Zurich would be among those selected for such an evaluation.

We will complement these evaluations with extensive tests among members of our own group and other SUMEX users who deal in one way or another with structure manipulations, including persons such as Wipke at Santa Cruz (SECS project) and members of the MOLGEN (molecular genetics) project at Stanford. We have begun evaluations in our own group, trying selected structural problems solved previously by CONGEN. In addition, we have in the past few weeks evaluated GENOA using as input substructures inferred from ^{13}CMR data of ethers and alkanes. ^{13}CMR represents a fertile area for GENOA applications because of detailed structural information which can be obtained on the environment of a given carbon atom. Obviously, such information is highly redundant.

C.2.b Integration of GENOA with CONGEN.

Initially, we will complete the development of GENOA with regards to topological representations of structure. Although the current version incorporates some considerations of symmetry of input substructures and of the representation of the problem itself, the construction procedure itself requires more development, primarily directed toward efficiency.

C.2.b.i Greater User Control of Overlaps.

The current version has an important limitation which needs to be overcome both for the proposed topological and stereochemical versions of GENOA. There is no control over which atoms in which substructures may (or may not) overlap, even though such information is frequently available from spectroscopic data. The flexibility of input permits one to work around some of these problems (for example, substructures known to be non-overlapping can be input to GENOA as a single substructure composed of several disjoint units), but a more general approach is required to prevent building undesired structures. We will accomplish this by associating with each atom a "uniqueness" tag which will declare that it is (or is not) to be considered in overlaps. Further work will permit one to allow designated sets of atoms to overlap with one another but not with another designated set. This capability will also be important in developing the automated inference/generation procedures described below.

C.2.b.ii Integration.

The next step will be to remove the generator/CONGEN separation of Figure 2 by integrating the output of GENOA with CONGEN itself to perform the final step of structure generation. The output of GENOA is a set of problems (see Figure 3), each of which is composed of non-overlapping structural fragments and can, therefore, be treated as a separate CONGEN problem. The combined results of all problems represent the complete and irredundant final set of structures. We will perform this integration by taking each individual problem from GENOA and casting it into the standard CONGEN superatom/atom framework.

This integration presents a number of potential difficulties which we have already recognized and will attempt to overcome. The set of CONGEN problems from GENOA may include problems which "contain" other problems. That is, the entire set of final structures from one CONGEN problem will be entirely contained in the final structures from another CONGEN problem. This sort of duplication will have to be recognized at this stage and the "contained" problem eliminated. Another potential difficulty is that some of the CONGEN problems will have identical superatoms and might give identical unimbedded structures after generation. These identical superatoms in separate problems will have to be recognized so that duplicates can be eliminated after generation but before imbedding. Some constraint information will have to be carried through and tested during generation and imbedding to free the user from having to input different constraints at different times. The final GENOA program will simply require as many statements (in a prescribed format) about the overall problem as the user wishes to input at the beginning and will apply the information throughout.

C.2.c A Stereochemical Version of GENOA.

The next step in development of GENOA will be to produce a version which generates structures consistent with both topological and stereochemical constraints. It is particularly important that GENOA have a full recognition of stereochemistry (both configurational and conformational). For many conceivable applications of stereochemical constraints it will be necessary to allow overlapping substructures. An example would be the substructures deduced from long-range couplings which will almost certainly contain substructures involved in short-range couplings. Data from ^{13}CMR will likely yield substructures which overlap those from proton NMR or CD spectra, etc. This effort to incorporate the stereochemical developments into the GENOA program will bring together two of the novel developments in our structure elucidation programs.

The exact method chosen for this development will probably depend on the current form of GENOA and the stereochemical programs at that time. However, a likely development would be in two stages. In the first version the user would input constraints with and without stereochemical information. The program would recognize the stereochemical constraints and save them until complete structures (topological) were generated. At this time the STEREO programs would be called and the constraints applied. After sufficient use and testing of this first version, the second version would be developed. The novel feature of this version would be the application of stereochemical constraints at the earliest possible time and in the most efficient way. This would necessitate some modification of the structure representation now used in GENOA and a "smart" constraints translator which would recognize those stereochemical constraints which affect possibilities for topological structures. An example of such a constraint would be the requirement of only trans double bonds eliminating structures with small rings containing double bonds. This final version would merge the stereochemical developments in such a way

that the user need not be concerned with any differences between topological and stereochemical constraints.

C.3 Development of Automated Approaches to the Exploitation of Spectroscopic Data.

Research will be undertaken into the development both of methods for deriving structural constraints for GENOA by automated spectral interpretation, and of methods for evaluating generated candidate structures by comparative analysis of predicted and observed spectral properties.

Spectral interpretation, to derive structural constraints, is appropriate only for those techniques in which, at least to a first approximation, spectral features can be correlated with fairly localized structural environments. Some spectrum/structure correlations are relatively weak; these can only be taken as suggestions, and not as absolute constraints. An example of such weak correlations would be any association of a particular group with an IR absorption below 1500cm^{-1} (for that region of the IR spectrum is usually dominated by combination and overtone bands due to vibrations of the entire molecule that render reliable interpretation difficult). Generally, inferences based on such weak spectrum/structure correlations should not be employed as constraints prior to structure generation but can be exploited when ranking structures subsequent to generation. In our proposed work on inferring structural constraints from spectral data, we shall be concentrating mainly on magnetic resonance techniques in which spectrum/substructure correlations are usually well defined.

Once structures have been generated, it is possible to exploit spectral data that relates more to the complete molecular structure than to any isolated subpart. We have already developed functions for predicting mass spectra using models, "theories", of how a given molecular structure might fragment [23,67]. Functions that predict magnetic resonance spectra, using a model of a complete molecular conformation, will in general be more accurate than those based purely on localized substructures, and consequently, will permit finer discriminations to be made between different candidate structures. Other post-testing of generated structures can exploit the suggestive evidence of spectrum/structure correlations that yield evidence for or against particular structural features but are not absolutely definitive tests.

C.3.a Carbon-13 Magnetic Resonance.

The chemical shift of a given carbon atom is sensitive to features of its local environment up to, and sometimes including, delta-neighbors. For some molecules, with rigid conformations, topologically more remote neighbors may also produce significant shifts through steric crowding. Although there is a great deal of information in ^{13}CMR data, frequently structure elucidation studies use the ^{13}CMR results just to determine gross features of the carbons in the structure such as their hybridization, degree of substitution and possible bonds to electronegative atoms. We intend to explore more constructive uses of CMR including both the inference of substructural parts prior to structure generation, and spectrum prediction and evaluation for complete, generated structures.

C.3.a.i Carbon-13 Spectrum Interpretation.

One aspect of the Meta-DENDRAL project in recent years was the development, by Mitchell and Schwenzer, of a system for the structural interpretation of ^{13}CMR spectra [16]. The Mitchell/Schwenzer system utilized rules that correlated particular resonance values with substructures; each rule involved a main prediction defining a fairly precise range in which a particular carbon in the substructure should resonate and a number of secondary and support predictions defining ranges in which the resonances should occur for the other constituent atoms of the substructure. An example of the type of rule used is shown in Figure 4. The rule is interpreted to mean that if a resonance is observed in the range 44.7-44.9, and if appropriate secondary/support predictions are satisfied, then the given substructure can be taken as a possible explanation for the resonance. These rules were abstracted from a set of spectra of standard alkanes and alkylamines and could be used for the identification of additional alkanes and amines not present in the training set.

44.7 ppm < delta(3) < 44.9 ppm =>

```
          1
          |
    5-4-3-2-7
          |
          8
```

Node	atomtype	secondary prediction	support prediction
1	C		27.1 < delta(1) <34.9
2	C	29.7<delta(2) <35.6	30.7 < delta(2) <33.4
3	CH2		
4	CH2	17.9<delta(4) <56.9	17.9 < delta(4) <27.6
5	C		15.4 < delta(5) <24.3
7	C		27.1 < delta(7) <34.9
8	C		27.1 < delta(8) <34.9

Figure 4. Meta-DENDRAL C-13 Spectrum Interpretation Rule

Structure elucidation was accomplished by determining the rules, and consequently the substructures, that could be associated with each individual resonance and then, in a complex search scheme, determining allowed combinations of these separate substructures. The Mitchell/Schwenzer system involved a heuristically guided, depth first search in which substructures were combined and expanded by identifying their allowed partial overlaps. The method has been illustrated through its analysis of a fully-decoupled spectrum of 3-3-dimethylhexane [16].

Methods of spectrum interpretation using data bases of ^{13}CMR spectra have been reported by both Bremser and Jezl [68,69,70,71]. These two systems allow for a number of search options; the data base can be searched in a conventional manner to identify reference compounds with spectra similar to that for an unknown, or individual resonances may be entered and the data base searched to retrieve all substructures showing similar shifts, or the data base may be searched to find the range of shifts associated with some given substructure (i.e. spectrum prediction). Bremser has illustrated how results of a standard file search may be interpreted

manually to yield information on substructural components of a compound not in fact present in his reference file.

In these two systems, the data characterizing each reference compound consists of atom-centered codes, (describing the topological environment of each carbon with an assigned shift), and associated shifts. Originally, both systems defined atom-centered codes that represented a tree-structure grown through the molecule from the central atom; tags were used in the Jezl system to indicate ring-membership and similar properties. These codes were ambiguous in that quite distinct, cyclic sub-structures could yield the same code. The Jezl scheme also suffered from the disadvantage that the codes were derived through relatively complex rules, handling many special cases, and the coding process could not be automated. The Bremser "HOSE" code was more simply defined as a Hierarchically Ordered description of an atom's Spherical Environment; generation of the code could be completely automatic. Bremser's original code had problems of ambiguity similar to those of the Jezl code; Bremser has subsequently modified his coding scheme but problems, such as truncation after a certain number of characters, remain.

We have started to develop a system for ^{13}CMR spectrum interpretation that is an extension of the second-type of search option in the Bremser and Jezl schemes. Our program takes, as input, the individual resonances of all carbon atoms in the unknown structure and searches a reference file to identify all atom-centered codes associated with similar resonances. Unlike the Jezl and Bremser systems, these retrieved codes are not the program's output, but are in fact the input to a more detailed stage of analysis in which the results for individual resonances are combined to yield much larger substructures. In essence, the subsequent processing steps correspond to the structure building procedures of the Mitchell/Schwenzer system; however, in this context it is possible to avoid some of their complexities relating to graph matching of possibly partially overlapping structures.

The current implementation of these algorithms is incomplete. We have systems for generating atom-centered codes and creating reference libraries of codes and shifts, and we have preliminary versions of functions for exploiting such reference libraries. The atom-centered code that we employ is a complete, canonical description of an atom's topological environment (out to a distance of three bonds). The coding system will be extended to include stereochemical information once such data has been incorporated into our standard structure definitions.

The following example is derived from data on isoterpinolene Figure 5:

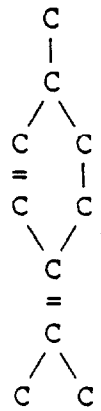


Figure 5. Isoterpinolene —— example structure for C-13 interpretation functions

The program requires the shift, multiplicity and allowed error in shift for every carbon in the structure.

```
| C-resonance : 21.4 4 0.5
| C-resonance : 30.6 2 0.5
| C-resonance : 133.1 2 0.5
| C-resonance : 125.0 2 0.5
| C-resonance : 127.7 1 1.0
| C-resonance : 125.7 1 1.0
| C-resonance : 25.7 3 1.0
| C-resonance : 31.7 3 1.0
| C-resonance : 20.5 4 0.5
| C-resonance : 19.6 4 0.5
| C13 reference file : crsncs
```

Figure 6. Spectral data input for Isoterpinolene

The program searches the file of substructures and associated shifts, retrieving those substructures with appropriate shifts and multiplicities. In this first pass through the reference file, each shift is considered individually.

```
| 3565 references read, 280 references written.
|
| allowed atom codes are :
|  -CH3  -CH2-  >CH-  -CH=  >C=  -C*H=  >C*=
|
| LINE TYPE .
| 1  -CH3
| 2  >CH-
| 3  -CH=  -C*H=
| 4  -CH=  -C*H=
| 5  >C=  >C*=
| 6  >C=  >C*=
| 7  -CH2-
| 8  -CH2-
| 9  -CH3
| 10  -CH3
```

(The symbol C* indicates an aromatic atom type). In a subsequent checks, through the reduced reference file, the program utilizes existing partial results regarding possible alpha-neighbors for each atom. Such checks are able to derive the following information:

```
allowed atom codes are :
 -CH3  -CH2- >CH-  -CH=  >C=

LINE TYPE
1  -CH3
2  >CH-
3  -CH=
4  -CH=
5  >C=
6  >C=
7  -CH2-
8  -CH2-
9  -CH3
10  -CH3

Current partial adjacency matrix:
      Hetero    1    2    3    4    5    6    7    8    9   10
    1      .    .    ?    .    .    ?    ?    .    .    .    .
    2      .    ?    .    *    .    .    .    ?    ?    ?    ?
    3      .    .    *    .    ?    ?    ?    .    .    .    .
    4      .    .    .    ?    .    ?    ?    ?    ?    .    .
    5      .    ?    .    ?    ?    .    *    ?    ?    ?    ?
    6      .    ?    .    ?    ?    *    .    .    .    ?    ?
    7      .    .    ?    .    ?    ?    .    .    ?    .    .
    8      .    .    ?    .    ?    ?    .    ?    .    .    .
    9      .    .    ?    .    .    ?    ?    .    .    .    .
   10      .    .    ?    .    .    ?    ?    .    .    .    .
```

Figure 7. Adjacency matrix derived for Isoterpinolene using Alpha-neighbors

(In the adjacency matrix, the "."s indicate no bond possible, "?"s indicate possibility of a bond and "*"s show where the program finds bonds to have been definitely established). The program has been able to determine that the none of the methyls can bond to the methylene groups, has established that the sole methine carbon must bond to one of the alkene units and identified the other alkene part. These restrictions could be input to GENOA (see Section C.2) and would considerably reduce the size of the structure generation problem. However, in this example, it was possible to derive further constraints by utilizing data on beta-neighbors provided by the retrieved codes. Using beta-neighbors leads to the following final connectivity matrix of Figure 8. The only uncertainty remaining concerns which of the methyl groups is attached to the methine and which are substituted onto the alkene Because the methyl groups exhibit different shifts, they are distinct within this program. As far as they are defined, these results do correspond to the correct structure.

Current partial adjacency matrix:

	Hetero	1	2	3	4	5	6	7	8	9	10
1	.	.	?	.	.	.	?	.	.	.	.
2	.	?	.	*	.	.	.	.	*	?	?
3	.	.	*	.	*	.	.	.	.	.	.
4	.	.	.	*	.	*	.	.	.	.	.
5	.	.	.	.	*	.	*	*	.	.	.
6	.	?	.	.	.	*	.	.	.	?	?
7	.	.	.	.	.	*	.	.	*	.	.
8	.	.	*	.	.	.	.	*	.	.	.
9	.	.	?	.	.	.	?	.	.	.	.
10	.	.	?	.	.	.	?	.	.	.	.

Figure 8. Adjacency matrix derived for Isoterpinolene using Beta-neighbors

The current program has several limitations (apart from the fact that the structure codes are purely topological with no stereochemical data). The major limitation is in the reference file which is derived from data on rather less than 500 compounds (standard alkanes, alkenes, terpenes, diterpenes and a few other natural products such as coumarins). Very frequently, the file does not contain any substructure equivalent, out to beta neighbors, to some feature in a new unknown compound. In such cases, attempts to utilize beta-neighbors will lead to a situation where no complete structure may be generated. However, matching alpha-neighbors does still provide a few usable constraints. We propose to obtain from other sources computer readable files of extensive collections of ^{13}CMR spectra. One such file is available through the NIH/EPA Chemical Information System. However, although most of the spectra are assigned, structures are not associated with the spectra. We are exploring ways to obtain the structures (in computer readable form) and integrate them with the spectra themselves. Bremser [70] has available a large collection of assigned spectra with associated structures. The charge is, however, DM 1.75 per spectrum (approximately 14,000 spectra are available). If we are unable to obtain the required data from the NIH/EPA system we will be forced to submit a supplementary application for funds to purchase Bremser's collection.

The current program is not interfaced to the rest of the CONGEN system, and some of the constraints that it can identify cannot be expressed in terms of the substructures that may be input to the current GENOA program. We propose to build those interfaces and to develop ways to translate constraints derived from the ^{13}C analysis into a form usable by GENOA. This work will initially relate only to the topological representations of structure. We will at the same time, however, begin a design of the required stereochemical extensions to both the interpretive aspects of the program and the interfaces to the structure generators.

C.3.a.ii Prediction of C-13 Spectra.

Several approaches to the computerized prediction of ^{13}CMR data have been reported. The CARBON-13 program of Munk, et al. [39] is concerned solely with the prediction of the number of distinct resonances that might be expected in the ^{13}CMR spectrum. Their prediction method is based on identifying topologically equivalent atoms with a few, ad hoc, extensions to identify anisochronous methyl groups in those diasterotopically related geminal methyls which are a commonly encountered feature of organic molecules. Our algorithms for determining the true stereochemical symmetry group would permit the correct enumeration of such resonance counts.

Other programs reported in the literature have been concerned with the

prediction of approximate resonance shifts for each carbon nucleus. Some attempts have been made to utilize additivity rules for ^{13}CMR spectrum prediction. Thus, Clerc has a program for estimating ^{13}CMR shifts of singly bonded carbons in acyclic structures containing combinations of non-cyclic functional groups [72]. Predictions for more complex cyclized structures generally require parameterized schemes devised especially for sets of related structures. While topological descriptors may suffice for certain applications, geometrical descriptors are also of value. Some advances in the computerized derivation of ^{13}CMR prediction functions, based on both topological and geometrical descriptors, have recently been reported [73].

Spectrum/substructure correlation rules of course provide another method of spectrum prediction. The spectrum prediction system of Mitchell [15] employs a slightly unusual representation of such a correlation table. In this system, 138 rules of the form "substructure -> shift range" were abstracted from a set of alkane and amine spectra. Spectrum prediction was achieved by graph matching each substructure to the given structure, and assigning to each carbon a shift range representing the common range of shifts predicted by each applicable rule. The interactive NIH ^{13}CMR system [74] can work in a somewhat similar manner; it differs in that the range of shifts, and details of their distribution, are derived from data base as required, rather than being abstracted into a "prediction rule". The substructure of interest is defined and then the program retrieves the shifts of all instances of that substructure in the current data base. (This has the advantage that expansion of the data base does not require re-analysis to derive new prediction rules). The search of the potentially large data base is made efficient by the use of fragment codes, bit screens etc; the final steps do, like Mitchell's system, require some graph matching.

Bremser has noted the use of his library of atom-centered codes and related shifts for spectral prediction; a paper with more details is due to be published. Essentially the same Mitchell/NIH correlation methods are used; atom-centered codes are generated for each atom in the given structure and the data base is searched to retrieve shifts associated with these codes. The graph-matching processes are here replaced by the string comparisons on the atom-centered codes. The Jezl/Dalrymple RACES system may be used for spectrum prediction in exactly similar fashion.

Currently, we have a similar spectrum prediction capability. We are choosing to concentrate at present on deriving substructural constraints from these spectrum/structure correlations, rather than explore spectrum prediction and ranking schemes. If the methods of interpreting ^{13}CMR data in terms of substructures are successful (and used to derive constraints for a structure generator), then spectrum prediction based on this type of model would not serve as a basis for further discrimination between structures. We will consider the use of more elaborate approaches to spectrum prediction that might be applicable once the CONGEN programs have been extended to provide complete conformational, stereochemical representations of structures. The conformations could be used as input to a force-field structure modeler and the resulting configurations analyzed to derive both topological and steric contributions to chemical shifts (as in the work of Smith and Jurs noted earlier).

C.3.b Proton magnetic resonance.

The chemical shift of a proton is, like a carbon nucleus, determined largely by features of its local environment. However, the factors determining the shift tend to be somewhat shorter range than those for ^{13}CMR. Not all influences on a resonance can be correlated with the immediate topological neighbors;

steric/conformation effects can result in protons experiencing strong shielding/deshielding due to topologically remote, but sterically close, pi-systems. Consequently, as with ^{13}CMR, much preliminary information can be derived, prior to structure generation, by interpreting ^{1}HMR in terms of local environments but sensitive discrimination between related structures may require spectrum prediction based on accurate models of molecular conformations.

Some attempts have been made to derive structural information from the coupling pattern exhibited in a proton spectrum either independently of [75], or in association with chemical shift data [76]. These methods assume well resolved, first order spectra. It is rare for compounds of bio chemical interest to show such simple spectral characteristics, and these interpretation methods are mainly of educational rather than practical interest.

Sasaki [41] uses conventional correlation chart approaches to derive structural data from the ^{1}HMR spectrum. He employs a table of about 200 standard fragments, each defined by the number of protons that should resonate in particular spectral regions. Analysis of a spectrum, by referencing this table, identifies maxima and minima for the numbers of groups of different types. For moderately large structures, the results from this type of ^{1}HMR interpretation are generally somewhat ambiguous; usually, many different combinations of substructures can be found that would satisfy these weak spectral constraints. However, even this simple interpretive process does capture the kind of negative evidence that chemists frequently fail to provide to CONGEN and other structure generating programs; the absence of required resonances will lead to the prohibition of the generation of certain substructures —— a prohibition that the chemist will typically only apply after seeing the generation of invalid structures.

We have made some preliminary experiments, within the INTERLISP version of CONGEN, on the use of additivity rules for the prediction of proton resonance spectra. This ^{1}HMR system consisted of functions for predicting the resonance values of protons attached to alkyl and alkene carbons, and functions allowing the user to define constraints on the number and type of protons in particular regions of the spectrum. The proton shifts were predicted by conventional additivity formulae using incremental shifts due to alpha-functional groups with some corrections for beta-groups and membership of small rings. The performance of this spectrum-prediction/structure-pruning system was erratic when applied to typical CONGEN problems. In some cases, more than 80% of generated candidate structures could be successfully eliminated using relatively weak constraints. More frequently, within the accuracy of the model, there were no significant differences in the spectra predicted for different candidate structures and pruning of the structure list was not possible. In a few cases, there were sufficiently large differences between the observed and predicted spectra of the true structure that pruning on the basis of what might have been assumed to be reasonable constraints would in fact have eliminated the correct structure. Prediction performance appeared to be poorest for protons on di- and tri-substituted alkenes; but, the scope and limitations of the additivity rules could not be clearly defined. Consequently, this proton magnetic resonance system was not made routinely available to CONGEN users.

Our current intention is to apply to proton NMR the same methods being investigated for ^{13}CMR. Priority is being accorded to ^{13}CMR in part because of the lack of suitable, machine-readable collections of assigned proton spectra.

C.3.c IR.

Spectral-structure correlations based on infra-red data have been exploited

in a number of other structure elucidation programs. Thus, both Sasaki [41,42] and Gribov [50] employ what are essentially correlation tables, while Munk has a system for inferring substructures by either "Pattern Recognition" or "Artificial Intelligence" IR-classification functions [37].

However, these IR interpretation schemes are limited. For example, Sasaki's system attempts merely to determine minima and maxima constraints on the number of oxygens in carbonyl, hydroxy and ether groups. Other groups with absorptions above $1500cm^{-1}$ might be characterizable by similar correlation rules. However, reliable identification is frequently not possible. Problems of interpretation result from extreme variability in intensity of "characteristic" absorption bands (e.g. the absence of a nitrile absorption in many alpha-hydroxy-nitriles), and from the fact that many of these absorptions lie in the same $1500-1700cm^{-1}$ region. Spectra below $1500cm^{-1}$ are frequently dominated by combination and overtone bands characteristic of the molecule as a whole rather than any isolated subparts; while such spectral data are valuable in file search oriented methods of identifying structures, they are of limited diagnostic value offering suggestive rather than definitive evidence. The same qualification —— suggestive rather than definitive evidence —— has to be applied to attempts to derive additional structural information from the particular position of absorption of a better characterized group such as a carbonyl. A carbonyl absorption around $1770cm^{-1}$ is suggestive of a five-membered lactone ring, but presumption of such a substructure, and its use as a constraint on a structure generator, would be ill-advised (similar absorption also characterize vinyl esters and several other substructures).

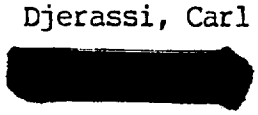
Since the identification of principal functionality has normally been completed long before a structure-elucidation problem is given to CONGEN, there appears to be relatively limited use for this type of IR-interpretive scheme in association with CONGEN/GENOA.

Gribov has worked on the evaluation of candidate structures using predicted IR spectra. The spectrum prediction is based on conventional approaches for analyzing the vibrational frequencies pf a molecule using parameterized bond strengths etc. As noted by Gribov, this approach is really only applicable to fairly small molecules. Conceivably, such an approach might be tried with the larger molecules analyzed by CONGEN.

The only use for IR currently envisaged is in a very simple ranking scheme. In the LISP version of CONGEN, functions were available that allowed the user to associate positive and negative scores with particular substructures and to let the program rank candidate molecules by combining scores associated with their constituent substructures. It is possible to implement a scheme that would derive scores for different substructures through the suggestive, but not definitive, IR-interpretation rules noted earlier. This would allow the association of plausibility values for each structure which might then be used with the simple ranking functions.

C.3.d Mass Spectrometry.

Interpretive processing of mass spectral data by means of spectrum/structure correlations is of limited utility. These correlations, e.g. (M-44 <=> anhydride), generally only provide weak evidence for, or against, the presence of functional groups more readily identified by other spectral/chemical techniques. This type of mass spectral processing is available as a minor component in Gribov's system, and may yet be incorporated in Sasaki's programs. However, it is not considered to be of value in the type of structure elucidation problem generally given to CONGEN.



The Mass Distribution Graph method does constitute a more general model for mass spectrum interpretation [67]. However, the applicability of the method to large structures is limited by the combinatorial nature of the algorithms. We have no plans currently to develop further this approach; in spite of the elegance of its conceptual base, a practical, useful program seems extremely difficult to develop.

Some of the more sophisticated mass spectral file search systems have the capability of identifying constituent substructures of molecules not actually present in the reference file [28]. McLafferty's group is exploring the possibility of using CONGEN in association with their STIRS file-based mass spectral interpretation system.

While methods for interpretation of mass spectra are still of limited applicability, our algorithms for mass spectrum prediction and structure ranking now seem well proven [67,23]. For mass spectral predictions, the use of a topological model of a chemical structure, as produced by current CONGEN, is generally quite satisfactory. The mass spectral processing algorithms can employ models, "theories", of fragmentation processes of varying degrees of specificity as may be appropriate to a particular application.

C.3.e Circular Dichroism and Magnetic Circular Dichroism

Circular Dichroism (CD) and Magnetic Circular Dichroism (MCD) are spectral techniques particularly useful in structural/stereochemical studies. Because of their strong dependence on stereochemistry, these techniques have heretofore not found much use in CONGEN and this represents a serious deficiency. With the addition of conformation information to CONGEN it will now be possible to incorporate structural and stereochemical information from these sources into computer-assisted structure elucidation using CONGEN and GENOA.

The interpretation and prediction of CD spectra based on substructural hypotheses is well established. Examples are the familiar octant rule originally derived for ketones [77], Brewster's rules for parrafinic hydrocarbons [78], and extensions to sector rules such as those of Kirk and Klyne [79]. In all cases the substructure considered must include both configurational and conformational information. The general method is to associate with each substructure an intensity (either positive or negative) which contributes to the overall observed intensity of the spectrum. The assumption is often made of additivity, but not always. Interpretation of CD spectra makes primary use of the observed sign and intensity of the spectrum. First of all, the fact that a CD spectrum was observed at all says that the molecule is chiral, which is a very useful constraint for CONGEN's structure generation. Rationalization of the observed intensity is best made with respect to possible structures, i.e., after a CONGEN generation of possible chiral structures and their conformations. The observed spectrum will not usually lead to structural hypotheses by itself since the (usually single) intensity can be caused by any number of substituents around the chromophore involved. The real use of the method is to rule out possible structures with specified conformation.

Useful, but somewhat coarse, predictions of CD and MCD spectra can be made with the use of a number of models proposed which correlate spectral intensity with substituents. Among these are the octant rule [80] and the "zig-zag" model which are most useful for the CD spectra of saturated ketones. A simple model has also been proposed for MCD spectra of saturated ketones [81]. These models generally are used for predicting spectra of structures having idealized geometries with substituents in nearby or well-defined locations with respect to the chromophore. These methods can therefore be applied to structures output from the proposed

conformation generator. We propose to develop a program which will allow us to discriminate between candidate conformations using these models and experimentally obtained CD or MCD spectra.

A more precise prediction of CD spectra requires a refined geometry and a more detailed energy calculation. If a flexible structure is in an equilibrium between two or more conformations, relative energies are needed to determine populations before computation of spectra. To make use of these methods, an interface between the output of CONGEN and the input to these already existing energy programs will be required. This work will be the second stage of this effort and will be guided by results of the first stage using the simpler models.

While CD spectroscopy in the ultraviolet region is long established as a useful tool for structure elucidation, recent work on MCD (Magnetic Circular Dichroism), VUVCD (Vacuum Ultraviolet Circular Dichroism), and VCD (Vibrational Circular Dichroism) indicate considerable promise for structure elucidation. While the methods of interpretation for these spectra differ, the end result is usually a correlation of substructure (including stereochemistry) with observed intensity and sign. We feel our proposed system for CD spectra will incorporate, with modest modification, substructural information from these newer methods as they become more common in biomedical structure elucidation.

C.3.f Combined Spectral Interpretation.

As we mentioned previously, structural information from different physical techniques is often complementary. Any program seeking to perform automated analysis of data from different techniques must eventually be capable of examining the data and resulting inferences from each technique in light of what has been learned from other techniques. We illustrated a very simple example of this approach in the example on interpretation of ^{13}CMR presented above. The assignment of specific substructures to observed resonances implies additional assignments, which can be made on the basis of further data analysis or logically, to cite the trivial example that assignment of one carbon to a C=C functionality means that another carbon must also be assigned to the double bond. We propose to generalize this method, perhaps using an adjacency matrix representation for the growing structure, with a program which acquires inferences from each technique and automatically searches data from other techniques for confirmation or denial of the inferred substructure. This approach will be complicated by the fact that some of the inferences will be tentative. We will evaluate schemes for assigning certainty factors to each inference to explore the possibility of generating structures with associated plausibilities (determined by appropriate combination of the certainty factors for each component part).

There is another useful application of methods for considering the collection of available data. As noted earlier, chemists using CONGEN frequently fail to specify all the structural constraints that, in fact, they know. For the most part, it is negative constraints that are forgotten. Because such constraints are neglected, some problems appear to be too large and many others are solved extremely inefficiently with the chemist generating and then pruning away large classes of structures. If some convenient method of entering spectral data can be devised, then relatively unsophisticated spectral interpretation techniques could be used to derive such negative constraints, effectively prohibiting the generation of particular substructures.

C.4 Conformation Generator for CONGEN.

A great deal of the information input to a structure elucidation problem is conformational in nature. Examples of such information are vicinal and long range couplings in NMR, steric shifts in ^{13}CMR such as the differentiation between axial and equatorial substituents, Circular Dichroism spectra, etc. At present, this information can be used indirectly in CONGEN only as it pertains to the constitution (bond-connectivity) or configuration (chiral centers and double bonds) of a proposed structure.

To eliminate this deficiency and enhance the value and scope of CONGEN, we propose to provide CONGEN with an exhaustive and irredundant generator of conformations based on a chosen discrete set of possible torsional states (i.e., positions of rotation) around rotatable bonds. The input to such a generator would be a CONGEN stereoisomer (specified constitution and configuration) and the chosen set of torsional states (which might be a default set contained in the program). The output would be the possible conformations based on these possible torsional states. Only torsional angles will be considered as variables for the conformation generation, the bond lengths and bond angles will be considered fixed.

While the conformation generator will indeed be exhaustive and irredundant, special attention will be given to common conformational situations such as six-membered rings. This will be done to improve the efficiency of the program for common problems while retaining the desired assurance for more complex problems (see Sec. C.4.b.i).

A conformation generator will provide the necessary link from CONGEN to existing computer methods which require input structures with coordinates. This link is shown schematically in Figure 9. Examples of such methods are molecular mechanics energy calculations, quantum mechanical energy calculations, graphics display programs, and finer grid (torsional angles) conformation generation. A conformation generator would also allow CONGEN to be applied in conformational analysis, another form of structure elucidation. We propose to develop this link to existing methods and applications.

We feel the proposed investment of resources into developing this conformation generator will be repaid many times over because:

1) The added versatility it will provide CONGEN to deal with all facets of conformational information in structure elucidation;

2) It will allow development of links to existing atomic coordinate based methods which permit many potential new biomedical applications. (See Sections C.4.C, F).

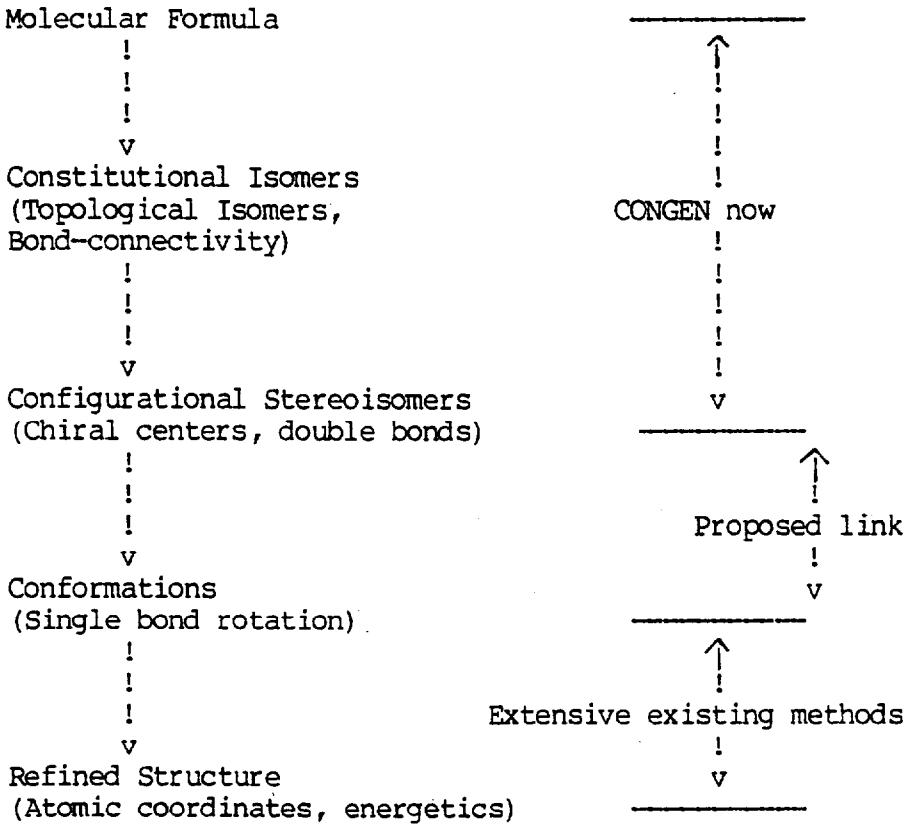


Figure 9. Proposed Link of CONGEN to existing coordinate-based methods

C.4.a Total Conformation Generation.

In its current state of development, CONGEN will generate, if desired, all possible configurational stereoisomers from an input molecular formula (Figure 9). This generation is complete and irredundant, that is, all possibilities consistent with chemical valence are included without duplication. We feel that complete and irredundant generation is one of CONGEN's strongest selling points as it assures no possibilities are overlooked in a structure elucidation problem. We propose to provide the same assurance for conformation generation. However, since there are infinitely many conformations possible for most structures, as torsional angles vary continuously, a complete and irredundant generation is possible only if a finite number of discrete values are allowed for the torsional angles. To accomplish the desired total conformation generation it will be necessary to develop:

i) a generation algorithm,

ii) a means of representing conformations,

iii) an assurance of irredundance.

C.4.a.i Generation Algorithm.

We propose to generate conformations for a CONGEN stereoisomer by labelling

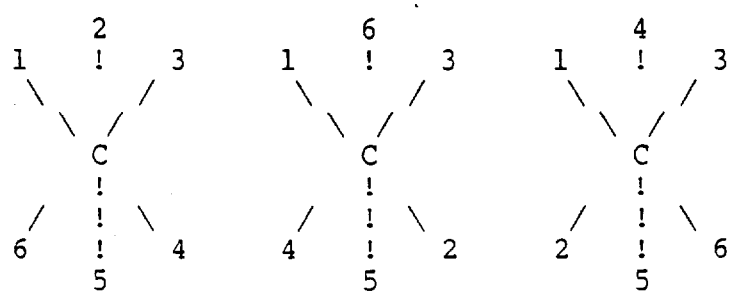
the bonds (edges) of the graph representing the stereoisomer with the possible bond torsional states. The algorithm will be developed so that any number of possible torsional states can be specified. The possible torsional states might be the familiar trans, gauche (+), and gauche (-) states which correspond to the minima of the torsional potential for single carbon-carbon bonds (e.g., the central bond of n-butane). Alternatively, the states might be the six (three staggered, three eclipsed) possible for a single carbon-carbon bond. The former choice might be made in cases where only local bond potentials are considered important, while the latter might be made when global potentials are important. The point here is that the algorithm would have to handle either case or any other choice of possible torsional states [82].

Part of the algorithm would involve recognizing the various kinds of rotatable bonds (e.g. sp3-sp3, sp3-sp2, sp2-sp2, etc.) and treating each accordingly. The problem of deciding which are rotatable bonds would be handled in a manner analogous to that used for determining stereocenters in the algorithm for configurational stereoisomer generation [25]. This method would start out by assuming every bond is rotatable and then proceed to reject those known to be fixed by intrinsic or input constraints (section C.4.b, below). Generation would proceed on the remaining bonds with further testing done as the conformations are generated. This method is similar to that used successfully for other structural generation problems in CONGEN.

We propose to do the generation of conformations as a labelling since our extensive experience with such algorithms suggests that these are very fast and efficient algorithms for problems of this kind. In particular, we believe that this method would be faster than an algorithm based entirely on atomic coordinates. The result of such a labelling would be a CONGEN structure represented as a graph with labels on some of the edges (bonds). The result of a generation of this kind would be a list of structures with bonds labelled in all possible ways and including structures not consistent with constraints (such as those imposed by intrinsic structural features such as ring closure). Labelling algorithms assure completeness since the labels are distributed in all possible ways. In our experience they are also constrainable, a particularly important consideration for conformations.

C.4.a.ii Canonical Representation.

In order to deal with conformations in the computer, some canonical (unique) representation for each conformation is required. We will likely construct such a canonical representation by making use of the atom numbering of the input stereoisomer. This is in analogy to the canonical representation used for the configuration of the stereocenters in the stereoisomer [24]. For example, consider the problem of uniquely describing the three staggered rotamers of a structure numbered as:



A method which would give unique representations for each of these three rotamers

would be to choose the lowest numbered substituent on each of the two atoms involved in the bond and express the bond state in term of them. In the example, this would be 1,2-gauche(+), 1,2-trans, and 1,2-gauche(-) respectively. The convention of designating clockwise rotation with a positive sign was used. If any or all of these were equivalent by symmetry, an ordering of the torsional states would be required, and the canonical representative would be the lowest. For example, if the ordering based on torsional angle were used: gauche(+) < trans < gauche(-), and all three were equivalent in the above case (because of equivalent substituents) the canonical representative would be 1,2-gauche(+). This method of canonical representation assumes that the numbering of the atoms was determined before the edge labels are added. This sequence corresponds to that which will be used in CONGEN since the atom numbering is done when the constitution (bond to bond connectivity) is determined.

C.4.a.iii Irredundance.

Assuring irredundance (no duplicate structures generated) in the labelling algorithm being proposed for conformation generation requires the proper use of the symmetries involved in the problem. Since CONGEN can produce structures of arbitrary symmetry, the conformation generator must be able to deal with any possible symmetrical structure. It is for this reason that symmetry must be dealt with so generally, not because of any claim of preponderance of symmetrical structures in biomedical problems. To our knowledge, this general problem of considering arbitrary symmetry in conformation generation has not yet been solved. The problems for which conformation generation has been treated either have no symmetry at all or symmetry of just one type (such as that in cyclic hydrocarbons). This problem can be solved by using a method similar to that used to irredundantly generate the configuration stereoisomers [24]. For the conformation generation problem, the symmetry group of the input CONGEN stereoisomer is represented by its effect on the possible torsional states for each rotatable bond. The equivalence classes of this group on the set of all possible conformations are the symmetry distinct conformations. A lowest representative of each class will be the canonical representative. This method will also readily give an enumeration formula for conformations which will be of use to the program since it will provide an independent check of the conformation generator and will allow faster computation of the number of conformations.

The method for irredundant generation of conformations will be illustrated on tetrachloromethylmethane $(ClCH_2)_4C$. The objective is to find the distinct rotamers of this structure with only perfectly staggered rotamers allowed. To do this it is necessary to represent the symmetry group of the structure by its effect on the possible torsional states of the rotatable bonds. The symmetry group of this structure is its Configuration Symmetry Group [24] and is isomorphic to the tetrahedral point group. This group is computed by the current stereoisomer generator in CONGEN. It is this group which must be represented on the possible torsional states of the bonds. This group has five "kinds" of symmetry operations: the identity, eight C3 rotation axes, three C2 rotation axes, six reflection planes, and six S4 alternating axes. Each of these five will be considered in turn to ascertain their effect on the possible torsional states. The identity operation has no effect on the possible torsional states. The C3 operations, however, do have an effect. Each C3 axis is colinear with one of the C-C bond axes. Rotation about this axis has the effect of interchanging the three possible torsional positions about that axis. It has no effect on the torsional positions about the other three axes. Each C2 axis interchanges pairs of bonds and the torsional states around each. With this much information it is possible to compute the number of possible rotamers using the method of Kerber [83]. The identity operation can be expressed

by the permutation (1)(2)(3)(4) of atoms. The C3 operations can be expressed as permutations like (123)(4). There are eight of these. The C2 operations can be expressed as permutations (12)(34) and there are three of these. The number of rotamers can be computed by substituting 3 for each orbit in the permutation and multiplying by the number of permutations, summing over all types of permutations and dividing by the order of the group. This is done as stated for the identity and the C2 axes. However, because the C3 axes interconvert the possible rotamers about one axis, this term contributes zero to the total. This conclusion can be reached without rationalization by simply using the formula derived by Kerber [83]. This yields (1/12)*((3)*(3)*(3)*(3)+3*(3)*(3)) =(1/12)*(81+27)=9 which is the correct number of rotamers for this structure [84]. The number of enantiomeric pairs can be computed by following through the procedure for the other symmetry operations. This procedure has computed the number of equivalence classes for this permutation group. A generation algorithm can be developed by simply taking each possible rotamer in turn and letting the symmetry group represented in this manner act on it. This process will actually construct the equivalence classes and the lowest member of each class is the canonical representative. While this is not the most efficient algorithm, it serves to illustrate the method.

C.4.b Constrained Generation of Conformations.

As in the previous efforts to develop generators of constitutional isomers and configurational stereoisomers, once a total generator is devised, it is necessary to modify it so that constrained generation can be done. We propose to develop a constrained generator of conformations in this manner. Constraints relevant to conformation generation include 1) intrinsic structural constraints such as that imposed by ring closure, 2) constraints input by the user based on partial structural information, and 3) constraints imposed by dynamics since most chemical structures exist in several conformations in rapid equilibrium.

C.4.b.i Intrinsic constraints.

The intrinsic constraints on conformation generation imposed by ring closure in polycyclic structures are probably the most difficult problem facing a conformation generation program of the type proposed here. Not all possible values of torsional angles for bonds are consistent with ring closure in these cyclic systems. For a polycyclic structure, the conformations possible for each ring are constrained by the conformations of adjacent rings, in addition to the constraint of ring closure. The relative constraints of overlapping rings can be determined by establishing the overlapping bonds and the configurations of any stereocenters involved. For example, in cis-decalin the common bond to the two six-membered rings has the same signed torsional angle (with respect to the ring carbons) in both rings. In trans-decalin, the common bond has opposite signed torsional angles in the two rings. For a given polycyclic structure these relationships can be computed by finding all rings in the structure and establishing the relative configurations of the stereocenters involved. The ring-finding can be done using a currently available function in CONGEN. Hence, the general problem reduces to that of determining the possible conformations for a single ring with constraints on the values for the torsional angles of some bonds. The ring-finding would have to be done only once for all the stereoisomers of given constitution. This is an efficiency because the list of structures coming out of CONGEN has all stereoisomers of the same constitution together. Another efficiency would be to consider only nonenveloping rings (those which do not "contain" smaller rings).

The process of generating conformations for a single ring with constraints

will be the "unit operation" of the conformation generator and will have to be optimized for efficiency besides being complete and irredundant. Six possible methods can be suggested based in part on the work of others.

1) The most direct method would be to systematically vary each bond in turn and check for end-to-end distance and overlapping atoms trigonometrically using coordinates. This would be a depth first generation with pruning when atoms overlap or the end-to-end distance becomes to large to allow closure.

2) A variation on this approach successfully used by Barry [64] would be to vary "fold axes" or dihedral angles which have as their axis the line connecting two nonbonded atoms. This method is more efficient because ring closure is more easily maintained but is not always exhaustive for one set of fold axes. The same distance checking would be done here.

3) A different method would be to use internal coordinates which measure the "pucker" of the ring and directly take account of the ring symmetry. These coordinates are usually based on displacement from a reference plane of the ring but could probably be expressed in terms of torsional angles. This method would probably have to be modified to assure exhaustion for large rings. This method also requires coordinates, but in a somewhat simpler fashion [58,60,61].

4) Another method makes use of ideas developed by workers in conformational analysis [85]. Families of conformations can be differentiated by the sequence of the signs of the bond torsional angles. For example, the chair conformation of cyclohexane has six torsional angle sign changes as one goes around the ring. The boat-twist family has four torsional angle sign changes. Similar patterns are evident for families of conformations of larger rings. This observation suggests that conformations might be generated by first discerning such families and then generating within each family. The families might be generated by first labelling the atoms with locations of torsional angle sign changes (only an even number of these are possible for any size ring) and then labelling the bonds with the values of the torsional angles. The first labelling reduces the number of possibilities for the second labelling. This is a reductionist approach which resembles the vertex graph method used so successfully in the first version of the CONGEN cyclic structure generator.

5) A related labelling method would make use of the polygon classification method for conformations developed by Dale [56]. Conformations are considered as polygons made of straight chain edges of varying lengths. This has the effect of reducing a ring of n atoms to a polygon of $m<n$ sides. Generation with a method like this would involve generating the possible polygons and labelling the edges in all possible consistent ways with numbers of bonds. This is also a reductionist approach similar to the one above.

6) A different method would be to make use of a "catalog" of the possible conformations for each ring size with a backup generator for cases involving rings larger than those in the catalog. Each ring in the structure would then be labelled with each of the possible conformations consistent with any constraints. This method effectively trades faster runtimes for a larger storage requirement.

These six possibilities have been ordered by their probable speed. The choice of one of these, a combination of them, or another method entirely will depend on speed, programmability, and constrainability along with an assurance of completeness and irredundance.

Particular attention will be paid to common conformational features such as six-membered rings. This will probably be done most efficiently by making use of a "catalog" of six-membered ring conformations (item 6 above). The possible six-membered ring conformations will be generated by reference to the catalog and any current constraints. This method will simplify the interpretation of conformational observations in this common case and will easily permit use of notions most familiar to chemists (chair, twist, axial, equatorial, etc.).

Other common intrinsic constraints involve rigid substructures such as multiple bonds and three-membered rings. These features are easily recognized with currently available parts of CONGEN which will simply be used in the conformation generator as well.

C.4.b.ii Input Constraints.

These are the constraints which will be input by the user when solving a particular structural problem. Common constraints of this type will arise from interpretation of NMR coupling data (vicinal and longrange), steric effects on C13 NMR (e.g. axial vs. equatorial substituent shifts) and interpretation of CD or ORD spectra (sector rules, Brewster's rules, etc.). Such constraints will be expressed as desired or undesired substructures which include designations for absolute configurations and bond torsional angles. They will be dealt with in a manner similar to that in CONGEN now, that is, by graph matching and pruning. Particular interest will be directed toward expressing constraints dealing with CD spectra because of ongoing efforts in Prof. Djerassi's research group on this topic and because of recent developments in new methods such as Magnetic Circular Dichroism (MCD) [86], Vacuum ultraviolet Circular Dichroism (VUVCD) [87] and Vibrational Circular Dichroism (VCD) [88]. The latter (VCD) is particularly interesting because this method will likely provide direct evidence about individual chiral environments of many chromophores. Some or all of these new methods will be particularly useful to structure determination of biomedically relevant molecules because almost all such structures are chiral. Another type of input constraint arises from observations about symmetrically equivalent atoms. The symmetry of conformations will always be less than or equal to the symmetry of the configurational stereoisomer from which they are generated. Thus, to take proper account of observations about the symmetry of a unknown structure, consideration of the possible conformations is crucial.

C.4.b.iii Dynamics.

A different sort of constraint arises in the conformation generation problem since most flexible molecules are in rapid equilibrium among several different conformations. Thus a structure determination of the type discussed here which includes conformations may lead to several final structures rather than just one. This sort of information could be expressed as a constraint which requires at least n conformations or requires only one be present. Alternatively, there may be an observation which requires that there be interconversion between two known partial substructures (chair-chair interconversion in cyclohexane for example). To make use of this information it will be necessary to predict flexibility in conformations. For certain situations such a prediction is fairly simple. For example, most acyclic substructures are flexible if not too heavily substituted. In many structure determination problems there may be no interest in flexibility of acyclic substructures, hence these could be recognized and ignored in further conformation generation leading to a savings in time and storage. Other kinds of flexibility which are fairly easily recognized involve ring pseudorotation, large unconstrained

rings or parts of rings, chair-chair interconversion, etc. However, predictions of flexibility in this way can only go so far since this property depends strongly on energetics. A typical structure elucidation problem might end with the observation of several final structures and the question of whether or not they can be interconverted and their relative populations. Resolution of this question would require an energy calculation of some type. The conformation generator will allow computation of all the possibilities, an important piece of information in problems of this kind.

C.4.c Interface and Applications.

CONGEN Interface.

The conformation generator program will eventually become part of the CONGEN program. Conformation generation will take place after configuration stereoisomer generation. Information will flow to the conformation generator which describes the constitution and configuration of the stereoisomer for which conformation generation is to be done. The conformation generator will construct a list of possible conformations subject to input constraints and will return information about continued candidacy of the input structure in the ongoing structure elucidation problem. Thus it may be possible to eliminate a stereoisomer or even a constitutional isomer from further consideration based on results of the conformation generation. The user interface will resemble that currently in CONGEN and in fact will use many of the program sections already written for this purpose.

Other Applications.

Besides improving CONGEN as a tool for structure elucidation, this proposed addition of a conformation generator will lead to potential new biomedical applications. The CONGEN program with a conformation generator will output structures complete with internal (torsional angle) coordinates. Since many existing programs require as input a structure with coordinates, the opportunity exists to interface CONGEN with these programs. Examples of such programs are empirical force-field energy calculations (molecular mechanics), quantum mechanical energy calculations (probably semi-empirical), graphics programs, and finer grid conformation generation (smaller torsional angle increments with van der Waals checking). These programs and methods frequently see biomedical applications; an example is in the determination of structure-activity relationships. By interfacing CONGEN to these existing programs, the opportunity for new biomedical applications will be expanded. Since these programs have been extensively developed by others, it will not be necessary for us to spend the time to develop them. Instead, we propose to use collaborative efforts to take advantage of these very interesting new applications. (See Section F, Collaborative Arrangements, section C.5.d Graphics Interface).

Another application would be to the field of conformational analysis. The conformation generator would be useful to such efforts by exhaustively and irredundantly generating the possible conformations which must be considered in such a study. We propose to explore this possibility collaboratively as well (Section F).

The conformation generator program will also be written to exist by itself (besides being a part of CONGEN) by providing a mechanism for inputting structures

from other sources. Since we expect this program to be fast and efficient and able to deal with large lists of structures, it is conceivable that the conformation generator could be used on lists of structures from other data bases to "upgrade" a list of structures which do not yet contain any conformational information. This could lead to applications in pharmacology or toxicology and other areas which make use of large chemical structure data bases.

C.5 Resource Sharing.

C.5.a Access to Programs via SUMEX and Local Dedicated Computer and Resource Management.

In Section A.2, Background, we discussed the relationship of our project to the SUMEX resource. In that discussion we outlined the conflicts between program development and extensive "production" use of resulting applications programs by collaborators both within and outside of the SUMEX community. We propose to resolve these conflicts by providing separate machines for development and production uses, each available to local and network users.

We propose that program development take place, as it has in the past, on the SUMEX PDP-10 system. This system provides the requisite facilities for languages, editors, message handling and so forth, to support effectively such development. In addition, SUMEX will provide the gateway for access to production use the proposed dedicated computer (see Figure 10). In the past, few collaborators have participated directly in program development. Their contributions have been, for the most part, indirect in that many suggestions resulting from trial use of CONGEN were incorporated by our group into programs, resulting in improved performance and greater chemical "intelligence". Now that more chemical and biochemical research groups are becoming sophisticated in their use of computers, we expect that, during this proposal period, our collaborators may desire more direct involvement, Cowburn at Rockefeller for example. We will encourage such collaboration and will carry out such work on SUMEX.

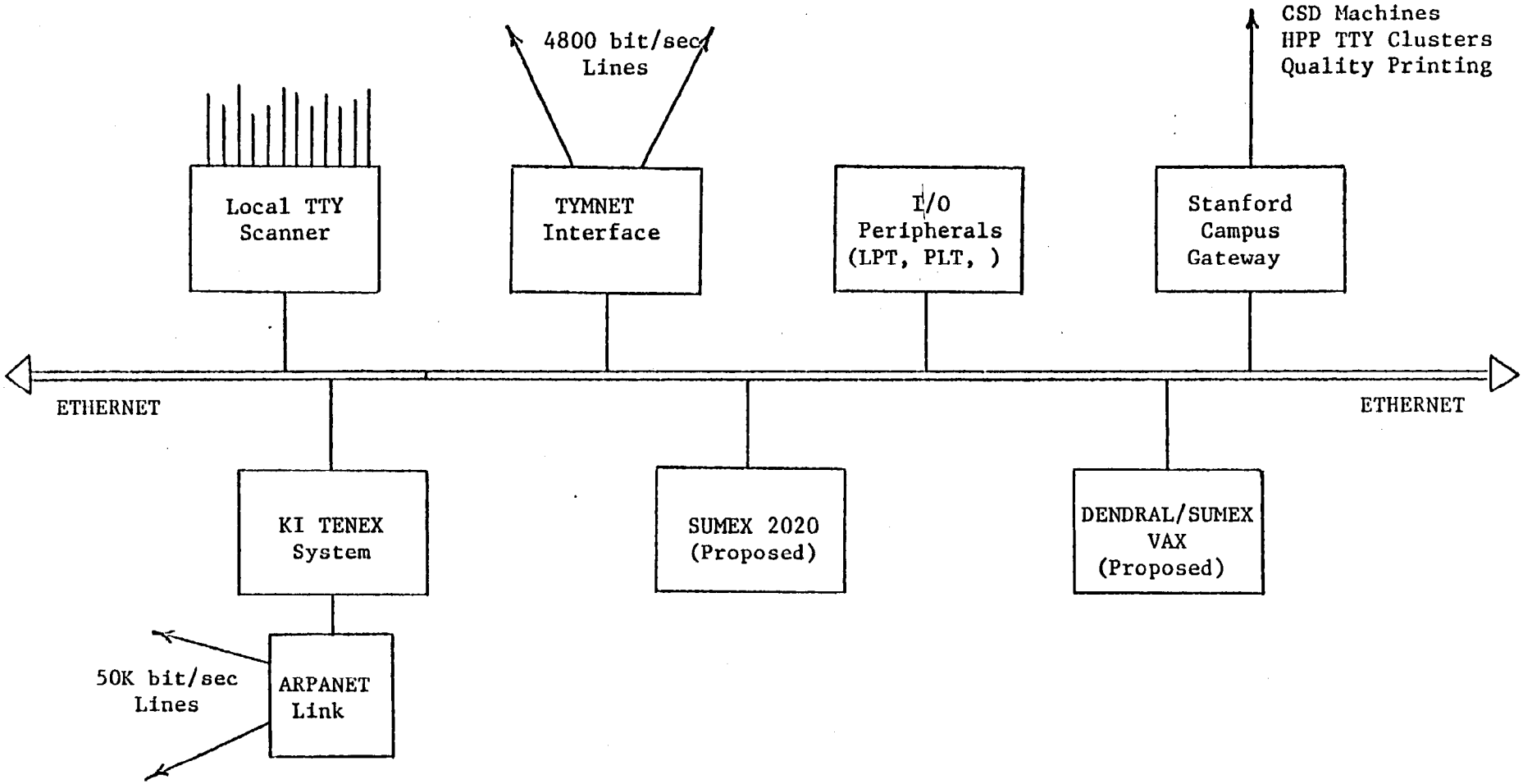


Figure 10. Access to DENDRAL programs on SUMEX

The production machine will provide more routine access to programs whose development and testing has progressed to the point where outside use of the programs by a wide community can be encouraged. Currently, CONGEN and auxiliary programs or functions, (STEREO for unconstrained generation of configurational isomers, SURVEY for examining structures, MSANALYZE for mass spectral prediction and ranking and REACT for simulation of chemical reaction sequences) are in this category. This use of our programs may involve local users, i.e. persons in our own group or local collaborators, or remote users accessing programs via TYMNET and SUMEX. We propose that access for applications, or production, use be handled by a computer linked to, but separate from the SUMEX system, specifically a Digital Equipment Corporation VAX-11/780. The schematic diagram of Figure 10 conveys our proposed design to allow users to access either machine, taking advantage of all existing network communications facilities on SUMEX. Whether computations will be carried out on SUMEX or the VAX will depend on the users and the program to be run. Only SUMEX users involved in DENDRAL will be enabled to run on the VAX and vice-versa. By exercising such administrative controls we will assure allocation of the DENDRAL machine for biochemical work while maintaining flexibility for network mail routing and approved file transfers. The VAX will have its own disc system, for system programs and page-swapping in the time sharing system of the VAX. The network configuration does not make the DENDRAL VAX dependent on SUMEX being up for access, yet still takes advantage of existing SUMEX communications hardware and other selected peripheral equipment (e.g., printer, plotter, TTY ports) without duplicating them.

We have discussed previously (see Budget Remarks) the reasons for our preliminary selection of the VAX computer. Briefly, in our opinion, it represents an optimum trade-off of several factors including interactive computing, large address space, compatibility with the biochemical and computer science communities with which we interact, long-term trends within the computing industry and cost-effectiveness. Several of the persons attending our workshops last year (see Annual Report, Appendix I) have, or soon will have, a VAX system. The National Resource for Computation in Chemistry will soon obtain a VAX. Taken together, these factors point towards the VAX as at least a good representative of the future shape of scientific computing appropriate to our applications.

The implementation of the VAX system as a part of the SUMEX resource, to be used as discussed above, will involve primarily the efforts of the two scientific programmers identified in the budget, under the guidance of the scientific staff and the Advisory Committee (see below). Their responsibilities, outlined in the Budget Remarks, are discussed here in more detail. We identify these programmers in the following discussion as befits their responsibilities, "systems programmer" and "applications programmer." Although we will be able to call upon the advice and expertise of the existing SUMEX staff in the implementation and continued maintenance and support of the VAX, in our opinion a full-time systems programmer and a full-time applications programmer (concerned primarily with the VAX) represent the minimum staffing required to support our VAX system. Initial implemetation of the VAX will include the following:

i) select either DEC's, or Bell Lab's UNIX, operating system (the responsibility of the entire staff).

ii) together with the SUMEX staff, make the necessary operating system modifications to allow access as outlined in Figure 10 (systems programmer).

iii) obtain, (or write) a BCPL compiler for the VAX in order to run the current CONGEN, GENOA and related programs written in BCPL (applications programmer).

While (i-iii) are being carried out, SUMEX will be used for applications under community controls in effect at the present time (GUEST access, strict control over SUMEX resource allocation (computer cycles) for local users). The applications programmer will coordinate this access and provide required documentation.

Once the BCPL compiler has been brought up, our next step will be to put the current production versions of CONGEN and GENOA on the VAX. At this stage we will have full availability of these programs and the system to the outside community, thereby relieving SUMEX of the bulk of the current production use (only programs in INTERLISP and experimental versions of new programs open for testing will be on SUMEX at this stage) (applications programmer).

At this time we will schedule a workshop, aggressively inform outside users of the system's availability and encourage them to thoroughly test the current versions of the programs and the VAX system. Based on our past experience with workshops and outside users we expect, because of a responsive system and many newly enthusiastic users (workshop participants), to have an active community of outside collaborators using our programs on the VAX. We expect this situation to occur in approximately late 1980 to early 1981 depending on the delivery time of the new system.

Following this initial period we will enter a phase of stepwise development of new programs on SUMEX, experimental testing of new applications programs by selected outside collaborators on the VAX and, finally, community-wide announcement of production versions of these programs on the VAX. (The scientific staff will be responsible for most of the new program development. The applications programmer, with the aid of the scientific staff and the systems programmer, will have important responsibilties in the resource sharing of experimental and production versions.)

Future DENDRAL program developments on SUMEX will be in languages for which compilers exist, or will soon exist, on VAX e.g. PASCAL, FORTRAN, MAINSAIL and BCPL. Obtaining these languages, and other system support packages such as text editors, will be the responsibility of the systems programmer. It should be noted that these efforts plus maintenance of the operating system and ancillary programs is a full-time, continuing job for a system of the complexity of the VAX.

There are proposals for an INTERLISP on VAX, which is probably a year or more away from implementation. Therefore, DENDRAL programs, such as REACT, MAXSUB and so forth, which are in INTERLISP and which we do not propose to convert to an exportable, algorithmic language, will remain accessible solely through SUMEX until an INTERLISP for VAX is available.

We plan, with the aid of the programmers requested in the budget, to implement on the VAX a number of chemistry-related programs which will directly support our new research and the work of our collaborators. We are particularly interested in molecular modelling programs such as MMI, CAMSEQ and PCILO, as an example. We plan to modify or adapt OMNIGRAPH to simplify our program developments for graphical input and output of structural information. We will work closely with others possessing VAX systems in order to share software and avoid duplication of effort as far as possible (e.g., VAX systems are being obtained by the National Resource for Computation in Chemistry with whom we are in contact, and Dr. David Pensak at DuPont who has been a collaborator with us in the past).

In a sense, we will be acting as brokers of some chemistry-related programs for the new generation of computers such as the VAX, for which little applications software exists currently. Our responsibilities as brokers would extend only to providing network access by qualified collaborators using our programs on the VAX,

and exporting applications programs to other VAX or VAX-compatible sites working on biomedically-relevant problems. We do not plan to compete with the quantum chemistry program exchange (QCPE), which distributes primarily programs that perform numerical computations. Our programs are primarily involved with symbolic computations. Also, we specifically do not intend to embark on language development for the VAX (e.g., INTERLISP), or to implement software which is not directly related to our goals or the goals of our collaborators. Although we may well utilize such programs and languages, such efforts are best done by other groups. We will monitor these efforts closely in order to take advantage of new software. Further into the grant period (two to three years) we expect that VAX and similar machines will be widely utilized within the biomedical community and that, as a result, we will be able to exploit the software available from the work of others.

Resource Management

To promote an orderly implementation of the dedicated computer system and to ensure that the respective goals of SUMEX-AIM and DENDRAL are being met, we propose to establish an Advisory Committee. This Committee will consist of the following persons:
Professor Edward Feigenbaum (chairman) - Principal Investigator - SUMEX
Professor Carl Djerassi - Principal Investigator - DENDRAL
Thomas Rindfleisch - Resource Manager - SUMEX
Dennis H. Smith - Co-Investigator - DENDRAL
Bruce G. Buchanan - Adjunct Professor, Computer Science - DENDRAL,MYCIN

The persons making up the committee are those with primary responsibilities for the conduct of both the SUMEX and DENDRAL efforts. They represent complementary knowledge and expertise, with Feigenbaum, Rindfleisch and Buchanan providing guidance on the computer science aspects of our research and on specific questions of the hardware and software interface with SUMEX and the dedicated machine, while Djerassi and Smith bring to the committee a strong emphasis on specific chemical and biochemical research problems of our group and those of our collaborators. Persons named above have worked extremely closely in the past (Prof. Feigenbaum, for example, was at one time PI on the DENDRAL grant); this committee will formalize an already existing close working relationship and focus attention on the specific area of resource sharing of our results.

This committee will advise on: a) planning the purchase, installation and development of the dedicated computer system; b) planning the hardware and software interfaces providing access to either the SUMEX or DENDRAL machines; c) assigning priorities for development of applications software on the VAX; d) promoting resource sharing activities such as lectures and workshops; and e) ensuring equitable access to the computer resources at Stanford by our collaborators. This committee will meet at two week intervals during the initial phases of the grant, and once per month thereafter.

C.5.b Exportable Programs.

To meet one of the goals of our present research, we have recently completed and begun distribution of an exportable version of CONGEN (see attached Annual Report). As we develop GENOA and the complete SASES system, we will ensure that they retain at least the same degree of exportability. The proposed version for the VAX computer will certainly improve the likelihood that outside scientists will be able to run the program in their own laboratories.

An intriguing philosophical question is why produce exportable versions for different machines when the program can be run at one site, accessible nationwide by computer networks. There are some practical reasons for this. Industrial firms require a high degree of secrecy for key structural problems and feel strongly that, to retain control and guarantee secrecy, they need a version operating on their own computers. Some academic laboratories have free access to mini/midi computer systems and cannot afford communication and commercial network/computer costs. When these practical reasons are exhausted, there remains a significant group of persons who simply want a version on their own machine under their own control

There is, however, another practical reason and that is, despite our efforts over the past few months, we have been unable to provide even experimental network access to CONGEN except the restricted access at SUMEX. There are two likely sites to which CONGEN can in principle be exported, both of which are easily accessible by computer network. The first is the NIH/EPA Chemical Information System (CIS), which operates a PDP-10 system on which CONGEN in its current form can be run without modification. For a year of more we have been negotiating, with persons responsible for CIS, for funds to support the conversion required to integrate CONGEN into CIS. Such funding has, we understand, received favorable Division of Research Resources/NIH review. However, we have recently learned that such funding will not occur now, and the future seems to be in question. It is unlikely that this situation will be resolved before the current proposal is submitted.

Another possible site is the Control Data (CDC) 6600 system, accessible at the Lawrence Radiation Laboratory (LBL) in Berkeley, through the National Resource for Computation in Chemistry (NRCC). We have arranged for some trials of the interactive BCPL system existing there to explore how CONGEN might operate in that environment. We do not expect these experiments to be fruitful because CONGEN relies heavily on intermediate files (to and from which structural information is communicated many hundreds of times during a problem); access to the file system is one of the primary bottlenecks of the LBL CDC system, access times being up to many seconds during heavy use of the system. However, the imminent purchase of a VAX system by the NRCC would change the situation dramatically. This factor, plus the uncertainties of the CIS lend considerable emphasis to our desires to implement readily-accessible versions of our programs on our own VAX system, with distribution to a wider community of persons handled eventually by the NRCC or the CIS.

C.5.c Workshops.

Our successful experience with workshops on the use of the new CONGEN program prompts us to propose a continuing series of such workshops designed to serve purposes similar to the previous effort. We want to introduce new collaborators to applications programs as they are developed, and to promote new applications and wider use of the programs. We want both new and current collaborators to evaluate new programs so that bugs, and clumsy or unclear aspects of the interaction can be resolved before development of the final version for application. Most importantly, these workshops will provide a mechanism for keeping our group up-to-date on requirements of the community for computer assistance in their efforts and provide a far more diverse set of structural problems than we would encounter in work here at Stanford. We have budgeted funds for one workshop per year, to be organized by our personnel and held here at Stanford University (or possibly elsewhere, the network providing access to the system).

C.5.d Graphics Interface.

We have invested a great deal of time and effort in the interface to CONGEN (and GENOA) with the lowest common denominator terminal, a teletype!, in mind. This proved to be a wise decision because many collaborators have only teletypes, or teletype-like hard copy terminals with which to access our programs. However, many structure drawings suffer in teletype form. Some can simply not be laid out within the confines of the teletype grid. In addition, the interface lacks characteristics enjoyed by most chemists, namely the capability actually to draw structural information. Elegant graphics interfaces have been produced for other systems, including Corey's [89] and Wipke's [90] organic synthesis programs, and the NIH Prophet system. In fact, OCONGEN offers capabilities for structure output to a variety of graphics terminals, taking advantage of NIH's Omnigraph package. Two things are clear from this experience. Few potential users can afford currently available terminals, such as the GT-40 series, even though graphical input and output would add considerably to the perceived utility of our programs. It would be out of the question to produce a useful graphics package for a wide variety of terminals, most of which will soon be obsolete.

Yet, we can hardly emphasize three-dimensional aspects of molecular structure and not provide capabilities for visualizing the resulting representations. At Stanford we can utilize the GT-40 purchased under the current grant for some of our work. We do not, however, wish to put a great deal of time and effort into interfaces for that terminal. It is nearly obsolete and compatible, newer versions are simply too expensive for mot research groups. We have requested funds to purchase one new graphics terminal in each of the first two years of the proposed grant. The terminal market is changing rapidly with plasma displays, storage displays, and raster displays, potential alternatives to the refresh displays such as the the GT-40. We will seek to purchase a display which is likely to become widely available at relatively low cost and invest our programming in that display.

D SIGNIFICANCE

There are several aspects of our proposed research which we feel are novel and especially significant including not only specific computer programs but also the methods by which they are shared.

One significant aspect of the proposed work, and a primary novelty, lies in the comprehensive computer treatment of both topological and stereochemical aspects of molecular structures in methods to assist scientists in elucidation of important biomolecular structures. By developing these method to include stereochemistry, we will have extended our approaches to computer-assisted structure elucidation to cover the key missing link of determining the actual spatial relationships of atoms in a structure rather than their mere connectivity. Our proposed methods for data interpretation and prediction applied to data collected on unknown structures offer several new treatments of topological representations of structure; the proposed stereochemical efforts, however, are certainly novel because no other similar system for structure elucidation utilizes stereochemistry in comprehensive form. The efforts are also necessary given the strong dependence of spectral properties on molecular configuration and conformation.

The proposed generator of conformers will be a novel solution to a long-standing problem in structural chemistry. Successful completion of a general constrained generator of conformation will be an important result in itself. However, its applications both to candidates for an unknown structure and to studies of conformations of (topologically) known structures will be a more important and novel result. In addition to the proposed collaborative efforts, we foresee several applications to problems relating structure to observed properties, including structure/biological activity relationships. We have not proposed such studies ourselves, but will collaborate with others who can make use of our results, under the resource sharing aspects of our proposal. The investment of resources into developing this program will be repaid many times over by increasing the versatility of CONGEN (as a tool for structure elucidation) and its scope of potential biomedical applications (by providing a link to existing methods based on atomic coordinates).

The GENOA program development, culminating in the SASES system, represents a general, and novel approach, to construction of structures with overlapping substructures, eventually constrained not only by topological but also by stereochemical constraints. These programs will be novel in:

a their modularity, so that they can be used alone or in concert with related programs, of our group or of our collaborators, via shared files of data,

b their comprehensive treatment of stereochemistry,

c the close invovlement of the scientist in the problem solving processes of the programs.

The last point perhaps deserves some more emphasis. The most interesting novel result of the proposed work lies in the use to which a well-designed system can be put in the hands of a well-trained chemist. The synergism of man and machine can be a powerful problem solving combination. The structure manipulations embodied in the SASES system represent a "dry" laboratory of functions for manipulating structures and associated data. Many aspects of structural chemistry besides the central task of structure elucidation can be studied utilizing component parts of SASES. In this way, complex questions can be posed and answered on the computer, before execution of actual laboratory experiments.

[redacted]

We feel that these features of our proposed research offer scientists capabilities for solving structures more accurately (in the sense of ensuring that all plausible alternatives have been considered) and in less time. In an era where detection and identification of trace-level organic compounds in environmental and biological milieus is of critical importance, new capabilities such as we propose can be of tremendous value.

We feel that the interest shown by structural chemists in our work, especially the CONGEN program, is already significant. We can point to the workshops, to persons requesting access to the programs at Stanford and to persons interested in the exportable version of CONGEN (see Section F, Collaborative Arrangements, and the Annual Report, Appendix I). However, we have been unable to meet the needs of many of the persons requesting access for the simple reason that methods of access have been limited. Some structural chemists and biochemists, because of personal preference or industrial secrecy, to name only two possible reasons, desire programs in their own laboratories on their own computers. Limitations in laboratory computers and CONGEN's currently limited exportability make export non-trivial. Those who access programs at SUMEX must compete with dozens of other persons for access to the machine and therefore obtain poor interactive service during normal working hours. One significant aspect of our proposal is to promote resource sharing in such a way (the dedicated computer) that export becomes simpler by utilizing exportable languages and more commonly available (and less expensive) computers, while at the same time providing a networked computer environment which greatly facilitates remote access and provides good interaction for those who do not have access to suitable computers in their own institutions.

E FACILITIES AVAILABLE.

This research will be carried out in our well-equipped laboratories in the Department of Chemistry and on the existing SUMEX-AIM computer resource at Stanford, supplemented by the dedicated machine requested in this proposal. We are able to support the proposed work within existing space allocations. Our primary equipment needs, besides the computer, are terminals. We have three character-oriented CRT display terminals, a Tektronix 4012 storage terminal and a GT-40 graphics terminal, plus access to various hard copy terminals when required. The additional terminals requested in the first two years of our grant will provide the required support for visualization of three-dimensional representations of molecular structure, currently possible in very limited ways with the existing Tektronix 4012 and GT-40 display terminals.

The SUMEX computer facilities, which will provide the support for the majority of our development efforts, are shown in Figure 11. This system has a complete complement of peripheral devices to support our needs and to interface with the dedicated system we propose. The system also provides extensive software support which is more than adequate for our proposed program developments, including text editors, a variety of languages, debugging facilities and an excellent staff to assist us in complicated problems related to the SUMEX system or one of its supported pieces of software.

Our personnel will become involved, as they have in the past, with structural problems related to research in my group or in the groups of our collaborators. Many of these problems will require chemical and spectroscopic data to be obtained at Stanford or elsewhere. Any costs associated with collection of these data will be borne by the individual research groups. However, it is important to note that Stanford has well-equipped chemical and spectroscopic laboratories in the Chemistry Department which will enable collection of high quality data in support of these structural studies.

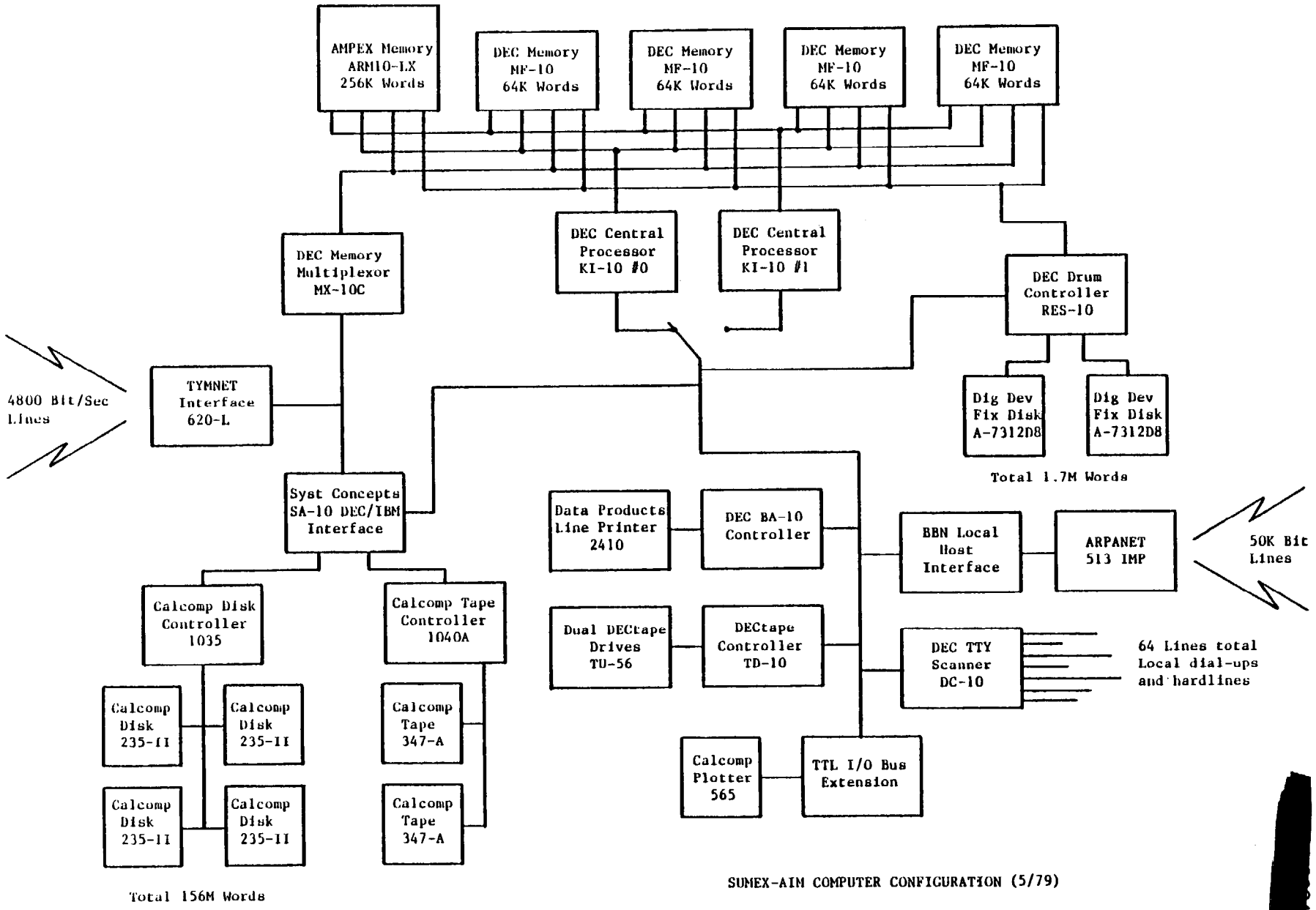


Figure 11. SUMEX hardware

F COLLABORATIVE ARRANGEMENTS.

The proposed improvements to our structure elucidation programs will, in addition, create the possibility of new biomedical applications which make use of existing methods and programs. We propose to promote these new applications via collaborative arrangements.

Collaboration with Prof. David Cowburn.

A very likely application for CONGEN enhanced with a conformation generator would be to the field of conformational analysis. This is the problem of determining the conformation of a structure with known constitution and configuration and is a general problem in describing the structures of molecules. The description of the conformation(s) of molecules of biological origin or of those possessing biological activity is of considerable importance in establishing more clearly the relationship of structure to function in the actions of drugs,hormones, and neurotransmitters on their natural receptors, the mechanism of enzyme action, and the rational design of new drugs. We propose to develop this application by collaboration with Professor David Cowburn and his coworkers at the Rockefeller University in New York. Professor Cowburn is actively engaged in determining peptide conformations using principally nuclear magnetic resonance studies of specifically designed and synthesized isotopic isomers of peptide hormones. These studies use the stable isotopes - deuterium, carbon-13, and nitrogen-15 [91]. Dr. Cowburn now has an account at SUMEX and would use the program remotely, at least at first. It is hoped that an effective collaboration can be developed in which Dr. Cowburn will investigate techniques for effectively rejecting chemically unreasonable conformations as they are generated. Those strategies that may be generally useful will then be adapted for CONGEN and incorporated. These techniques will be related either to general considerations(e.g. insufficient degrees of freedom for cyclization of a particular ring system, from a partially generated conformational state) or to the specific molecules being examined (e.g. restrictions stemming from experimental data such as nmr vicinal coupling constants). Some research using small programs outside CONGEN would be expected to be useful in investigating this area, and some possible techniques are outlined in Dr. Cowburn's letter (pp. -). CONGEN equipped with a conformation generator, would likely be useful to Prof. Cowburn's research in at least three ways:

1) The program would be able to generate all the possible conformations for a given problem with input constraints based on NMR couplings. Such a generation is a difficult task for,e.g, compounds containing large rings. The value of CONGEN would be to provide assurance of exhaustion and to explicitly construct all the possibilities.

2) The program would be able to generate all possible isotopic isomers for a given constitution and configuration. if a pruning technique was available, then the generated list would be extremely useful to Dr. Cowburn in considering the strategies of synthesis and nmr experimentation. The avoidance of particularly costly or time consuming steps is of considerable importance in that experimental work.

3) In conjunction with the spectral interpretation and planning modules proposed, CONGEN may be able to generate strategies for patterns of enrichment or for nmr experiments which are optimum for conformational determination. Some additional programming would probably be necessary to accomplish this.

Collaboration with Prof. Gilda Loew.

Since our proposed conformation generator will output structures with internal (torsional angle) coordinates, it is possible to obtain further information about these structures by doing quantum mechanical energy calculations. By developing a link to these methods, the usefulness of CONGEN should be considerably increased. Since a great deal of work has been done by others on such methods it is not necessary for our group to develop programs of this kind. Instead we propose to develop this link by collaborating with Prof. Gilda Loew and her group in the Dept. of Genetics at Stanford Medical School. Professor Loew's work has involved the use of semi-empirical quantum mechanical energy calculations to derive structure-activity for a variety of drug types [92]. The first step in such a collaboration would be to construct the interface necessary to link the CONGEN output structures with the input for the PCILO (Perturbation Configuration Interaction using Localized Orbitals) program. This program requires as input, structures with internal coordinates. This will be the form of the output from the proposed conformation generator with an assumption of bond lengths and angles. Once this link has been made then we see at least two areas where CONGEN might be helpful to Professor Loew's ongoing research.

1) It will be possible to generate systematically variants of a structure with respect to its constitution, configuration, and conformation. Each such structure would then be given to PCILO for an energy calculation, the results of which are used to help explain potency variations [92]. The advantage of using CONGEN in this way is that an exhaustive generation can be guaranteed which assures no possibilities are overlooked.

2) Professor Loew has been considering the conformational variations caused by the intercalation of ethidium into nucleic acids [93]. The observed stability of such intercalated structures has been related to conformational changes in parts of the DNA structure, in particular, the sugar moieties. The application of CONGEN to such a study would again be a systematic variation of possibilities with particular emphasis on the more difficult cyclic structures.

GUEST Access.

We currently provide access to new users of our programs via the GUEST account at SUMEX. In addition, we have provided GUEST access to several of our recent workshop participants. A list of these participants appears in our annual report (Appendix I). We receive many inquiries about our programs and access to them, far in excess of the current capacity at SUMEX. A list of persons to whom such access has been granted in the past year appears in our annual report (Appendix I). More recent inquiries have been received from (and GUEST access given to):

Prof. Glenn D. Prestwich
Dept. of Chemistry
State University of New York at Stony Brook
Stony Brook, New York 11794

Dr. Andrew Stuper
RCG Group
Rohm and Haas
Norristown and McKean Roads
Springhouse, Pa. 19477

Prof. E. F. Domino
Department of Pharmacology
M6322 Medical Sciences Building
Ann Arbor, Michigan 48109

Djerassi, Carl

[redacted]

Prof. John D. Roberts
Div. of Chem. and Chem. Engr.
Calif. Inst. of Tech.
Pasadena, Calif. 91125

THE ROCKEFELLER UNIVERSITY
1230 YORK AVENUE · NEW YORK, NEW YORK 10021

10 May 1979

Dr. J. Nourse
Department of Chemistry
Stanford University
Stanford, California 94305

Dear Dr. Nourse:

GENERATION OF CONFORMATIONS BY CONGEN

In the past, you and the other members of the Dendral group at Stanford have produced a unique tool for assisting chemists in the interpretation of data, principally in the area of organic structure determination. This tool, the program CONGEN, has recently been augmented by your inclusion of configuration in the descriptive quantities that the program can accept and process.

The possible extension of CONGEN to include conformational values is then the next logical step in the development of this system. In addition to the potential uses of conformer generations in the area of organic structure determination - e.g. inclusion of structural constraints based on experimental observations from NMR or Mass Spectrometric studies, such a program would be of considerable use as a research tool in the applications of conformational analysis in a number of areas. The program would be very useful in exhaustively and precisely defining conformational states in small molecules. In larger molecules (molecular weight >300) the description of standard states (Dunitz 72) and the conformational features associated with certain kinds of rings or regularly repeating substructure are areas of active and significant research endeavor (e.g. [Anet 78], [Dale 75], [Cremer 75]). For polymers, particularly biopolymers, the assumption that a somewhat irregular structure (e.g. a protein) can be described by the cataloguing of standard conformational states associated with the substructures has been widely made ([Desantis 65], [Zimmerman 77]). This assumption has been tested to some degree in a few highly - resolved crystal structures. It has not been tested widely with respect to structures in solution, or to those dynamic structures interconverting relatively rapidly. Such structures, and their analysis, are of very considerable importance in understanding more thoroughly the structure/function relationships of hormones, neurotransmitters and drugs.

It is probable that an enhanced CONGEN could be a very effective tool in these developing fields, for several purposes. For a set of standard states, the program will be able to completely generate all the possible combinations of the conformers of the substructures. This is of considerable importance when a number of rather different substructures are considered in any one molecule, where the permutational task is not straightforward. The powerful features of CONGEN in creating positive and negative conditions for an allowed structure during its generation will provide an exceptionally sound base for testing selections of standard conformational states, for generating a restricted list of possible conformers, and for investigating

Dr. J. Nourse 10 May 1979

conformational classes in sets of related molecules.

The enhanced CONGEN would be useful to our research in all the above ways. I hope that during the period of introducing these enhancements, it will be possible for us to cooperate and collaborate in testing various algorithms that might be incorporated into the final product. These algorithms include various standard modules concerning generation of cartesian coordinates, forced cyclization of coordinates of end groups in rings systems, empirical energy calculations, etc., and modules concerning less well known problems. In this latter area, we are particularly interested in developing techniques for effectively reducing the number of produced conformers during a constrained generation. Methods for doing this may be of a quite general nature. For example, a technique for selection based on the ability of a growning structure to correctly refold to permit cyclization has been described [Dirkx 79]. Exact descriptions for certain regular rings are also possible [Cremer 75]. These pruning techniques may, alternatively, be quite specific to the structure under consideration, e.g. incorporation of known conformational features, or restrictions based on maxima or minima for certain interatomic distances, or restrictions of torsion angles based on predicted effects or specific neighboring groups.

We look forward to being able to pursue this area of research with you, while specifically applying these techniques to the studies of the dynamic conformations of peptide hormones currently under investigation here. The people associated with this research here include Professors William C. Agosta, and David H. Live, Mr. William Wittbold, and myself. Our research in this area is supported by NIH AM-20357.

Sincerely,

David Cowburn

DC:mmh

David Cowburn
Associate Professor

[Anet 78] Anet F. A. L.; Rawdah T. N. "The conformational energy surface of trans,trans,trans-1,5,9-cyclodecatriene." JACS, 1978, 100, 5003-5007.

[Cremer 75] Cremer D.; Pople J. A. JACS 1975, 97, 1358.

[Dale 75] Dale J. "Multistep conformational interconversion mechanisms." Topics in stereochemistry, 1975, 9, 199-270.

[DeSantis 65] DeSantis P.: Giglio E.; Liquori A. M.; Ripamonti A. Nature, 1965, 206, 456-461.

[Dirkx 79] Dirkx J.; Knappenburg M.; Dufour P. "A program for generation of possible conformations of cyclic molecules." Comp. Prog. Biomed. 1979, 9, 63-68.

[Dunitz 72] Dunitz J. D.; Waser J. "Geometric constraints in six and eight membered rings." JACS 1972, 94, 5645-5650.

[Zimmerman 79] Zimmerman S. S.; Pottle M. S.; Nemethy G.; Scheraga H. A. "Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP" Macromolecules 1977, 10, 1-9.

G PRINCIPAL INVESTIGATOR ASSURANCE

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

Date Principal Investigator

H REFERENCES.

1) J. Lederberg,
"DENDRAL-64-A System for Computer Construction, Enumeration and Notation of Organic Molecules as Three Structures and Cyclic Graphs," (technical reports to NASA)
(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, G.L. Sutherland, B.G. Buchanan, E.A. Feigenbaum, A.V. Robertson, A.M. Duffield, and C. Djerassi, J.Am.Chem.Soc., 91, 2973, (1969).

3) R.E. Carhart, D.H. Smith, H. Brown and C. Djerassi, J.Am.Chem.Soc., 97, 5755, (1975).

4) A.M. Duffield, A.V. Robertson, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum, and J. Lederberg, J.Am.Chem.Soc., 91, 2977, (1969).

5) A. Buchs, A.B. Delfino, A.M. Duffield, C. Djerassi, B.G. Buchanan, E.A. Feigenbaum and J. Lederberg, Helv.Chim.Acta, 53, 1394, (1970).

6) 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, Org.Mass Spectrom., 4, 493, (1970).

7) (a) L.M. Masinter, N.S. Sridharan, J. Lederberg and D.H. Smith. J.Amer.Chem.Soc., 96, 7702, (1974).
(b) L.M. Masinter, N.S. Sridharan, R.E. Carhart and D.H. Smith, ibid, 96, 7714, (1974).

8) H. Brown, L. Hjelmeland, and L. Masinter, Discrete Mathematics, 7, 1, (1974).

9) H. Brown and L. Masinter, Discrete Mathematics, 8, 227, (1974).

10) H. Brown, SIAM Journal of Applied Math, 32, 534, (1977).

11) D.H. Smith, B.G. Buchanan, R.S. Engelmore, A.M. Duffield, A. Yeo, E.A. Feigenbaum, J. Lederberg, and C. Djerassi, J.Am.Chem.Soc., 94, 5962, (1972).

12) D.H. Smith, B.G. Buchanan, W.C. White, E.A. Feigenbaum, C.Djerassi, and J. Lederberg, Tetrahedron, 29, 3117, (1973).

13) B.G. Buchanan, D.H. Smith, W.C. White, R.J. Gritter, E.A. Feigenbaum, J. Lederberg, and Carl Djerassi, J.Am.Chem.Soc., 98, 6168, (1976).

14) R.E. Carhart and D.H. Smith, Computers in Chemistry, 1, 79, (1976).

15) T.M. Mitchell and G.M. Schwenzer Organic Magnetic Resonance, 11, 378, (1978).

16) G.M. Schwenzer and T.M. Mitchell, in D. Smith, (ed.), Computer Assisted Structure Elucidation, ACS Symposium Series, 54, Washington, D.C., 1977, p58.

17) (a) C.J. Cheer, D.H. Smith, and C. Djerassi, B. Tursch, J.C. Braekman, and D. Daloze, Tetrahedron, 32, 1807, (1976).
(b) D.H. Smith, Anal.Chem., 47, 1176, (1975).
(c) D.H. Smith, J.Chem.Inf.Comp.Sci., 15, 203, (1975).
(d) D.H. Smith, J.P. Konopelski and C. Djerassi, Org.Mass Spectrom., 11, 86, (1976).

18) R.E. Carhart, S.M. Johnson, D.H. Smith, B.G. Buchanan, R.G Dromey, J. Lederberg, in P. Lykos (ed.), Computer Networking and Chemistry, American Chemistry Society Symposium Series, 19, Washington, D.C., 1975, p192.

19) T.H. Varkony, R.E. Carhart and D.H. Smith, in W.T. Wipke and T. Howe, (Eds), American Chemical Society Symposium Series, 66, Washington, D.C., 1977, p188.

20) (a) T.H. Varkony, D.H. Smith, and C. Djerassi, Tetrahedron, 34, 841, (1978).
(b) R.M.K.Carlson, S. Popov, I. Massey, C Delseth, E. Ayanoglu, T.H. Varkony, and C.Djerassi, Bioorg.Chem., 7, 453, (1978).

21) T.H. Varkony, R.E. Carhart, D.H. Smith, C. Djerassi, J.Chem.Inf.Comp.Sci., 18, 168, (1978).

22) C. Djerassi, D.H. Smith and T.H. Varkony, Naturwissenschaften, 66, 9 (1979).

23) N.A.B. Gray, D.H. Smith, T.H. Varkony, R.E. Carhart, and B.G. Buchanan. "Use of a Computer to Identify Unknown Compounds. The Automation of Scientific Inference," Chapter 7 in "Biomedical Applications of Mass Spectrometry," G.R. Waller (Ed.), in press.

24) James G. Nourse,
J.Am.Chem.Soc., 101, 1210 (1979)

25) James G. Nourse, Raymond E. Carhart, Dennis H. Smith, and Carl Djerassi,
J.Am.Chem.Soc., 101, 1216 (1979).

26) (a) M. Bachiri and G. Mouvier,
Org.Mass Spectrom., 11, 1271, (1976).
(b) G.M. Pesyna and F.W. McLafferty,
Determination of Organic Structures by Physical Methods",Vol 6,
F.C. Nachod, J.J. Zuchermann, and E.W. Randall (Eds),
Academic Press, New York, N.Y., 1976, p91.

27) F.W. Mellon,
in "Mass Spectrometry",Vol 4,
R.A.W. Johnstone, Sr. Reporter,
The Chemical Society,
Burlington House, 1977, p89.

28) (a) K.S. Kwok, R. Venkataraghavan, and F.W. McLafferty,
J.Am.Chem.Soc., 95, 4185 (1973).
(b) H.E. Dayringer, G.M. Pesyna, R. Venkataraghavan,
and F.W. McLafferty.
Org.Mass Spectrom., 11, 529, (1976).

29) S.R. Heller, G.W.A. Milne and R.J. Feldmann,
Science, 195, 253 (1977).

30) R.C. Fox,
Anal.Chem., 48, 717 (1976).

31) D.L. Dalrymple, C.L. Wilkins, G.W.A. Milne, and S.R. Heller,
Org.Magn.Res., 11, 535, (1978).

32) (a) P.R. Naegli and J.T. Clerc,
Anal.Chem., 46, 739a, (1974).
(b) J. Zupan, M. Penca, D. Hadzl and J. Marcel,
Anal.Chem., 49, 2141, (1977).

33) D.H. Smith, M. Achenbach, W.J. Yeager, P.J. Anderson, W.L. Fitch,
and T.C. Rindfleisch.
Anal.Chem., 49, 1623, (1977).

34) (a) M. Senn, R. Venkataraghavan, and F.W. McLafferty,
J.Am.Chem.Soc., 88, 5593, (1966).

(b) K.Biemann, C.Cone, B.R. Webster and G.P. Arsenault,
J.Am.Chem.Soc., 88, 5598, (1966).

35) A. Mandelbaum, P.V. Fennessey and K.Biemann,
Proc.Ann.Conf.Mass Spectrom and Allied Topics, 15th, 111, (1967).

36) A. Kundered, R.B. Spencer, and W.L. Budde,
Anal.Chem., <43>, 1086, (1971).

37) H.B. Woodruff and M.E. Munk.
Anal.Chim. Acta, 95, 13 (1977).

38) H.L. Surprennant and C.N. Reilley,
in "Computer Assisted Structure Elucidation",
D.H. Smith (Ed.),
ACS Symposium Series, 54,
American Chemical Society (1977).

39) C.A. Shelley and M.E. Munk,
Anal.Chem., 50, 1522, (1978).

40) C.A. Shelley, H.B. Woodruff, C.R. Snelling, and M.E. Munk.
in "Computer Assisted Structure Elucidation",
D.H. Smith (Ed.),
ACS Symposium Series, 54,
American Chemical Society (1977).

41) S. Sasaki, Y. Kudo, S. Ochiai and H. Abe,
Mikrochim.Acta, 726 (1971).

42) S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai,
K. Saito, and T. Yamasaki,
J.Chem.Inf.Comp.Sci.,18, 211, (1978).

43) (a) B.R. Kowalski, P.C. Jurs, T.L. Isenhour, and C.N. Reilley,
Anal.Chem., 41, 1945, (1969).
(b) H.B. Woodruff, G.L. Ritter, S.R. Lowry, and T.L. Isenhour,
Appl.Spectrosc., 30, 213, (1976).

44) C.L.Wilkins, R.C. Williams, T.R. Brunner and P.J. McCombie,
J.Am.Chem.Soc., 96, 4182, (1974).

45) P.C. Jurs and T.L. Isenhour,
"Chemical Applications of Pattern Recognition",
Wiley-Interscience,
New York, N.Y., (1975).

46) J. Schechter and P.C. Jurs,
Appl.Spectrosc., 27, 30 (1973).

47) W.E. Brugger, A.J. Stuper, and P.C. Jurs,
J.Chem.Inf.Comp.Sci., 16, 105 (1976).

48) (a) M.E. Munk, C.S. Sodano, R.L. McLean, and T.H. Haskell,
J.Am.Chem.Soc., 90, 1087, (1968).
(b) C.A. Shelley, T.R. Hays, M.E. Munk and R.V. Roman,
Anal.Chim. Acta, 103, 121 (1978).

49) J.E. Dubois,
in "Computer Representation and Manipulation of
Chemical Information",
W.T. Wipke, S.R. Heller, R.J. Feldmann and E. Hyde, (Eds.),
Wiley-Interscience,
New York, N.Y., 1974, p239.

50) L.A. Gribov, M.E. Elyashberg and V.V. Serov,
Anal.Chim.Acta, 95, 97, (1977).

51) R. P. Smith,
J. Chem. Phys., 42, 1162 (1965).

52) P. J. Flory, U. W. Suter, and M. Mutter,
J. Am. Chem. Soc., 98, 5733, (1976)
and earlier cited references.

53) L. E. Scales and J. A. Semlyen,
Polymer, 17, 601, (1976).

54) J. B. Hendrickson,
J.Am.Chem.Soc., 86, 4854, (1964).

55) M. Dygert, N. Go, H. A. Scheraga,
Macromolecules, 8, 750, (1975).

56) J. Dale,
Acta. Chem. Scand., 27, 1115, (1973).

57) M. Saunders,
Tetrahedron, 23, 2105, (1967).

58) D. F. Bocian, H. M. Pickett, T. C. Rounds, H. L. Strauss,
J.Am.Chem.Soc., 97, 687, (1975).

59) J. E. Kilpatrick, K. S. Pitzer, R. Spitzer,
J.Am.Chem.Soc., 69, 2483, (1947).

60) D. Cremer and J. A. Pople,
J.Am.Chem.Soc., 97, 1354, (1975).

61) C. Altona and M. Sundaralingam,
J.Am.Chem.Soc., 95, 2333, (1975).

62) J. Dirkx, M. Knappenburg, P. DuFour,
Comp. Prog. Biomed., 9, 63, (1979).

63) A. Murakami and Y. Akahori,
Chem.Pharm. Bull., 25, 2870, (1977).

64) C. D. Barry, J. A. Glasel, R. J. P. Williams, A. V. Xavier,
J. Mol. Biol., 84, 471, (1974).

65) F. A. Gorin and G. R. Marshall,
Proc.Nat. Acad. Sci., 74, 5179, (1977).

66) R.E. Carhart, T.H. Varkony and D.H. Smith,
in "Computer Assisted Structure Elucidation",
D.H. Smith (Ed),
American Chemical Society Symposium Series, 54,
Washington, D.C., 1977, p126.

67) Dennis H. Smith and Raymond E. Carhart,
in "High Performance Mass Spectrometry: Chemical Applications",
M.L. Gross, (Ed.),
American Chemical Society Symposium Series, 70,
Washington, D.C., 1978, p325.

68) W. Bremser, M. Klier and E. Meyer,
Org. Magn. Res., 7, 97, (1975).

69) W. Bremser,
Fesenius Z.Anal.Chem., 286, 1, (1977).

70) W. Bremser,
Anal.Chim. Acta, 103, 355, (1978).

71) B.A. Jezl and D.L. Dalrymple,
Anal.Chem., 47, 203, (1975).

72) J.T.Clerc and H.Sommerauer,
Anal.Chim. Acta, 95, 33, (1977).

73) D.H.Smith and P.C.Jurs,
J.Am.Chem.Soc., 100, 3316, (1978).

74) J. Zupan, S.R. Heller, G.W.A. Milne and J.A. Miller,
Anal.Chim. Acta, 103, 141, (1978).

75) D.H. Sleeman,
Int.J.Man Machine Studies, 7, 183, (1975).

76) G. Beech, R.T. Jones and K. Miller,
Anal.Chem., 46, 714, (1974).

77) W. Moffitt, R. B. Woodward, A. Moscowitz, W. Klyne,
and C. Djerassi.
J.Am.Chem.Soc., 83, 4013, (1961).

78) J.H. Brewster,
Tetrahedron, 30, 1807, 1974.

79) D.N. Kirk and W. Klyne,
J.Chem.Soc. Perkin I, 1076, (1974).

80) L. Seamans, A. Moscowitz, G. Barth, E. Bunnenberg, C. Djerassi,
J.Am.Chem.Soc., 94, 6464, (1972).

81) R. E. Linder, K. Morrill, J. S. Dixon, G. Barth,
E. Bunnenberg, C. Djerassi, L. Seamans, and A. Moscowitz,
J.Am.Chem.Soc., 99, 727, (1977).

82) W. D. Hounshell, D. A. Dougherty, and K. Mislow,
J.Am.Chem.Soc., 100, 3149, (1978).

83) A. Kerber,
"Representations of Permutation Groups", Vol II,
Springer-Verlag, N. Y., 1975.

84) R.S. Cahn, C. Ingold, and V. Prelog,
Agnew.Chem.Int.Ed.Eng, 57, 385 (1966).

85) F. A. L. Anet,
Fort. Ch. Forsch., 45, 169, (1974).

86) J. H. Dawson, J. R. Trudell, R. E. Linder, G. Barth,
E. Bunnenberg, and C. Djerassi,
Biochemistry, 17, 33, (1978).

87) W. C. Johnson,
Ann. Rev. Phys. Chem., 29, 93, (1978).

88) C. Marcott, H. A. Havel, J. Overend, A. Moscowitz,
J.Am.Chem.Soc., 100, 7088, (1978).

89) E.J. Corey and W.T. Wipke,
Science, 166,178 (1969).

90) W.T. Wipke, H. Braun, G. SMith, F. Choplin, and W. Sielser,
in "Computer Assisted Organic Synthesis",
W.T. Wipke and J. Howe, Eds.,
American Chemical Society Symposium Series, 61,
Washington D.C., 1977, p97.

91) (a) D. Cowburn, A.J. Fischman, D.H. Live, W.C. Agosta,
H.R. Wyssbrod,
Proc.Fifth Amer. Peptide Symp.,
Ed. M. Goodman and J. Meinhofer, 322, (1977).
(b) A. J. Fischman, M. Rieman, D. A. Cowburn,
Febs. Letts., 94, 236, (1978).
(c) D. H. Live, and 5 authors ,
J. Am. Chem. Soc., 101, 474, (1979).

92) (a) G. Loew and J.R. Jester,
J. Med. Chem., 18, 1051, (1975).
(b) G. Loew, D.S. Berkowitz, and R.C. Newth,
J. Med. Chem., 19, 863, (1976).
(c) G. Loew and R. Sahakian,
J. Med. Chem., 20, 103, (1977).

93) G.R. Pack and G. Loew,
Biochim. et Biophys. Acta, 519, 163, (1978).

APPENDIX I

DENDRAL 1978-1979 ANNUAL REPORT

Table of Contents

Section	Subsection		Page
1.	Objectives		97
	1.1	Overall Objectives	97
	1.2	Goals for Current Year	97
2.	STUDIES and RESULTS		99
	2.1	CONGEN	99
	2.2	CONGEN Developments	103
	2.3	RESOURCE SHARING	107
	2.4	Stereochemistry	110
	2.5	Structure checking functions for CONGEN	112
	2.6	Meta-DENDRAL	121
	2.7	REACT and MAXSUB Programs	123
	2.8	High Resolution GC/MS System	124
	2.9	References	125
3.	SIGNIFICANCE		127
4.	RESEARCH GOALS 1979-1980		128

1 Objectives

1.1 Overall Objectives

This progress report covers the second year of our three year grant on computer applications in chemistry, with particular emphasis on techniques of computer-assisted structure elucidation and applications of these techniques to problems of biomolecular structure characterization. To meet this primary objective we have focussed our attention on development of interactive computer programs which are designed to act as chemists' assistants in exploration of the potential structures for unknown compounds. These programs take into account structural information derived from a variety of sources including both physical and chemical methods. We are focussing our research on those aspects of structural analysis which are most difficult to perform manually. We are extending the interpretive power of these programs to enable them to draw meaningful structural conclusions from chemical data. To meet these objectives we are developing a series of computer programs, described in more detail below, which emulate several important aspects of manual approaches to structure elucidation, We are applying these programs to structural problems in our own laboratories and laboratories of others including utilization of the mass spectrometry resource supported under this grant for structural assignment based on mass spectral data.

In order to promote dissemination of the results of our research to the biomedical community, we are developing methods for better access to our programs, including both remote access to the SUMEX resource via nationwide computer networks and exportable versions of important programs including just recently the CONGEN program discussed below.

1.2 Goals for Current Year

Our goals for the current year included the following:

i) develop and test an exportable version of the CONGEN program and begin investigation of actual export;

ii) develop CONGEN to utilize techniques of constraint interpretation developed in year 1 and to incorporate other features useful in structure elucidation (see below);

iii) promote resource sharing utilizing SUMEX, through export of programs and through workshops held at Stanford;

iv) interface the stereochemistry package to CONGEN and provide useful output of the STEREO program to

v) the chemist; under Meta-DENDRAL, explore automated rule formation for both mass and ^{13}C spectra, and use such rules to predict spectra and rank candidate structures for an unknown compound based on agreement between predicted and observed spectral properties;

vi) to develop approaches to automated examination of large numbers of candidate structures to aid in experiment planning;

vii) apply the REACT program to investigation of biosynthetic pathways in the marine sterol field;

viii) develop a program for detection of structural similarities in a set of diverse structures;

ix) exploit the high resolution, combined gas chromatography/mass spectrometry system for identification of new natural products.

We have met these goals during the past year, although the methods used and the programs which resulted reflect some changes in emphasis based on requirements of our own applications and those of outside persons using the programs. We have endeavored to place our emphasis on those aspects of the computational methods which were required for certain key structural problems AND which appeared to be of sufficient generality to warrant including in the programs for future use by other persons. Our approaches to experiment planning, use of stereochemistry, definition of aromaticity and mass spectral prediction and ranking all reflect such exposure to real problems and the results represent what we think are generally useful solutions.

2 STUDIES and RESULTS

2.1 CONGEN

2.1.1 Exportable CONGEN

2.1.1.1 Reprogramming CONGEN

Our previous annual report discussed preliminary development of some portions of CONGEN in the BCPL programming language, specifically the structure generation algorithm. This early experience showed BCPL to be a compact and efficient language containing all of the basic features needed for the full reprogramming effort. Continued development has produced a version of CONGEN in BCPL which contains nearly all of the features of the INTERLISP/SAIL/FORTRAN version. The primary exception is the perception of aromaticity, and this feature is currently being implemented. The BCPL version has the following advantages over the previous one;

a) It requires less than 10% as much computer memory, due partly to the more compact coding and partly to the use of an overlay structure;

b) It uses about 2-5 times less computer time on typical cases than the most highly-tuned (block compiled) previous version of CONGEN;

c) The redesigned front-end provides significantly more error checking, a simpler and more flexible input format, and a more thorough "help" facility;

d) It can easily deal with problems an order of magnitude larger.

e) It is exportable to a variety of different computers.

2.1.1.2 Overlay structure

As portions of CONGEN were developed in BCPL, estimates of its eventual size could be made and it became obvious that the entire program would occupy a somewhat larger amount of memory than is usually available at many installations (on the order of 100-200 K words, still much smaller than the roughly 450 K words needed by the prior version).

Because the processing in CONGEN falls naturally into several independent activities (generating intermediate structures, imbedding, defining substructures, etc.), the program can easily be broken into separate overlay segments which need to communicate only relatively small amounts of information. In the interest of transportability, though, it was decided not to rely upon the overlay mechanism provided by any particular operating system or language. The safest approach seemed to be to divide the overall program into completely independent, separately runnable modules capable of starting one another and communicating with one another via disk files. The drawback of this approach is that there may be a significant overhead in creating and reading files, and in switching from one module to another. But because all information needed to describe a CONGEN session is maintained on file, the program is unusually robust; even if an error causes the program to crash, CONGEN can simply be restarted and it will restore the complete environment which existed before the erroneous command was issued. Also, a particular operating system may offer some means of accomplishing overlays efficiently, and by interfacing the modules through a small control program, it should be possible to take advantage of such facilities. Under TENEX on the PDP-10, for example, a program may control a large number of forks (independent virtual address spaces) each containing a separate program. We have successfully interfaced the CONGEN modules through a small fork-manipulating program so that the overhead of starting a particular module is paid only once for each CONGEN session.

The current CONGEN is composed of eight modules, the largest of which occupies about 46 K words of memory and the rest of which fall in the range 15-36 K words. We are exploring ways of reducing the size of this largest module (SURVEY - see below) to bring it into this range also. The modules and their functions are as follows:

a) CONGEN (35 K) - Main control module, user interaction, error checking;

b) EDITS (19 K) - User interaction for defining substructures;

c) GENERA (26 K) - Generation of intermediate structures;

d) PRUNE (15 K) - Elimination of structures based on structural features;

e) IMBED (36 K) - Expansion of superatoms in intermediate structures;

f) DRAW (26 K) - Output of structural drawings to the user's terminal;

g) SURVEY (46 K) - Examination of large structure lists for frequency of occurrence of standard structural features; and

h) STEREO (24 K) - Generation of stereoisomers.

2.1.1.3 Export status

We are concentrating our export effort on machines which have a significant number of users in the chemical community. We decided to pay special attention to machines which our CONGEN users have access to now and to those which they have indicated to us that they will have access to in the near future. We are also strongly guided by the persons attending the workshops (see Section ???) and the machines to which they have ready access.

Since many of our users have Digital Equipment Corporation PDP-10 computers this was our first priority. The program was designed to run on the Tops-10 operating system since there is a compatibility package which allows programs which run under Tops-10 to also run without change on Tops-20 and the Tenex operating systems. We got a version running on the Tops-10 and exported it to Rutgers where it ran on the Tops-20 system, and to two different Tops-10 sites: Smith,Kline Research and Ely Lilly Research. Since we have a Tenex operating system at Stanford we have now verified that CONGEN does run on all three and that the compatibility package is robust.

We are continuing negotiations for a contract, separate from this proposal, to provide a version of CONGEN accessible through the NIH/EPA Chemical Information System (CIS). That system is currently operating on DEC equipment so that direct export of the current version of the program to CIS will be simple. Complete integration of the program into the CIS framework of programs and their intercommunication is, however, a much more difficult task. This task will be pursued and funded separately from the current grant because it is essentially a mechanical programming and documentation effort with little research content. However, the resulting documentation will be available for all persons to who we export the program, thus benefitting our DENDRAL work.

We decided on two other machines with their associated operating systems for reasons that will be discussed briefly below. The two machines are the IBM 370 running the new Virtual Machine Conversational Monitor System Operating System (VM/CMS) (CONGEN running on this system will also run on the next generation of IBM machines the 3100 and the 3300 because they will run an identical (to the user) version of VM/CMS), and the Control Data Corporation 6600 computer at the National Resource for Computation in Chemistry at the Lawrence Berkeley Laboratory.

A study was made of all of the different operating systems for IBM 370 series computers. We looked carefully at those operating systems with virtual memory. Of these, only VM/CMS seemed to be a reasonable short range possibility since its interactive, time sharing system is very similar to the PDP-10 Tenex system. We applied and were admitted as a project to a Stanford/IBM Joint Study project. This project is slated to provide us with access to a IBM 370 computer some time in the spring. After we get CONGEN running on VM/CMS we plan to investigate in more detail the other 370 virtual memory operating systems and also to investigate the IBM series 1 mini-computer which has a virtual memory operating system. We estimate that 3 man months will be spent on this project.

The NRCC currently has a computer complex consisting of a CDC 7600,6600, and 6400 together with a PDP-8E mini-computer. The 6600 has been dedicated to interactive computing consisting of both interactive programs and an interactive system for preparing batch programs for the 7600. The 6400 serves as an input / output machine and the PDP-8E manages the terminals and teletypes for the 6600. A project at LBL called the real time systems group has brought up a version of BCPL on the 6600 and they have expressed interest in helping us bring CONGEN. We did a detailed calculation and determined that an average CONGEN session on the 6600 conducted over the Tymnet network would cost about 100 dollars at normal priority at week day rates. This seemed reasonable to us and we will be applying to NRCC for a grant for the computer time necessary to bring up CONGEN. We estimate the task to be about one and a half man months.

During our discussion and research we considered a significant number of other machines and other programming languages. We have so far been unable to find any solution to the problem of a version of CONGEN for a mini-computer system such as the DEC PDP-11 series machines. Address space limitations of 32K words make the task prohibitive in terms of effort. Even with systems with memory management, the job of rewriting CONGEN to fit into 64K 16 bit words is probably beyond our present means in terms of programming time required.

We are discussing with Varian Associates, Palo Alto, the prospects for a mini-computer version of CONGEN in the PASCAL language. If they decide to undertake such an effort we may have access to a mini-computer version in a language which is rapidly gaining popularity and already enjoys significant transportability among machines.

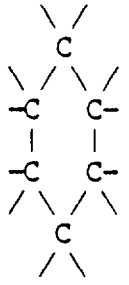
Several other computer systems and languages have been explored for suitability for CONGEN. So far all have suffered from language deficiencies which do not allow the heavy recursion required for CONGEN's basic algorithms (e.g. FORTRAN), from lack of transportability (the BLISS and C languages), or from being implemented on a machine

which is not widely available in the chemical and biochemical community (e.g., Honeywell, Hewlett Packard). These investigations will continue because new machines like the DEC VAX-11/70 will have an increasing number of users in the future and versions of the BCPL compiler will be available for popular systems.

2.2 CONGEN Developments

The reprogramming effort has been far from a transliteration of existing algorithms into BCPL. In many portions, the basic algorithmic approach taken in the previous version was reformulated to allow for a more effective representation and solution of the problem. Aside from the development of and proof of correctness for a new structure-generation technique (related to that of Sasaki) which we discussed in last year's report, and aside from the work described elsewhere in this report on stereochemistry (Section 2.4) and the SURVEY function (Section 2.5), the major milestones in CONGEN development which have paralleled the reprogramming are as follows:

1) Imbedder. The mathematical technique for expanding superatoms in intermediate structures developed by Brown was reexamined and reformulated to allow for a more compact representation. The primary difference in our new approach is that the topological symmetry group of the atoms, rather than the free valences, is used in the computation. For example, the superatom A below, with twelve free valences, has twelve topological symmetry operations



A

interchanging its atoms, but because of the pairwise interchanges between free valences on each atom, the free-valence group has 64x12=768 symmetry elements. The BCPL version of the imbedder carries the symmetry information as 11 permutations of 6 objects (the identity permutation is not explicitly represented)

requiring 66 words of memory, rather than as 768 permutations of 12 objects requiring 9216 words of memory. By implicitly representing interchange symmetry among free valences, among the termini of internal bonds being allocated to the superatom and among monovalent atoms being attached to the superatom, the new version is able to use a drastically smaller amount of space for the storage of symmetry information.

Neither of these approaches to imbedding can perceive all possible sources of duplicate structures, so it was necessary also to develop a final filter package to canonicalize the imbedded structures and compare them for duplicates. However, the new version stores the structure representations on an external random-access file rather than in the computer's memory as was done before, and only a list of pointers to these filed structures is stored internally. As a result, the new imbedder can deal with thousands rather than hundreds of imbedded structures using only a modest amount of memory.

2) Constraints. The basic structure generation and imbedding algorithms are of little practical use without the ability to constrain their output based on the presence or absence of structural features. The graph matcher and cycle finder, which accomplish constraint testing, were translated with little change from their INTERLISP counterparts. Inclusion of constraints in the imbedder, where they serve only as a filter on the final output structures, was straightforward. In the structure generator, however, the constraint-testing mechanism was merged much more intimately with the generation process. The main aspects of this merging are as follows:

> a) As soon as hydrogen atoms are distributed among the non-hydrogen atoms (the first activity of the generator), the distributions are checked against the constraint substructures to determine which distributions can be ruled out a priori. If a substructure is required to be present and contains three methine carbons (CH), for example, the generator will immediately discard hydrogen distributions which do not

contain at least three such carbons. Many constraints supplied to the generator place restrictions on the possible distributions of hydrogen atoms, and by this mechanism such constraints are tested most efficiently.

b) The order in which the generator assembles its atoms is influenced by which atoms appear in the constraints. If a substructure forbidding the construction of peroxides (O-O) is present, the generator will be encouraged to consider possible interconnections among oxygen atoms first so that the presence of peroxides can be avoided early in the computation. Because different constraints may encourage different starting atoms, a scoring scheme has been developed which is used to establish the overall order of atom assembly, taking all constraints into account.

3) Interactive aids. Much effort has been directed toward the development of a robust and helpful interactive system to allow a user easily to define a CONGEN problem and to make use of the basic algorithmic tools. The primary accomplishments in this direction have been as follows:

a) The development of LINSTR, a package of BCPL functions for interactive input from the user, accessed by all of the interactive CONGEN modules. The line-input and prompting functions in LINSTR provide for three levels of help information which can easily be passed from the main program. The first level consists of prompts which are typed to the user when information is required by the program. The novice may step through the prompting sequences supplying one piece of information at a time in response to these prompts, while the expert user may anticipate the prompts

and type ahead his responses on the line to avoid the prompts. This, together with the ability of the LINSTR functions to accept unambiguous abbreviations for keywords, allows a great deal of flexibility in the form of the input. For example, the following two sequences accomplish the same effect in the program (user's inputs are underlined):

Step-by-step input;

```
 DEFINE
DEFINITION TYPE:SUBSTRUCTURE
NAME:R6
(NEW SUBSTRUCTURE)
>RING 6
>DONE
R6 DEFINED
```

Condensed input;

```
 DE S R6;R 6;DO
(NEW SUBSTRUCTURE)
R6 DEFINED
```

A second level of help is provided by the '?' facility which can can be evoked at any prompt in the program. At these points, the '?' input will cause helpful information passed by the main program to LINSTR to be typed to the user. The third level of help is provided by a similar '??' facility, which will cause the program to refer to a much more extensive on-line help document to give a full description of the expected information, and the context in which it will be used. This third level is still under development; the basic mechanism has been developed but we have not yet constructed the on-line documentation.

b) The simplification and extension of the basic commands. The

number of basic CONGEN commands has been reduced from 29 to 14 by the consolidation of commands with similar function (e.g., SHOW is now a general-purpose method of obtaining information about the session and replaces six previous commands) and eliminating little-used options (e.g., TREEGEN). The number of EDITSTRUC commands has likewise been reduced from 23 to 17. Also, previous concepts which were somewhat artificial have been removed. For example, a user does not now need to distinguish between superatoms and patterns when he defines a substructure. The representations for these two types of substructure have been consolidated and a defined substructure can be used in either context. As another example, the user does not need to place substructures on BADLIST any more - the new input sequence allows him to express the presence or absence of substructural features in a natural statement such as 'exactly 3' or 'at most 1' or 'none'. The new command structure seems easier for users to remember and work with.

2.3 RESOURCE SHARING

2.3.1 CONGEN Workshops

In early December, 1978, we held at Stanford a series of mini-workshops on the use of an exportable version of the CONGEN program. Invitees included members of the chemical and biochemical community who are actively engaged in solving the structures of unknown chemical compounds encountered in research in industrial, academic and government research laboratories. The primary purpose of these workshops was to introduce experts in the field of structure elucidation to the first version of the exportable program. These persons were chosen for their chemical and biochemical expertise; few had significant experience with computers previously. Thus, they represented what we think is a good cross-section of the community of potential users of CONGEN. We held three three-day sessions of the workshop so that we could offer access to a computer terminal for all

the persons at one session and so that we could provide close supervision and assistance as they began to learn and use CONGEN. We also implemented a recording scheme so that an interactive session at the terminal could be recorded as a text file and available after the problem was completed for close scrutiny for the chemist and for ourselves. Such scrutiny reveals, for example, common difficulties in certain portions of the user interaction thereby pointing out areas for improving the interaction.

The persons who attended the workshops, their affiliation and a summary of their reactions to the program are summarized in Appendix I. We also include in that Appendix persons who were not able to attend the workshops but desire, on the basis of our contacts with them with regards to the workshops, a copy of the exportable CONGEN. A copy of the original letter sent to one of the invited persons is included as Appendix II to describe our purposes in more detail.

Although the version of CONGEN used in the workshops was not complete, enough of the program existed in close to final form to allow us to fulfill our other purposes. We wanted to ensure that any remaining program errors could be detected and fixed prior to making the program more widely available. The best way we have found to do this once a program is essentially debugged is to confront the program with a wide variety of problems from many different users. We also wanted to determine if there were major deficiencies in any part of the program which made it difficult to understand or use. Eliminating such deficiencies would ensure that an exported version would meet the needs of the persons attending the workshop, i.e., that some minimum standards of acceptability could be determined and met. Finally, we needed to determine the computing facilities available to this group and in detailed discussions to explore opportunities for export to their own laboratories. This allows us to set some priorities on developing versions for various makes of computers. The facilities of each attendee and the current and future state of export to each laboratory are summarized in Appendix I.

2.3.2 Conclusions from the Workshop

There are several conclusions which can be drawn from the workshop experience. The reaction of all persons attending the workshop was very positive, not only concerning organization and intellectual stimulation, but also with the problem-solving capabilities of the program. The following are major positive aspects of the workshop experience:

a) we were able to meet our goal of demonstration of exportability by utilizing CONGEN on two different computers during the workshop;

b) every participant found the program of sufficient utility to express an interest in obtaining a version in some way for his or her own laboratory;

c) the interface to CONGEN, extensively modified based on experience with the old version of the program, proved much simpler to use, much more chemically logical and consistent and much more helpful to the user in providing guidance and error checking;

d) several new problems were analyzed successfully at the workshops, either by verification of the unambiguous nature of the structural assignment or by obtaining a list of candidate solutions to guide further experimentation;

e) installation of the exportable version has been completed successfully at two different sites, Lilly Research and Smith, Kline and French Research, and several more will follow in the next two months.

There are some common criticisms expressed by the persons attending the workshop which, in our opinion, represent points of focus for the remainder of the grant period and for a renewal application. Briefly, the major deficiencies were as follows:

a) The requirement of specifying non-overlapping structural units is non-intuitive and thus unnatural. Other programs, like CONGEN, share this difficulty, but we are in a position to remedy it based on recent research so that future versions may be easier to use;

b) The program is very complex and lacks sufficient documentation or internal 'help' facilities. We recognize this and to some extent it is a reflection of the lack of maturity of the new version. We plan to provide better on-line help facilities accessible from within the program and a much more comprehensive program guide with examples.

c) The teletype oriented drawing program produces some drawings which are difficult (if not impossible) to interpret. Providing the chemist with a connection table of such drawings, as we can do currently, is no long-term solution. Here we face the problem of diminishing the exportability of the program if we restrict its use to certain types of graphics terminals (there are many types, each requiring different programs to operate). Currently there is no graphics terminal which is competitive in price to character-

oriented terminals. One way to solve this problem is to encourage collaborators to provide their own graphics packages which we can then in turn offer to others.

2.4 Stereochemistry

2.4.1 SAIL Program

The stereoisomer generator program written in SAIL and discussed in last year's annual report has been improved in several ways. The program has been modified to process lists of structures to count and/or generate the possible stereoisomers. Thus with the existing CONGEN structure generator it is now possible for the first time to generate all the possible stereoisomers for a given empirical formula completely and irredundantly. These stereoisomers are represented in a compact canonical form and are written onto a disk file by the program along with other information about the structure. Three additional features which were proposed in the last annual report have been added to this program. First, at the user's discretion, the program will compute cis and trans double bond designations for the stereoisomers and write these on the file. Second R and S designations for tetravalent stereocenters based on the Cahn-Ingold-Prelog conventions are computed for stereocenters which are not fixed by any nontrivial symmetry element. These designations were thought to be the most useful and most stable with respect to future changes of the R/S nomenclature system. Third, the ability to handle stereochemistry of common heteroatoms with valence less than 5 has been added. A small interactive package has been added for deciding whether trivalent nitrogen atoms are free to invert. The user is given a choice for each such nitrogen atom.

This program has been included with the current LISP version of CONGEN (it runs as a separate fork) and is available to all users who can access SUMEX. It has been extensively tested on well over 1000 structures. Further details can be found in the publications cited in this report.(HPP-78-8,HPP-78-9)

2.4.2 BCPL program

Since the CONGEN program has been recently reprogrammed into BCPL to create an exportable version, it was decided to also reprogram the STEREO program into BCPL and carry on further developments in that language to ensure compatibility and exportability. With the exception of the parts of the program which compute R/S symbols and handle

heteratoms interactively, this reprogramming has been accomplished. Further developments on this program include a fairly extensive interactive package which allows the user to obtain information about the generated stereoisomers. The user may obtain drawings of projected stereocenters showing absolute configurations of stereocenters (e.g., Fischer projections, Newman projections, double bonds) or obtain drawings of linear segments of structures showing all the configurations of the included stereocenters. The user may also obtain information about the symmetry and equivalent atoms in any stereoisomer. This program is currently running with the BCPL version of CONGEN and was available and tested during the recent series of workshops. This program has been exported with this version of CONGEN.

The experimental version of the BCPL program has been modified to allow for some constrained generation of stereoisomers as proposed in the last annual report. The algorithm and program for exhaustive generation were written with this eventuality in mind. An additional interactive session has been added to the stereoisomer generator which allows the user to add constraints before generating the stereoisomers. At present, the user may input constraints on the absolute or relative stereochemistry of any stereocenters. Thus if part of the stereochemistry of a structure is known, it is possible to constrain the stereoisomer generator to produce just those isomers consistent with the known stereochemistry. This parallels the procedure in the structure generator of CONGEN.

2.4.3 European speaking trip

At the invitation (and expense) of the Center for Interdisciplinary Research at the University of Bielefeld in West Germany, one of our group, J. G. Nourse, talked about recent developments in the CONGEN program. Besides the lecture at Bielefeld at a conference on the applications of permutation group theory to Chemistry, Physics, and Biology, a lecture was given at the University of Bremen (also W. Germany) to a conference on applications of graph theory to Chemistry. In addition invited lectures were given in Berlin (Free University) and twice in Zurich (ETH and University). A great deal was learned about current efforts by others in both the US and Europe on computer applications to chemical structure elucidation, synthesis, and data bases. Considerable interest in our programs resulted. Besides continuing correspondence, this is evidenced in part by the presence of Prof. Andre Dreiding of the University of Zurich at one of our recent CONGEN workshops. The contents to some of these lectures are included in references 14 and 20.

2.5 Structure checking functions for CONGEN

2.5.1 Introduction

A program, "STRUCC", has been developed to provide functions for checking sets of structures for desired substructural features or for compatibility with recorded mass-spectral or nmr data. While primarily devised for processing sets of structural isomers produced by means of CONGEN , STRUCC can also take as input sets of structures created through the REACT program or defined though an extension of CONGEN's EDITSTRUC function.

The main structure checking functions currently available through STRUCC are:

1) EXAMINE: This EXAMINE function is an extended version of that available in standard CONGEN. Amongst other extensions are facilities for checking for specified ring-fusions or spiro-junctions within structures.

2) MSA: The MSA ("Mass Spectral Analysis") functions provide a means for using mass spectral data to rank candidate structures. The MSA functions can employ either ordinary "half-order theory", or a model of fragmentation in which bond break plausibilities are related to specified substructural features.

3) LOOK: The LOOK (1) functions are intended to assist a user in investigating the utility of proposed experiments for differentiating between candidate structures. LOOK provides a mechanism for determining the various different ways in which particular superatom parts are incorporated into candidate structures.

4) TSYM: The TSYM function allows some simple forms of symmetry constraint to be defined. These constraints use only topological symmetry.

5) RESONANCECHECK: The RESONANCECHECK function is intended for checking that all constraints have been given to the structure generator. The function can identify differences in candidate structures that would be associated with features in the 1Hnmr or 13Cnmr that

(1) (The LOOK functions incorporate some of the features of the PLAN functions described in last year's report).

one might reasonably expect to be fairly obvious (e.g. different numbers of hydroxy protons, different numbers of carbonyl carbons etc). Generally, such differences are found in cases where the user has forgotten to specify substructural features incompatible with the observed data, or has misapplied the constraints so that not all instances of unwanted features are eliminated.

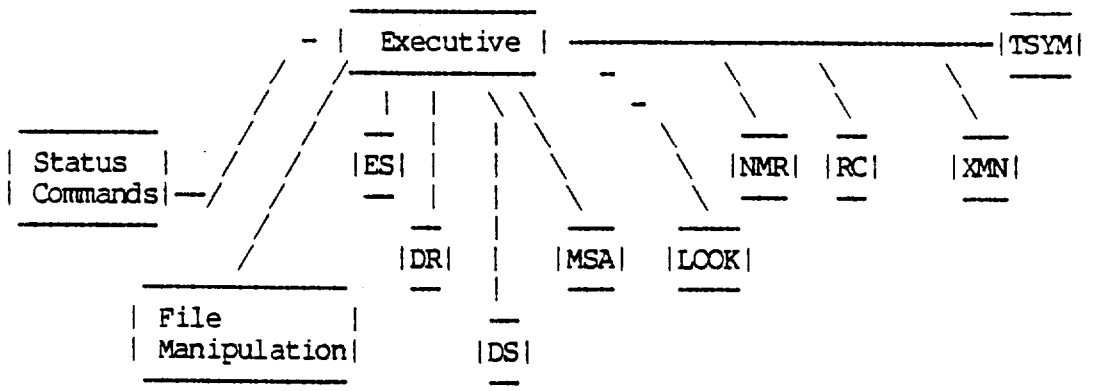
6) NMRFLT: The NMRFLT functions represent a first attempt at developing a system for predicting proton resonance spectra of candidate structures, and for using differences between predicted and observed spectra as a basis for pruning the structure list.

The STRUCC system is also used as a test-bed for new structure evaluation functions. When functions are considered to be sufficiently developed to be of use, top-level calls to those functions are added to STRUCC's repertoire of commands.

STRUCC has a user-interface similar to that of CONGEN and incorporates many of the same subsystems (e.g. EDITSTRUC and DRAW).

2.5.2 The Form of the STRUCC Program

The following diagram indicates schematically the overall form of STRUCC:



1) Status commands:

i AR?: Lists the "aromatics templates" used in CONGEN's last GENERATE or IMBED step.

ii CLEAR: Restarts the program.

iii CM?: Gives the structure composition.

iv CN?: Lists the contents of the "Global Constraint list" used in CONGEN's last GENERATE, IMBED or PRUNE step.

v CT?: Gives the current number of structures.

vi EF?: Gives the empirical formula (if defined).

vii EXIT: Ends the program.

viii UA?: Displays all user defined superatoms and patterns.

2) File Manipulation:

i RESTORE: Reads in a CONGEN-file (or a REACT-file) containing defined superatoms, composition, constraints and structures.

ii BCPL: Reads in a file of structures created through the new BCPL version of the CONGEN program in order that they may be analyzed through MSA, EXAMINE etc.

iii APPEND: Adds all structures from a CONGEN save-file to the set currently in memory and then eliminates any duplicates. This option is useful for combining results from problems where the structure generation process was performed several times with different assumptions about starting superatoms etc.

iv SAVE: Creates a CONGEN-file containing current superatoms, structures etc.

3) EDITSTRUC (ES): CONGEN's standard EDITSTRUC

function is available. See the CONGEN user's manual for details of this function.

4) DRAW (DR): CONGEN's standard DRAW function is available. See the CONGEN user's manual for details of this function.

5) DEFINE-STRUCTURES (DS): The DS function provides various extensions to standard EDITSTRUC that are useful when creating sets of related structures. The DS function is used when the structures to be processed by one of the analysis functions (e.g. MSA) have not been created by REACT or CONGEN.

6) MSA: Mass spectral analysis.

7) LOOK: (for assistance in experiment planning etc).

8) RESONANCECHECK (RC): (Simple checks for omitted constraints).

9) EXAMINE (XMN): Extended version of EXAMINE.

10) NMRFLT (NMR): Prediction and checks on proton resonance spectra of candidate structures.

11) TOPSYM (TSYM): TSYM will prune the structure list to retain only those structures in which some, user defined, substructure has a given minimum number of symmetrically equivalent images.

On starting, or on restarting subsequent to a "CLEAR" command, STRUCC first lists any news bulletins about new options/bugs etc and then asks whether the structures might vary in composition. Many of the processing functions use checks on composition and have to be informed as to whether these checks have to be performed just once, or, for each structure being processed. If the structures were generated by CONGEN then all will have the same composition but structures produced by REACT or entered manually through DEFINE-STRUCTURES may vary in composition.

2.5.3 STRUCC's HELP System

STRUCC has a primitive on-line documentation system. This subsystem is invoked by giving the command "HELP" in reply to a prompt from the program. If the command "HELP" is used alone, then the program retrieves information supposedly useful within the current context.

Arguments can be used with the "HELP" command, e.g. "HELP TAG UNTAG" would result in HELP trying to find information on the EDITSTRUC TAG and UNTAG commands. The HELP files do contain some commented examples of the more complex functions.

2.5.4 DEFINE-STRUCTURES

The DEFINE-STRUCTURES (DS) command allows you to define complete structures by means of an extended EDITSTRUC system. Typically, the DS-command would be used to enter a set of structures that are to be processed by one of the analysis routines —— such as MSA —— but which have not been generated by CONGEN.

Generally, sets of structures that are being entered by means of the DS-system will share common substructures. For example, the structures might consist of steroidal compounds based on one or two nuclear skeletons and half a dozen sidechains. The DS-system allows you to use substructures, (previously defined as Pattern-type Superatoms in EDITSTRUC) when creating new structures.

2.5.5 MSA, The Mass Spectral Analysis Functions.

The MSA functions utilize an extended version of DENDRAL's "half-order theory of mass spectrometry", (described in previous reports), and can provide the following forms of mass-spectral analysis:

1) PREDICTION: prediction of spectra on the basis "half order theory". The program has to be given:

i Parameters controlling the fragmentation process

ii A minimum plausibility value for ions to be listed

iii The minimum mass of interest.

All structures on the structure list are processed and their spectra are listed at the terminal.

2) ANALYSIS: In this mode, MSA can be used to list all possible rationalisations for observed ions. The program lists, for each ion, the breaks, neutral losses and H-transfers necessary for it to be generated from a given structure. In general, this is a large amount of data; consequently, the program only processes a user-defined subset of the structures. Each structure in the subset is processed in turn with the fragmentation

analysis being listed at the terminal. The program has to be given:

i) The index numbers of the structures to be processed

ii) The observed spectrum

iii) Fragmentation control parameters.

iv) A minimum plausibility value on processes that are to be reported.

3) RANKING: For ranking structures, the program has to be given:

i) The observed spectrum.

ii) Fragmentation control parameters.

iii) The form of the scoring function. The contribution to a structures' score from a recorded ion being predicted is given as the product of the predicted plausibility and one of:

a) 1 (presence/absence of ion is all that matters)

b) The ion's mass

c) The ion's observed intensity

d) The product of the ion's mass and intensity.

All structures are processed; optionally, their scores can be listed as they are processed. Once all have been processed, the program produces a ranked listing of the structures. It is then possible either to simply prune away those with inadequate scores, or to enter EXAMINE with these scores. Within EXAMINE, the results of MSA-scoring can be combined with substructural features to

form selection criterion based on overall agreement with the spectrum and presence of desired features.

4) EXAMINE: In this mode, the program identifies all structures for which the observed ions are predicted. The information is converted into a form that can be used by EXAMINE. The observed ions can then be used as EXAMINE-selection keys, just like substructural features; so, one can select structures with
>C8H6Ø1 AND C6H1ØN1O3 AND C4H8N1O2
The number of ions that can be rationalized in terms of a given structure is used as a score for that structure. This score is available in EXAMINE. So, as well as checking for structures that can explain particular ions, it is possible to request those which can explain a given number of the observed ions. In EXAMINE mode, MSA requires the same data as when in RANKING mode.

The basic set of parameters which may have plausibilities adjusted in the "half order theory" are:

1) the plausibility of single bond breaks, (e.g. 1)

2) the plausibility of aromatic bond breaks, (e.g. Ø)

3) the plausibility of double bond breaks, (e.g. Ø)

4) the plausibility of bonds of higher order breaking, (e.g. Ø)

5) the plausibility of adjacent breaks, (e.g. Ø.25)

6) the plausibility of the molecular ion being observed, (? class dependent)

7) if multi-step processes are permitted, then taking the plausibility of single step processes as 1, values must be given for relative plausibilities of more complex processes
e.g. two step processes (e.g. Ø.7)
three step processes (e.g. Ø.4)

8) if H-transfers or neutral losses are specified

then plausibility values must be given for each transfer/loss.

MSA functions allow substructural patterns (created using EDITSTRUC) to be used to define bond environments to which special break plausibilities are to be assigned. The program works by checking whether any of these substructural patterns match a structure, and if so which bonds in the structure correspond to those for which special break plausibilities have been designated. Then, when the program is fragmenting that structure to predict ions, it can check if any of the bonds it has broken are in the list of those having special break plausibilities.

As well as allowing these more general mechanisms for defining the plausibility of bond breaks, the MSA functions let the plausibility assigned to a predicted ion to be adjusted according to how well it is likely to localize charge. The basic "half order theory" does not make allowance for factors such as Nitrogen being able to better stabilize a charge than Carbon and, consequently, Nitrogen-containing ions being more plausible than those without Nitrogen. In MSA, relative charge-localization plausibilities may be defined for different atom-types. The plausibility assigned to a predicted ion is then modified by the maximum charge-localization plausibility of any of its constituent atoms.

2.5.6 LOOK

Frequently, a chemist can conceive of additional experiments that could serve to probe the structural environment of one of the superatom parts that he has used in defining a CONGEN problem. Such experiments might involve a reaction at the site of the superatom part or a series of proton decoupling measurements for "walking along alkyl chains" from some identifiable starting point. Generally, the utility of such experiments depends on there being some significant structural difference between candidates within some relatively small radius of the already known superatom part. The LOOK functions are intended to assist the chemist in finding such differences.

Basically, LOOK takes the starting superatom (or any other substructural pattern that the user may wish to define), maps it into each structure, expands it by including neighboring atoms, creates a canonical representation of the expanded part and groups candidates according to these canonical representations. LOOK then reports on the different expanded features that have been identified and allows the user to further inspect these larger features. The user can choose for a part to be further expanded to achieve some finer discrimination or can investigate differences relating to ring-systems involving the new feature etc. In LOOK, the substructure expansion process is controlled through user specified options.

2.5.7 The Proton NMR Functions

Some simple functions are now available that can be used to specify features in the proton resonance spectrum and prune the structure list to obtain only those candidates that appear to provide a rational for the selected features.

These functions use an "additivity of shifts" model for predicting the proton resonance spectrum of a candidate structure. This model ignores all steric effects; including such important influences as shielding/deshielding through close proximity to an unsaturated system. Further, as shift values in reference tables represent averages over many different types of (usually acyclic) compounds, they can provide but a poor model for any given structure. One can hope that the predicted resonances of methylene groups will generally be within about Ø.6ppm of the observed values while methines should be within 1.5ppm.

The formulae used are:

$$Delta_{CH2} = Ø.2 + C1 + C2$$

$$Delta_{CH} = Ø.2 + C1 + C2 + C3$$

$$Delta_{C=CH} = 5.2 + Z_{gem} + Z_{cis} + Z_{trans}$$

where the Ci and Zi values are supposedly additive constants.

The resonances of methyl, aromatic, alkyne, aldehydic, hydroxy and some other classes of protons are not computed but taken from standard tables. For some of these classes, e.g. hydroxy and aromatic protons, the resonance values are given as a range rather than any typical value.

If the approximate prediction methods appear tolerably accurate for a given class of candidate structures, then the functions can be used for pruning the structure list by tests that predicted spectra satisfy user-defined constraints. These constraints take the form of requirements for specified (minimum) numbers of protons resonating in (possibly overlapping) regions of the spectrum.

2.5.8 BCPL versions of STRUCC

The more useful components of the STRUCC system are being converted to BCPL so that they may be available to future users of the BCPL-CONGEN system.

2.6 Meta-DENDRAL

2.6.1 META-DENDRAL PROGRESS

2.6.1.1 INTSUM

The INTSUM program for the analysis of spectra has been improved by using confidence factors in the place of many of the original program constraints. This feature allows association of liklihoods with fragmentations. It thus allows consideration of a much wider range of possible processes while limiting the final explanations for spectrum peaks to the most plausible explanations.

Additional improvement of the program allows logical separation of the concepts of H-transfers and neutral composition transfers. This provides a better correlation between the explanations provided by the program and those expected by the chemist.

2.6.1.2 RULEGEN

A significant problem in generalizing the INTSUM explanations has always been reducing the size of the search space so as to be able to produce interesting rules in a reasonable amount of time. In addition to the constraints already provided, the RULEGEN program now allows use of existing rules to filter the peak explanations to be considered. This is an important step in allowing the program to focus on rules which account for peak explanations not yet encompassed by existing rules. As an aid in better understanding the process of rule formation, the program is now capable of generating additional information about the search space. This information serves as data for other programs which can then analyze and present to the user compact descriptions of the rule search done by RULEGEN.

2.6.1.3 EDITSTRUC INTERFACE

The latest versions of the structure editor, EDITSTRUC, and the structure drawing programs have been interfaced to allow their use in all appropriate places in INTSUM and RULEGEN. The newest programs for conversion of EDITSTRUC structures recognize a larger subset of the structural features which may be specified within EDITSTRUC. This allows the user greater flexibility in the specification of substructures in user-created rules.

2.6.1.4 PREDICTION and RANKING

The programs allowing the entry and use of user-defined rules have

been extended to allow prediction of the molecular ion and inclusion of confidence factors in the rules.

The process of spectrum prediction from Meta-DENDRAL rules has previously involved the matching of rules against only those sites in the molecules considered as possible breaks. With the use of user-entered rules, and program developed rules containing greater structural detail, the program was generalized to allow prediction based on graph matching alone, without the prior generation of possible break sites.

2.6.1.5 HUMAN ENGINEERING

Many minor improvements have been made in the program's interaction with the user. In general, these improvements have been designed according to the following criteria: 1. Messages should be informative yet not excessively long or wordy; 2. User typing should be kept to a minimum; 3. Programs should behave in ways which people find understandable; 4. During execution, programs should provide occasional information concerning their progress.

2.6.2 RESULTS

The practical value and capability of new programs are best evaluated by applying them to real, non-trivial problems. In our case, we have chosen the biologically important marine sterol compounds. Their mass spectra are predominant in the structure elucidation of new compounds in spite of the fact that relatively few of the fragmentation mechanisms are known. Often very similar spectra are recorded due to the great similarity of common skeletons. Our study involves the comparison of predicted spectra of known structures with the observed spectra of unknown compounds. We want to compare the usefulness of different methods of forming the rules used for spectrum prediction. We distinguish 3 methods: 1) Half-order theory (can be supplemented by functional group rules). 2) Class-specific rules (selected by the chemist) 3) Computer-generated rules Our results were obtained using nine selected 4-demethylsterols (six isomers of composition C29H48O, two C28H46O and one C27H44O). Each spectrum of the nine selected marine sterols was considered to be the observed spectrum and ranked against 23 candidate structures (the 23 candidates contained 17 different C7 - C11 sidechains and three 4-demethylsterol skeletons). For the half-order theory an overall average performance of (2.4 0.9) was obtained. The first number gives the number of candidates ranked better than the correct one, the second represents the number of candidates ranked equally with the correct one. In this case the average value is not very representative, as its value is strongly reduced by a compound which was ranked in 17th place. This compound, the 23-demethylgorgosterol, contains a cyclopropane in the sidechain for which no special fragmentation processes are considered in the simple half-

order theory. The ranking can be greatly improved by providing fragmentation rules for cyclopropane rings. The results of the second method (class specific rules), depends on the quality and number of selected rules. For this study we selected about 17 skeleton breaks (observed in more then 70 percent of the structures) from the INTSUM results of 23 marine sterols to which we added 13 known fragmentation processes. These processes (associated with neutral transfers, intensity range, and a confidence factor) were entered using the new rule editor program. The overall performance of these rules was (0.3 0) which means that, with the exception of three compounds, which were ranked in the second position, the correct structure was always ranked first. A further improvement is seen when the distribution of the scoring values is considered. For these rules, much better separations were observed than with the half-order theory. Also, the quality of the predicted spectra are sufficient to consider the creation of a library which could be visually compared without the need of a computer. For the third method no results can be summarized here, as the computer generated rules are still being developed. The improvement of this last step will be a main goal of the next year.

2.7 REACT and MAXSUB Programs

2.7.1 REACT

During the last year there have been no additional developments in the REACT program. Rather, it has been used extensively in applications to both structure elucidation problems, and, more effectively, in mechanistic studies involving plausible biochemical cyclization and rearrangement pathways.

A major paper describing the REACT program and the underlying algorithms which allow it to interpret automatically structural constraints applied to reaction products appeared this year (9). This paper was concerned more with describing the program for interested persons, but did include a simple example application involving the structure elucidation of a sesquiterpenoid alcohol isolated from a marine organism. The program was also described in more "chemical" terms for a general audience in a review paper which will appear shortly (15).

In conjunction with the work on Meta-DENDRAL and spectrum prediction and ranking applied to analysis of marine sterols (see Section 2.f), we have employed the REACT program to generate biochemically plausible sterol side chains. As we described in the previous annual report, reaction mechanisms thought to be applicable to side chain modification, including cyclizations, rearrangements and

degradations, were supplied to REACT as, effectively, constraints on the variety of side chains which are theoretically possible. For example, CONGEN can be used to generate isomeric C-11 sterol side chains possessing one double bond. There are 7769 (!) of them. Using REACT, however, only 76, less than one percent, are predicted as plausible.

Recent papers have illustrated this approach for both extended (7) and shortened (19) side chains. Recently we showed (15) that seven new structures were all predicted by the program, adding some support to the hypotheses of biochemical transformations.

2.7.2 MAXSUB Program

The function of the MAXSUB program is to detect common structural features in a potentially diverse but related set of compounds. This problem is one faced by chemists engaged in structure/activity studies, particularly in design of new, biologically active compounds based on known compounds with known activities. However, any problem involving an "activity" related to structure, including spectral signatures, is in principle amenable to analysis by MAXSUB. MAXSUB, by determining common features of structures displaying common activities, is presumably focussing on those aspect of the structures which are related to the activity. However, in its current state, the program is only experimental. Many types of activity are intimately connected with stereochemical aspects of structure and MAXSUB does not include any stereochemistry. It does represent a foundation for further study of the problem because the algorithms can in principle deal with three-dimensional descriptors of atoms and bonds. Some work may be done on this program in the next grant period. The existing program will be described in detail in a publication which will appear soon (18).

2.8 High Resolution GC/MS System

For the current grant period we deemphasized further development of our GC/MS and GC/HRMS system as requested by the study section and focussed our attention on maintenance of the existing system and applications of the system to a variety of mass spectral and structural problems of ourselves and our collaborators. In addition we have in press a major paper describing in detail our approach to both GC/low resolution mass spectrometry and GC/high resolution mass spectrometry (17). In this paper we describe methods of data acquisition, reduction and preliminary analysis, a description which includes all major elements of our mass spectrometer/computer systems and a variety of computational details where our approaches differ from those of other workers in the field. In a companion paper we describe how resulting mass spectral data have been utilized in computer-assisted structure

elucidation (16). The following list summarizes the samples we have analyzed in various operating modes of the instrument during the past year:

1) High Resolution analyses:

a) DENDRAL-related	134
b) Outside collaborators	45

2) High Resolution GC/MS

a) DENDRAL-related	86
b) Outside collaborators	13

3) Low resolution GC/MS 45
(these samples were primarily marine sterol mixtures from our laboratory and from several other groups, which did not require HRMS analysis)

2.9 References

In this section we summarize recent publications supported wholly in part by the current grant. This list includes a few publications published at the end of 1977 to help set the context for more recent publications which build on the previous results.

(1) T.H. 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., 1977, p. 188.

(2) "Computer-Assisted Structure Elucidation," D.H. Smith, Ed., American Chemical Society, Washington, D.C., 1977.

(3) R.E. Carhart, T.H. Varkony, and D.H. Smith, "Computer Assistance for the Structural Chemist," in "Computer-Assisted Structure Elucidation," D.H. Smith, Ed., American Chemical Society, Washington, D.C., 1977, p. 126.

(4) D.H. Smith, M. Achenbach, W.J. Yeager, P.J. Anderson, W.L. Fitch, and T.C. Rindfleisch, "Quantitative Comparison of Combined Gas Chromatographic/Mass Spectrometric Profiles of Complex Mixtures," Anal. Chem., 49, 1623 (1977).

(5) B.G. Buchanan and D.H. Smith, "Computer Assisted Chemical

Reasoning," in "Computers in Chemical Education and Research," E.V. Ludena, N.H. Sabelli, and A.C. Wahl, Eds., Plenum Press, New York, N.Y., 1977, p. 401.

(6) D.H. Smith and R.E. Carhart, "Structure Elucidation Based on Computer Analysis of High and Low Resolution Mass Spectral Data," in "High Performance Mass Spectrometry: Chemical Applications," M.L. Gross, Ed., American Chemical Society, 1978, p. 325.

(7) T.H. Varkony, D.H. Smith, and C. Djerassi, "Computer-Assisted Structure Manipulation: Studies in the Biosynthesis of Natural Products," Tetrahedron, 34, 841 (1978).

(8) D.H. Smith and P.C. Jurs, "Prediction of 13C NMR Chemical Shifts," J. Am. Chem. Soc., 100, 3316 (1978).

(9) T.H. Varkony, R.E. Carhart, D.H. Smith, and C. Djerassi, "Computer-Assisted Simulation of Chemical Reaction Sequences. Applications to Problems of Structure Elucidation," J. Chem. Inf. Comp. Sci., 18, 168 (1978).

(10) D.H. Smith, T.C. Rindfleisch, and W.J. Yeager, "Exchange of Comments: Analysis of Complex Volatile Mixtures by a Combined Gas Chromatography-Mass Spectrometry System," Anal. Chem., 50, 1585 (1978).

(11) W.L. Fitch, P.J. Anderson, and D.H. Smith, "Isolation, Identification and Quantitation of Urinary Organic Acids," J. Chrom., in press.

(12) W.L. Fitch, E.T. Everhart, and D.H. Smith, "Characterization of Carbon Black Adsorbates and Artifacts Formed During Extraction," Anal. Chem., in press.

(13) W.L. Fitch and D.H. Smith, "Analysis of Adsorption Properties and Adsorbed Species on Commercial Polymeric Carbons," Environ. Sci. Tech., in press.

(14) J.G. Nourse, R.E. Carhart, D.H. Smith, and C. Djerassi, "Exhaustive Generation of Stereoisomers for Structure Elucidation," J. Am. Chem. Soc., in press.

(15) C. Djerassi, D.H. Smith, and T.H. Varkony, "A Novel Role of Computers in the Natural Products Field," Naturwiss., in press.

(16) N.A.B. Gray, D.H. Smith, T.H. Varkony, R.E. Carhart, and B.G. Buchanan, "Use of a Computer to Identify Unknown Compounds. The Automation of Scientific Inference," Chapter 7 in "Biomedical Applications of Mass Spectrometry," G.R. Waller, Ed., in press.

(17) T.C. Rindfleisch and D.H. Smith, in Chapter 3 of "Biomedical Applications of Mass Spectrometry," G.R. Waller, Ed., in press.

(18) T.H. Varkony, Y. Shiloach, and D.H. Smith, "Computer-Assisted Examination of Chemical Compounds for Structural Similarities," J. Chem. Inf. Comp. Sci., in press.

(19) R. Carlson, et al., Bioorg. Chem., 7, in press.

(20) J.G. Nourse, "The Configuration Symmetry Group and its Application to Stereoisomer Generation, Specification and Enumeration," J. Am. Chem. Soc., in press.

3 SIGNIFICANCE

There are several results obtained during the past year which are especially significant for both the advancement of our techniques for biomolecular structure elucidation and the dissemination of results of our research. Perhaps the most significant result is in the category of dissemination because for the first time we can offer directly to the biomedical community the results of our research in the form of exportable computer software for computer-assisted structure elucidation. We have discussed in the past the differences between our work and many other forms of health-related research. The latter can generally be described in the literature in sufficient detail for others to duplicate the procedures and results and, more importantly, build on those results in extensions of the work. When one of the primary "results" of research is a set of complex computer programs, transfer of these results becomes a formidable task. Certainly the programs and even some of the algorithms can be described in the literature. We have done so, and the recent references at the end of the previous section represent such publications. However, unless an interested person can actually gain access to a program in an executable, or runnable, form, such descriptions are inadequate. We now have the capability of offering the new, smaller, faster version of CONGEN to the community at SUMEX on a trial basis and exported to their own computers for those who wish to make more extensive use of the program. Therefore, we can share our results with our collaborators much more easily now than in the past and persons receiving the program have a foundation on which to build.

The workshops were also significant in that we received constructive suggestions and criticisms from a cross-section of potential users of our computational techniques, while at the same time exposing a variety of persons active in structure elucidation to CONGEN working on real problems. The success of this effort has encouraged us

to make available in the exportable CONGEN a variety of other structure manipulation tools which we and the persons at the workshop perceive as useful as adjuncts to CONGEN. Thus, as we have developed capabilities for exploration of large numbers of structures in the STRUCC program, and new ways to use and display the results of the STEREO program, we have begun incorporating these capabilities into CONGEN. This effort will continue, using CONGEN as the focal point for further developments in the area (see next section).

Other significant advances have been summarized in the sections outlining various aspects of our research, above. These include new mathematical developments which made possible the reduced size and increased efficiency of CONGEN, completion of the STEREO program and integration of it into CONGEN, new approaches to rule formation in Meta-DENDRAL and better methods for predicting mass spectra and ranking structures based on those predictions. The last item deserves further comment, because it represents a general approach to structural analysis of which we can now take advantage because of the efficiency of the new CONGEN. Spectrum prediction for a variety of spectroscopic techniques, including mass, NMR, IR, etc., including now even chiroptical methods for future work, represents a valuable method for structure determination. Now that we can deal with hundreds or even thousands of structures in intermediate stages of a problem, such spectrum prediction techniques provide powerful filters to separate plausible from implausible structures.

4 RESEARCH GOALS 1979-1980

We have several goals for CONGEN development and export during the next grant year. Because several were summarized previously, we give only a brief list here. The reprogramming effort will continue, first to incorporate aromaticity into the program and second to include capabilities for constraints interpretation and translation. These capabilities have been experimental efforts in the old, LISP version of CONGEN. Some additional development must be done and the BCPL version brought up to date with these features. Such developments will allow much more intuitive use of the program (see section on Conclusions from the Workshops) and greatly improve the ease with which structural problems can be specified to CONGEN.

We will be pushing very hard for further export of CONGEN either by developments here at Stanford or by assisting others at remote sites. We can now supply version for DEC-10 and DEC-20 under TOPS-10 or 20 or TENEX operating systems. The contract for an NIH/EPA CIS version will make CONGEN available on that system soon. Our Joint Study project will hopefully allow us to develop an IBM version at

minimal cost. A version for the CDC systems at the National Resource for Computation in Chemistry is under study. The whole area of minicomputer versions is under investigation. We expect that these efforts will make the current version widely available. However, it does leave open the question of long-term maintenance and, particularly, upgrading older versions as new developments ensue. These issues will be dealt with at length in our renewal proposal for funding subsequent to the coming year.

In further response to the workshop requests, we will be developing an extended "help" facility for the program which, together with improved documentation, will improve the ease with which new persons can become familiar with CONGEN. We have already begun investigations of how to improve the current structure drawing facilities in CONGEN, with our first goal to improve the teletype drawings so that all users can benefit. We will also be considering methods for aiding those with graphics terminals to exploit the improved structure input and output possible with such devices.

Further work on the STEREO program will be to develop further the constrained stereoisomer generator and to improve interaction with the main CONGEN program and the user. Specifically, constraints involving the chirality of structures, patterns of equivalent atoms based on either stereoisomer or topological symmetry, and undesirable structural features such as trans double bonds in small rings will be implemented. The flow of information within the program will become two-way between CONGEN and the stereoisomer generator. At present, this information only flows to the stereoisomer generator. The user interaction will be improved so the user can more easily visualize the stereoisomers and the stereochemistry of substructures and can input stereochemical information more intuitively.

We will emphasize several lines of investigation with regards to examination and evaluation of structural candidates from CONGEN, through the STRUCC program. Experiment planning will be approached from the current PLAN and LOOK commands, which search the environment of program- or user-specified structural features and give the experimentalist information on how the structural candidates vary with respect to those environments. The mass spectrum prediction and ranking functions will be completed and transferred to BCPL versions running as part of the CONGEN program. The concept of prediction and ranking will be extended to proton and carbon NMR data using, first, simple additivity rules and, later, refining the predictions based on more detailed examination of the structural relationships among the protons and carbons whose signals are being predicted. We hope that this approach will be effective in choosing a small subset of highly plausible structures based on agreement between predicted and observed spectra, just as the mass spectrum analysis functions have proven useful.

We plan no further development of the GC/MS/Computer system, or the REACT and MAXSUB programs. Rather, these will be used in applications to current structural problems in conjunction with compounds or data from other sources. Incremental improvements will be made as necessary if a particular new application demands it.

Appendix I. CONGEN Workshop Attendees, Affiliation, Research Interests, Comments on Program and Export Status.

1) Dr. Henry Stoklosa, E.I. DuPont de Nemours. Dr. Stoklosa has been affiliated with a group at DuPont involved with computer applications to chemical problems, including computer-aided organic synthesis. He will soon be involved in another group which might also be able to make use of the REACT program in addition to his more general interest in CONGEN.

2) Dr. G.W.A. Milne, National Institutes of Health. Dr. Milne is currently in charge of the National Institutes of Health contribution to the NIH/EPA Chemical Information System. His interests included not only evaluation of the utility of the program but also exploration of ways in which CONGEN might be interfaced to the Chemical Information System. This effort is described in more detail in Section 2.3. Dr. Milne offered both praise and criticism. He praised the program itself but wants good user-level documentation and a large number of sample problems for future workshops.

We are currently exploring with him the interest of NIH in obtaining a version of the current program for use on the NIH PDP-10 facility.

3) Dr. William Brugger, International Flavors and Fragrances. Dr. Brugger represents the key person at IFF Research responsible for computer applications in their laboratories. Structure elucidation is a major activity of this company not only in analysis of natural and synthetic products but also in assessing the relationships between chemical structure and toxic properties affecting human health. Dr. Brugger has access to both DEC VAX and IBM computers, the former representing the laboratory computer. The status of a version of CONGEN for these machines has been discussed previously. Meanwhile Dr. Brugger and his colleagues are evaluating CONGEN at SUMEX via the GUEST access facility.

4) Dr. Douglas Dorman is head of the NMR laboratory at Lilly Research Laboratories and works closely with mass spectroscopists and other chemists in solving structures of a variety of compounds related to existing or new products. Dr. Dorman has been familiar with the "old" (non-exportable) version of CONGEN and

thus was able to critique the new program not only on its merits but also on comparison with the old version. His detailed critique is included in Appendix III. It is worthwhile to point out that we are in the process of implementing two of the important new features he mentions because of their general importance. The SURVEY and EXAMINE commands will allow both user-specified features used to explore structural possibilities and Boolean statements used to select structures with combinations of features. The MSRANK facility will perform the selection of plausible structures based on agreement of predicted and observed spectra. Dr. Dorman has access to a PDP-10 computer operating under the TOPS-10 operating system. About two weeks ago he received a copy of the exportable CONGEN and is now using the program in his own laboratory. He has agreed to provide us with continuing comments and criticism in exchange for receiving updated versions of the program.

5) Dr. Jon Clardy, Cornell University. Dr. Clardy is a recognized leader in development and applications of the technique of X-ray crystallography in structure elucidation. His attendance of the workshop was based on an interest in learning about alternative, computer-based approaches to the problem. As his letter points out, the ability to use CONGEN to help solve structures before expenditure of time and effort in X-ray analysis would be an important benefit. A more important outcome of the workshop for future research were discussions on the possibility of coupling CONGEN-suggested structures to Patterson search techniques. In principle, each of the candidates suggested by CONGEN could be used in turn to guide the search of the electron density maps for a fitting of the structure to the maps. The correct structure should yield the least ambiguous fit. Dr. Clardy will access CONGEN at SUMEX via the GUEST facility. His ability to use CONGEN at Cornell depends on the success of the efforts of Mr. In Ki Mun (next section).

6) Mr. In Ki Mun, Cornell University. Mr. Mun attended the workshop representing Prof. Fred McLafferty at Cornell. Prof. McLafferty's group has had for many years an interest in use of computer techniques to help solve structures, based primarily on mass spectral data. His research in this area has led to programs which suggest the presence of functionalities in an unknown molecule. CONGEN can, in

principle, complete such a schema for analysis by piecing together the inferred functionalities. Mr. Mun attended to explore the feasibility of this approach and to learn how best to use CONGEN on the IBM system at Cornell. He is currently evaluating the effort of modifying CONGEN for an IBM version of BCPL. If successful then this version should be available for other persons who have good interactive services available at their IBM installations.

7) Dr. Reimar Breuning, Munich. Dr. Breuning learned of the existence of the workshops from discussions with Prof. Djerassi at the IUPAC meeting on natural products. Dr. Breuning had the opportunity to attend as he was in the process of arranging a post-doctoral appointment with Prof. Nakanishi (see next section). He is actively involved in natural products structure elucidation at Munich and expects these interests to continue. Dr. Breuning can access CONGEN via Guest for the near future from Munich. At Columbia he will have the opportunity to take advantage of access to the facility there.

8) Dr. David Lynn, Columbia University. Dr. Lynn attended the workshop representing Prof. Koji Nakanishi, the latter a recognized expert in the area of structure elucidation of a number of classes of natural products of relevance to human health. Dr. Lynn is to act as the focal point for introduction of the computer methods to that research group. Prof. Nakanishi's group has access to a DEC-20 system operating under the TOPS-20 operating system. We are currently arranging for a version of CONGEN to be installed there. We anticipate no problems because of our successful experiment during the workshop of running CONGEN on a DEC-20 system at Rutgers.

9) Dr. Y. Gopichand, University of Oklahoma. Dr. Gopichand attended the workshop representing the marine natural products group of Prof. Francis Schmitz. This group specializes in structure elucidation of halogenated terpenoid molecules possessing a variety of biological activities and marine sterols representing intermediates or end products in steroid biosynthesis. As evidenced by the letter of critique, this group represents an excellent example of classical approaches to structure elucidation; sufficient data are collected such that the number of structural possibilities is reduced to a very small number. As the letter also

indicates, CONGEN should be useful at least to check the rigor of their structural assignment. What should prove more interesting is whether or not such a group, with little computer expertise, can use CONGEN earlier on in the process of structure elucidation to guide subsequent collection of data.

10) Ms. Wendy Harrison, University of Hawaii. Ms. Harrison attended the workshop representing the marine natural products group of Prof. Paul Scheuer at Hawaii. This group is engaged in structure elucidation problems similar to those encountered in Prof. Schmitz's laboratory, although focus is on different classes of organisms. Their letter of critique mentioned that due to absence of critical high resolution mass spectrometric data to establish molecular formulas, they have been unable to use CONGEN this past month to help on their problems. This situation is expected to be resolved soon. Prof. Scheuer plans to use CONGEN as an aid in a course in structure elucidation this coming semester. One difficulty is that this group only has access to a DEC PDP-11 system. This introduces all of the problems mentioned earlier on a version of the program for mini-computers, including proliferation of manufacturers and operating systems. For example, there is a version of BCPL for PDP-11 series machines, but only under the RT-11 operating system, whereas the Hawaii group runs under The RSX-11 operating system. For the near future, access from Hawaii will be to CONGEN at SUMEX running under the GUEST directory.

11) Dr. Laszlo Tokes and Dr. Michael Maddox, Syntex Research. Drs. Tokes and Maddox are, respectively, in charge of the mass spectrometry and NMR laboratories at Syntex Research. They are responsible for the majority of structure elucidation problems which rely on physical methods. Their interest in CONGEN is that it might help them solve certain problems in less time than required by manual methods. Syntex research has access only to laboratory mini-computers such as the DEC PDP-11 series machines. They, too, must await a smaller version of CONGEN suitable for mini-computers before being able to use the program in their own laboratory.

12) Dr. John Figueras, Kodak Research Laboratory. Dr. Figueras attended representing the Analytical Sciences Division of Kodak's Research Laboratory. This

division is responsible for data collection and analysis in support of the structure elucidation activities of the Laboratory including not only new developments in the photographic process but also the new technology of thin-film bound enzymes systems for clinical analyses. His role was to evaluate what part CONGEN could play in the on-going automation of of the Division. The Division will be obtaining a DEC-20 system in the near future on which CONGEN will run directly. Meanwhile we have offered GUEST access in return for continuing critique on utility of CONGEN for large, poly-heteroatomic, aromatic molecules of the types encountered in their research.

13) Dr. Charles Snelling, University of Illinois. Dr. Snelling attended the workshop representing Prof. Kenneth Rinehart in the Chemistry Dept. at Illinois. Prof. Rinehart is also an acknowledged expert in structure elucidation with emphasis on macrolide antibiotics, halogenated terpenoids and other classes of natural and synthetic products of relevance to human health problems. Dr. Snelling had several comments and criticisms about the difficulties facing a novice user of such a complicated system. Although he found the program useful and wishes to get it running at Illinois, he would like to see some changes made to simplify use of the program for the chemist. There are many computer systems available at Illinois and choice of which system on which to mount CONGEN will depend on the ease of the task and access to a particular system. This is currently under study by them.

14) Dr. Gilles Moreau, Roussel UCLAF. Dr. Moreau attended the workshop representing the French pharmaceutical concern Roussel UCLAF. This company maintains an active group in computer applications in chemistry and wished to evaluate CONGEN for its use in their structural problems. The company is quite interested and will explore use of the program via GUEST access. They are concerned about issues of secrecy for new problems and we have made it quite clear that GUEST access represents public knowledge of their problems. Therefore, assuming their interest continues, we will be arranging some alternative to SUMEX for use of CONGEN. They have access to IBM equipment and we are awaiting further description of the form of access possible.

15) Prof. Andre Dreiding, Zurich. Dr. Dreiding

has been interested in both the problem-solving and the pedagogical aspects of CONGEN for some time. He had previously used the old version and was gratified to see the improvements in the new version. He would like to see much more attention paid to the actual structures (i.e., in three dimensions) of molecules rather than simply their constitutions as CONGEN, with the exception of the STEREO command, currently represents structures. We are, of course, working very hard to introduce concepts of stereochemistry into our computational procedures. Prof. Dreiding will probably be able to access CONGEN at Zurich on an existing PDP-10 installation. If so, export to his group will be trivial.

16) Dr. James Shoolery and Dr. Michael Gross, Varian Associates. Dr. Shoolery is in charge of Varian's NMR application laboratory and Dr. Gross is in charge of computer software for Varian's NMR/computer systems. Their respective interests match their responsibilities. Dr. Shoolery feels that CONGEN could be a valuable assistant in helping solve structures based primarily on proton and carbon NMR data. In fact, he has been able to demonstrate such an approach on some recent problems of persons outside Varian. Dr. Gross is interested in a mini-computer version of CONGEN for incorporation into their existing data system. Although we cannot ourselves support such an effort, we have agreed to provide them with program listings and documentation to explore the feasibility of a small machine version.

17) Dr. Daniel F. Chodosh, Smith, Kline and French. Dr. Chodosh was not actually invited to the workshop, but happened to visit our group during one of the sessions. He was sufficiently impressed that he procured a tape to carry away a copy of the program with him. He has now been supplied with a version of CONGEN for the PDP-10 and as of last week has it running at the SKF research laboratories in Philadelphia. This is an interesting experiment because as an informal attendee of a small portion of the workshop, he is learning the program almost from scratch based on existing help facilities and documentation. He will begin introducing others to the program when he has developed sufficient familiarity to be comfortable with the program.

In addition, the following persons during the past year have asked

for information about and access to CONGEN. For the most part we have granted access through the GUEST directory, setting up an account only for those users with more than occasional log-ins.

Dr. David Cowburn
Physical Biochemistry
The Rockefeller University
New York City

We have sent Dr. Cowburn information on access to CONGEN and are currently discussing how to use some of our computational methods or extensions to them for assistance in his problems of elucidating peptide conformations.

Douglas Henry
School of Pharmacy
Oregon State University
Corvallis, Oregon

He has been sent our programs for structure drawing for use on his own computer.

The following have asked for and received information on access to CONGEN at SUMEX via the GUEST facility.

Dr. H. Kating
Institut fur Pharmazeutische Biologie
Der Universitat
Bonn, Germany

Dr. Kerber
Lehrstuhl D fur Mathematik
Aachen, Germany

Dr. Brenda J. Kimble
Radiobiology Laboratory
University of California
Davis, California

Dr. J. Neubuser
Lehrstuhl D fur Mathematic
Aachen, Germany

Dr. George Padilla
Dept. of Physiology
Duke University Medical Center
Durham, N.C.

Dr. W. Sieber
Sandoz Ltd.

Basel, Switzerland

Dr. Babu Venkataraghavan
Lederle Laboratories
Pearl River, New York

We also helped him bring up the Fortran draw program on the DEC-10 system at Lederle

Dr. Stephen Wilson
Dept. of Chemistry
Indiana University
Bloomington, Indiana

Appendix II. Sample Letter of Invitation Sent To Prospective Workshop Attendees.

COPY

STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DEPARTMENT OF CHEMISTRY

August 24, 1978

Professor Kenneth Rinehart
Department of Chemical Science
University of Illinois
Urbana, Illinois 61801

Dear Professor Rinehart:

I am writing to determine your interest in a mini-workshop we plan to hold at Stanford on use of computer programs for computer-assisted structure elucidation. Over the past three years we have been involved in a research effort directed in part to exploring the feasibility and utility of interactive computer programs as tools to help chemists solve unknown structures. The most highly developed of these research tools is the CONGEN program, and I know that you have in one way or another been exposed to this program.

We have made CONGEN available on the SUMEX computer here at Stanford via communications networks to a selected group of chemists because we know that the only way to improve the program and to make it useful is to apply it to a variety of real structural problems in chemistry. We have learned a great deal from this experience and have designed a production version of CONGEN with the primary goal of eliminating the deficiencies in the research version. In particular the new program is much simpler to use, much smaller and faster and can be exported to certain other computers.

There is a significant amount of work which remains to be done to make the exportable version of CONGEN a truly useful chemist's "assistant". Some of this work is underway now. But before investing a great deal of time and effort in polishing a new version to make it more useful and acceptable, we would like to expose a group of several chemists experienced in structure elucidation to this version of the program and base future research and development efforts on their perceptions of deficiencies remaining in the program and its ultimate utility. We feel its usefulness can be significant in terms of exploring alternative structural possibilities and in guaranteeing no plausible alternative has been overlooked. However, in order to demonstrate that utility we need assistance in developing a version which large numbers of chemists can use easily and productively in their own laboratories and on their own computers. We feel that by participating in this workshop you can contribute in an important way.

We plan an approximately four to five day informal workshop here at Stanford, to be attended by you or one of your experienced co-workers. During this time you would be able to use the exportable version of CONGEN here at SUMEX to work on recent structural problems encountered in your laboratory, which we hope will include some which would be

unknowns. In turn, we would be learning how to finish development of the first version of the program for export to your laboratory with some guarantee that major problems or deficiencies would be eliminated based on the experience of the workshop. We have not yet established a time for the workshop, but we are tentatively considering a date between October and December, 1978. We are also trying to arrange for travel support for non-industrial persons. We are seeking now only an expression of interest on your part.

There are only two requirements on your part, other than your expressed interest in participating. The first requirement is that you must be capable of accessing CONGEN at SUMEX or preferably be capable of running CONGEN on your own computer system. The second requirement is that you actually be interested in using the program in the future (assuming the workshop was worthwhile) to provide us with some feedback on whether improvements meet your criteria for acceptance.

Currently the new version of CONGEN can be used on Digital Equipment Corp. PDP-10's and 20's with little problem. We are now exploring use of IBM 360 and 370 series machines. We have already demonstrated that small segments of the program run on suitably configured IBM computers. However, for CONGEN to be used interactively requires a time-sharing operating system. If you have access to an IBM computer but are uncertain about the operating system, find out from your computer systems people the name of the operating system and call or write to us with that information.

While at Stanford you would also be able to take a close look at some of our other research efforts which are not quite so far along as CONGEN, including a recently finished stereochemical structure generator, the REACT program and a variety of tools for ranking structures based on comparison of predicted and observed spectra, facile examination of large sets of structural candidates for common and unique features and methods for assistance in experiment planning. We would appreciate your evaluation of these efforts also. The main emphasis, however, will be on CONGEN.

Obviously we do not expect a firm commitment for attendance at this time, but if at all possible I would appreciate a reply. Indications of your interest are important in this regard.

Yours sincerely,

Carl Djerassi
Professor of Chemistry