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 will emphasize in the new grant period 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 research 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 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 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 have embarked on a concerted effort to extend to the scientific community access to our programs, and critical parts of our research effort 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 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.

I.B. Medical Relevance and Collaboration

The medical relevance of our research lies in the direct relationship between molecular structure and biological activity. The sciences of chemistry and biochemistry rest on a firm foundation of the past history of well-characterized chemical structures. Indeed, structure elucidation of unknown compounds and the detailed investigation of stereochemical configurations and conformations of known compounds are absolutely essential steps in understanding the physiological role played by structures of demonstrated biological activity. Our research is focussed on providing computational assistance in several areas of structural chemistry and biochemistry, with primary attention directed to those

aspects of the problem which are most difficult to solve by strictly manual methods. These aspects include exhaustive and irredundant generation of constitutional isomers, and configurational and conformational stereoisomers under chemical, biological and spectroscopic constraints with a guarantee that no plausible stereoisomer has been overlooked.

Although our programs can be applied to a variety of structural problems, in fact most applications by our group and by our collaborators are in the area of natural products, antibiotics, pheremones and other biomolecules which play important biochemical roles. In discussions of collaborative investigations involved with actual applications of our programs we have always stressed the importance of strong links between the structures under investigation and the importance of such structures to health-related research. This emphasis can be seen by examination of the affiliations of current DENDRAL-related investigators and the brief description of current collaborative efforts in Interactions with the SUMEX-AIM Resource.

I.C. Highlights of Research Progress

In this section we discuss briefly some major highlights of the past year and research currently in progress.

I.C.1. Past Year

1) Exportable version of the CONGEN program for computer-assisted structure elucidation. CONGEN is an interactive computer program whose task is to provide to the structural biochemist all chemical structures which are possible candidates for the structure of an unknown chemical compound. Based on this information, experiments can be designed to pinpoint the correct structure, thereby facilitating rapid and unambiguous identification of novel, bioactive chemicals. During the previous grant year we have completed an exportable version of the CONGEN program and have begun to export it to a variety of structural analysis laboratories in academic, private and industrial research organizations. CONGEN is being utilized at Stanford and at export sites in the hands of investigators who use it as a tool in solving their own structural problems. Even though we have been exporting versions of CONGEN for only six months, already the program has been used for new structures and recent results have formed the basis for at least four formal lectures by users of CONGEN at remote sites.

2) Version I of the GENOA program for structure generation with overlapping atoms. GENOA is an outgrowth of CONGEN whose purpose is to suggest candidate structures for an unknown based on redundant and ambiguous structural inferences. This program, which utilizes CONGEN as an integral part of the computational procedures, is far simpler to use by the practicing biochemist. This results from GENOA's capability to construct structures based on substructural information obtained from a variety of spectroscopic, chemical and biochemical techniques. The program itself considers the structural implications of each new piece of structural data and automatically ensures that all overlaps are considered, thereby freeing the investigator from concerns about the potential for overlapping, or redundant substructural information. In addition, GENOA is the ideal tool

for interfacing to automated procedures for spectral interpretation, because the necessity for manual intervention in the assignment of substructures is no longer required as it was for CONGEN.

3) Exhaustive and irredundant generation of stereoisomers. During the current grant period we have solved the problem of computer generation of configurational stereoisomers. These are isomeric chemical structures that differ from one another in the arrangement of atoms in three-dimensional space. Previously, CONGEN and GENOA were capable only of generation of constitutional isomers which convey no information about the structure in three dimensions. The interaction of biomolecules with biochemical systems is based on their three dimensional nature, not simply their constitution. Therefore, this new development is crucial to use of computational techniques in structural studies. It is interesting to note that this particular problem remained unsolved, until the present work, since it was originally proposed by Van't Hoff more than 100 years ago.

I.C.2. Research in Progress

1) Programs for Interpretation and Prediction of Spectral Data. We are actively pursuing several novel approaches to the automated interpretation of spectral data, concentrating on carbon-13 magnetic resonance (CMR), proton magnetic resonance (PMR) and mass spectral (MS) data. These approaches utilize large data bases of correlations between substructural features of a molecule and spectral signatures of such features. Our approaches are unique in that: 1) we can incorporate stereochemical features of substructures into the data bases; and 2) we can use the same data bases for both interpretation and prediction of data.

The stereochemical substructure descriptors are absolutely essential, especially in magnetic resonance data, for either interpretation or prediction. Resonance positions are a strong function of the local environment of a resonating atom, including position in space relative to other neighboring atoms. Descriptors which include the three dimensional relationships among atoms in a substructure are required in order to obtain meaningful correlations.

The data bases can be used to interpret spectral data to obtain substructures to be used in CONGEN and GENOA, the structure generating programs. Automation of this aspect of structure elucidation could significantly ease the burden on the structural biochemist because the computer-based files are much more comprehensive and easier to use than correlation tables or diffuse literature sources. The same data bases can be used to predict spectral signatures in the context of a set of complete molecular structures. Comparison of predicted and observed spectra allows a rank-ordering of candidates and will be very useful in directing the attention of the investigator to the most plausible alternatives.

This effort marks the beginnings of the SASES system, a general, automated system for computational assistance in several phases of structure elucidation.

2) Constrained Generation of Configurational Stereoisomers. We have just completed an experimental version of a program, designed to be used with the structure generation programs CONGEN and GENOA, capable of constrained generation of stereoisomers. This means that, for the first time, a computer program can be used to begin with the molecular formula of an unknown compound and using constraints on both molecular connectivity and configuration arrive at a set of structural alternatives which include potential stereochemical variability. This capability allows use of spectral data whose interpretation (see Highlight 1) depends strongly on stereochemical features of molecules. Most importantly, it gives us a structural representation and methods for structure generation and manipulation which represent the foundations for future developments of the one important remaining aspect of structural analysis, treatment of molecular conformations.

I.D. List of Recent Publications

(1) 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.

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

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

(4) 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).

(5) 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).

(6) 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).

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

(8) 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.

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

(10) T.H. Varkony, Y. Shiloach, and D.H. Smith, "Computer-Assisted Examination of Chemical Compounds for Structural Similarities," J. Chem. Inf. Comp. Sci., 19, 104 (1979).

(11) J.G. Nourse and D.H. Smith, "Nonnumerical Mathematical Methods in the Problem of Stereoisomer Generation," Match, (No. 6), 259 (1979).

(12) N.A.B. Gray, R.E. Carhart, A. Lavanchy, D.H. Smith, T. Varkony, B.G. Buchanan, W.C. White, and L. Creary, "Computerized Mass Spectrum Prediction and Ranking," Anal. Chem., in press (1980).

(13) A. Lavanchy, T. Varkony, D.H. Smith, N.A.B. Gray, W.C. White, R.E. Carhart, B.G. Buchanan, and C. Djerassi, "Rule-Based Mass Spectrum Prediction and Ranking: Applications to Structure Elucidation of Novel Marine Sterols," Org. Mass Spectrom., in press (1980).

(14) J.G. Nourse, D.H. Smith, and C. Djerassi, "Computer-Assisted Elucidation of Molecular Structure with Stereochemistry," J. Am. Chem. Soc., submitted for publication.

(15) J. G. Nourse, "Applications of Artificial Intelligence for Chemical Inference. 28. The Configuration Symmetry Group and Its Application to Stereoisomer Generation, Specification, and Enumeration.", J. Amer. Chem. Soc., 101, 1210, (1979).

(16) J. G. Nourse, "Application of the Permutation Group to Stereoisomer Generation for Computer Assisted Structure Elucidation.", in "The Permutation Group in Physics and Chemistry", Lecture Notes in Chemistry, Vol. 12, Springer-Verlag, New York, (1979), p. 19.

(17) J. G. Nourse, "Applications of the Permutation Group in Dynamic Stereochemistry" in "The Permutation Group in Physics and Chemistry", Lecture Notes in Chemistry, Vol. 12, Springer-Verlag, New York, (1979), p. 28.

(18) J. G. Nourse, "Selfinverse and Nonselfinverse Degenerate Isomerizations," J. Am. Chem. Soc., in press (1980).

(19) N. A. B. Gray, A. Buchs, D. H. Smith, and C. Djerassi, "Computer-Assisted Structural Interpretation of Mass Spectral Data," Helv. Chim. Acta, submitted for publication.

(20) N. A. B. Gray, C. W. Crandell, J. G. Nourse, D. H. Smith, and C. Djerassi, "Computer-Assisted Interpretation of C-13 Spectral Data," J. Org. Chem., in preparation.

(21) N. A. B. Gray, J. G. Nourse, C. W. Crandell, D. H. Smith, and C. Djerassi, "Stereochemical Substructure Codes for C-13 Spectral Analysis," Org. Magn. Res., in preparation.

I.E. Funding Support

I.E.1. Title

RESOURCE RELATED RESEARCH: COMPUTERS IN CHEMISTRY (grant)

I.E.2. Principal Investigator

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

Dennis H. Smith (Associate Investigator), Senior Research Associate, Department of Chemistry, Stanford University

I.E.3. Funding Agency

Biotechnology Resources Program, Division of Research Resources, National Institutes of Health

I.E.4. Grant Identification Number

RR-00612-11

I.E.5. Total Award and Period

Total - 5/1/80 - 4/30/83 --------- $641,419

I.E.6. Current Award and Period

Current - 5/1/80 - 4/30/81 -------- $221,255

II. Interactions with the SUMEX-AIM Resource

In the coming period of our research, our computational approaches to structural biochemistry will become 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 (see Justification and Requirements for Continued SUMEX Use for additional details).

For these reasons, in our new grant period we have identified the SUMEX-AIM resource as the resource to which our research is related. 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 is one which goes far beyond mere consumption of cycles on the SUMEX machine. It has been the goal of the SUMEX project 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 in the DENDRAL Project to yield more generally useful programs will simply magnify the problem.

We will, therefore, under the new relationship between SUMEX-AIM and our project, 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 utilize the small machines requested as part of the SUMEX renewal. Our project will benefit by being able to provide more extensive operational access to our existing and developing programs using these machines, and to provide a test environment for adapting our programs to a more realistic laboratory computing environment than the special-purpose SUMEX resource (see Justification and Requirements for Continued SUMEX Use for additional information). SUMEX will benefit by moving a substantial part of the DENDRAL production load to more cost-effective systems, thereby freeing the SUMEX resource for new program 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 new machines. New program developments will 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 small machines as integral part of the resource sharing aspects of our relationship to SUMEX. The DENDRAL project is one of the first of the SUMEX-AIM projects to have developed sufficient maturity to require additional computer facilities to support production use and to facilitate export of its programs to be applied to real-world, biomedical structural problems. 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 realistic mechanisms for dissemination of their software to meet the computational needs of their collaborators. Cooperating with SUMEX in the use of small machines, 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 explore the utility of machines which will be much more

widely available in Department and laboratory environments than DEC-10's and -20's. There are currently very few 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 share of the SUMEX-AIM computing resource.

II.A. Scientific Collaboration and Program Dissemination

II.A.1. Scientific Collaborations

Several of our research goals involve problems in structural analysis whose solution is of interest to other research groups with specific, health-related problems in structural biochemistry. The following is a brief description of collaborative efforts that have been taking place or will soon commence in the use of DENDRAL programs for various aspects of structural analysis.

1. Dr. David Cowburn, The Rockefeller University. 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 will develop this application in 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. CONGEN equipped with a conformation generator, would likely be useful to Prof. Cowburn's research in at least three ways:

a) 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.

b) 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.

c) 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.

2. Dr. Gilda Loew, Stanford Research Institute and The Rockefeller University. Since our 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 will develop this link by collaborating with Prof. Gilda Loew and her group. 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. 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.

a) 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.

b) Professor Loew has been considering the conformational variations caused by the intercalation of ethidium into nucleic acids. 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.

3. Drs. Larry Anderson and Elliott Organick, Depts. of Fuels Engineering and Computer Science, University of Utah. Dr. Anderson's research is in establishing the structure of coal and related polymers via various thermal and chemical degradation schemes. The degradation products are of interest to both energy and environmental studies. Professor Organick is responsible in part for the computer and graphics facility on which CONGEN and related programs can be run. We are exploring with them structure representations based on the Superatom concept in CONGEN as a means of representing families of structures. Access to our programs is primarily via the computer facility at Utah.

4. Dr. Raymond Carhart, Lederle Laboratories. Dr. Carhart (a former member of our group) is engaged in research concerned with computer applications to structure/activity relationships. Program development is done jointly between Lederle and Stanford with free exchange of software. Lederle applications are carried out on their own computer facility.

5. Dr. Janet Finer-Moore, University of Georgia. Dr. Finer-Moore is engaged in structure analysis of alkaloids in Dr. Peletier's group at Georgia. This research makes extensive use of 13C NMR. Our collaboration involves the development and application of our 13C interpretive and predictive programs in structure elucidation of new compounds based on an extensive set of 13C data available on closely related compounds. Access is via network to our programs at Stanford. Recent use of our programs has aided her in correcting erroneous assignments of 13C resonance shifts to known structure and aided in the solution of the structures of new diterpenoid alkaloids.

6. Dr. Brenda Kimble, University of California, Davis. Dr. Kimble's research is in structural analysis of compounds which are present in trace amounts in environmental milieus and which show mutagenic activity. Many of these compounds are largely aromatic. We are developing the capabilities of our programs to deal efficiently with large, polynuclear aromatic compounds. Access to our programs is via network to Stanford.

7. Dr. Fred McLafferty, Cornell University. Dr. McLafferty's research is involved with instrumental and analytical aspects of mass spectrometry. We are working with him on the development and application of an interface between his STIRS system and CONGEN/GENOA for structure determination based on mass spectral data. Part of this collaboration is development of IBM versions of some of our programs. Access is in part to Stanford, shifting primarily to Cornell as development proceeds.

II.A.2. Program Dissemination

Because one of our goals is dissemination of our programs to a wide community of collaborators, we have made use of several of the mechanisms provided by SUMEX-AIM to introduce new investigators to our work and to encourage close collaboration in the study of important structural problems. Generally speaking, introduction of new persons and the development of collaborative projects has followed the course outlined below:

1) GUEST Access. The GUEST account mechanism of SUMEX-AIM is normally used when persons from the outside community contact us to learn more about our programs. We provide to them a special packet of information on network access and connection to the GUEST account, together with documentation of specific programs in which they are interested. This is a simple way of performing a "try it and see" experiment to determine the utility of the programs to the individual investigator. The following persons have used this method of access the past year:

Dr. Robert Adamski - Alcon Labs

Dr. A. Bothner-by - Carnegie Mellon University

Dr. Reimar Bruening - Institut fur Pharmazeutische Arzneimittellehre der Universitaet, West Germany

Dr. William Brugger - International Flavors and Fragrances

Dr. Raymond Carhart - Lederle Laboratories

Dr. Robert Carter - University of Lund, Sweden

Dr. Francois Choplin - Institut Le Bel, France

Dr. Jon Clardy - Cornell University

Dr. Mike Crocco - American Hoechst Corp.

Dr. V. Delaroff - Roussel UCLAF, France

Dr. Dan Dolata - University of California at Santa Cruz

Dr. Bruno Frei - Laboratorium f. Organische Chemie, Switzerland

Dr. Y. Gopichand - University of Oklahoma

Ms. Wendy Harrison - University of Hawaii at Manoa

Dr. Richard Hogue - University of California at Santa Cruz

Dr. David Lynn - Columbia University

Dr. In Ki Mun - Cornell University

Dr. Koji Nakanishi - Columbia University

Dr. Suba Neir - Washington University, St. Louis

Dr. J.D. Roberts - California Institute of Technology

Dr. Joseph SanFilippo - Rutgers University

Dr. Babu Venkataraghavan - Lederle Laboratories

Dr. W.T. Wipke - University of California at Santa Cruz

Dr. Michael Zippel - Institut fur Biochemie Zentrale
Arbeitsgruppe Spectroskopie, Germany

2) EXODENDRAL Accounts. SUMEX-AIM has set aside a special account group called EXODENDRAL designed to give each collaborator, whose initial GUEST experience has proven fruitful, an account of his or her own. These accounts facilitate both access to a variety of our experimental programs (not generally available through GUEST) and communication using the various message and bulletin board programs. For persons who use exportable versions of our programs on their own computer facilities, EXODENDRAL accounts are used primarily for rapid contact and exchange of messages.

Dr. Jean-Claude Braekman - Universite Libre de Bruxelles,
Belgium

Dr. Hartmut Braun - Organische-Chemisches Institut der
Universitaet Zurich, Switzerland

Dr. Roy Carrington - Shell Biosciences Laboratory, England

Dr. David Cowburn - The Rockefeller University

Dr. Douglas Dorman - Lilly Research Laboratories

Dr. Andre Dreiding - Organische-Chemisches Institut der
Universitaet Zurich, Switzerland

Dr. Janet Finer-Moore - University of Georgia

Dr. Kenneth Gash - California State College at Dominguez Hills

Dr. Steven Heller - Environmental Protection Agency

Dr. Martin Huber - Ciba-Geigy, Switzerland

Dr. Peter W. Milne - CSIRO Division of Computing Research,
Australia

Dr. James Shoolery - Varian Associates

Dr. William Sieber - Sandoz Ltd., Switzerland

Dr. Mark Wood - Rutgers University

3) Program Export. SUMEX-AIM is also the facility which is used to develop and perform experiments with exportable versions of our programs. Wherever possible we encourage collaborators to run our programs on their own computers to decrease the computational burden on SUMEX-AIM as much as

possible. This year we have distributed CONGEN to a number of laboratories owning computers on which the exportable version can now execute. These currently include DEC PDP-10 and -20 systems operating under the TENEX, TOPS-10 and TOPS-20 operating systems, and more recently, the beginnings of a version for IBM systems. The following persons are currently running CONGEN on their own laboratory computers:

Dr. Larry Anderson - University of Utah

Dr. Hartmut Braun - Organische-Chemisches Institut der Universitaet Zurich, Switzerland

Dr. Raymond Carhart - Lederle Laboratories

Dr. Roy Carrington - Shell Biosciences Laboratory, England

Dr. Robert Carter - University of Lund, Sweden

Dr. Daniel Chodosh - Smith, Kline & French Laboratories

Dr. Douglas Dorman - Lilly Research Labs

Dr. Martin Huber - Ciba-Geigy, Switzerland

Dr. Carroll Johnson - Oak Ridge National Laboratory

Dr. G. Jones - ICI Pharmaceuticals, England

Dr. Peter W. Milne - CSIRO Division of Computing Research, Australia

Dr. James Morrison - Latrobe University, Australia

Dr. Fred W. McLafferty - Cornell University

Dr. David Pensak - E.I. duPont de Nemours and Company

Dr. Gretchen Schwenzer - Monsanto Agricultural Products Co.

Dr. William Sieber - Sandoz, Ltd., Switzerland

Dr. M.D. Sutherland - University of Queensland, Australia

Dr. R.O. Watts - Australian National University

4) Industrial Affiliates Program. The high level of interest shown by industrial research laboratories in our programs has always presented us with delicate questions about access to SUMEX-AIM. In the past we have granted access for trials of our programs under the conditions that access is necessarily limited and that the recording mechanisms of our programs be used to ensure that all such trial use be in the public domain. As of

April, 1980, we have begun solicitation of interested industrial organizations to participate in a DENDRAL Project Industrial Affiliates Program. We intend to use this program as a means by which we can offer collaborations with our on-going research to industrial organizations separate from SUMEX-AIM. Although EXODENDRAL accounts to such organizations may be used to facilitate communication and sharing of new programs and concepts of interest with the community as a whole, all significant and certainly all proprietary use of our programs will be carried out on their own computational facilities. As of the writing of this portion of the SUMEX-AIM renewal proposal we have not had any organizations formally take up membership.

II.B. Interactions with Other SUMEX-AIM Projects

We routinely collaborate with other projects on SUMEX most closely related to our own research. In particular, these collaborations have taken place with the CRYSALIS project, MOLGEN, SECS and have begun with Dr. Carroll Johnson at Oak Ridge.

CRYSALIS is concerned with new approaches to the interpretation of X-ray crystallographic data. X-ray crystallography is another approach to molecular structure elucidation. One of our long-term interests is exploring ways in which CONGEN or GENOA generated structures might be used to guide the search of electron density maps. We are also communicating with Prof. Jon Clardy at Cornell on this problem. It is hoped that having narrowed down the structural possibilities for an unknown using physical and chemical data, the few remaining candidates can be used to guide interpretation of such maps.

Most of the structural problems investigated by MOLGEN involve much larger molecules than the size normally investigated in DENDRAL research. Thus, structural representations involving higher levels of abstraction are of utility in MOLGEN, making our structure manipulation tasks quite different. However, many of the ways in which MOLGEN manipulates its structural representations drew on past experience in DENDRAL in developing algorithms to perform these manipulations.

We collaborate frequently with the SECS project in a number of ways. Although our research efforts are in one sense directed toward opposite ends of work on chemical structures, SECS being devoted to synthesis, DENDRAL being devoted to analysis, the underlying problems of structural manipulation share many common aspects. We have exchanged software where possible, particularly in the area of chemical structure display. We have held several discussions in joint group meetings and at several symposia including the AIM Workshops on common problems, including substructure searching, canonical representations and representation and manipulation of stereochemistry. Persons visiting one laboratory often take the opportunity to visit the other. For example, recent visitors to both laboratories have included Prof. Andre Dreiding, Zurich, Dr. Martin Huber, Basel, and Prof. Robert Carter, Lund.

Dr. Carroll Johnson has collaborated on the CRYSALIS project in the past. More recently he has taken an interest in the use of knowledge-based

programs for certain problems in spectral data interpretation. For this reason he is exploring the AGE and EMYCIN systems as frameworks for his program structure, and is involved in discussions with DENDRAL to see where common areas of data interpretation can be identified so that he can draw on our experience and programs. This effort is just beginning at this time; we plan to meet early in May at Stanford to continue discussions.

II.C. Critique of Resource Management

The SUMEX-AIM environment, including hardware, system software and staff, has proven absolutely ideal for the development and dissemination of DENDRAL programs. The virtual memory operating system has greatly facilitated development of large programs. The emphasis on time-sharing and interactive programs has been essential to us in our development of interactive programs. Our experience with other computer facilities has only emphasized the importance of the SUMEX environment for real-world applications of our programs. To run CONGEN, for example, in a batch computing environment would make no sense whatever because the program (and our other, related programs) is successful in large part because an investigator can closely monitor and control the program as it works toward solution. We have no complaints whatsoever about the computing environment.

We do have, however, significant problems with SUMEX-AIM capacity, both in available computer cycles and on-line file storage. In a sense DENDRAL suffers from its success. The rapid progress made during the last grant period and now continuing into the next period has led to development of many new programs as adjuncts to CONGEN and GENOA and at the same time has inspired many persons in the scientific community to request some form of access to our programs. The net result is that it is often very difficult to carry on at the same time development and collaborations involving applications of our programs to structural problems due to high load average on the system.

The current overcrowding we see on SUMEX creates two major problems for us in the conduct of our research. First, it diminishes productivity as many people compete for the resource; the "time-sharing syndrome" leads to idle, wasted time at the terminal waiting for trivial computations to be completed. Second, the slow response time of the system is an aggravation to an outside investigator who is anxiously trying to solve a structural problem. At some point even the most interested persons will give up, log off the computer and resort to manual methods where possible.

We have taken many steps within our project to try to work around heavy use periods on SUMEX. Our group works a staggered schedule, both in terms of the actual hours worked each day and in terms of what days each week are worked. This results in some problems in intra-group communication, but fortunately the message and other communication systems of SUMEX help alleviate that situation. We try to run all demonstrations on the DEC-2020 to help ease the burden on the dual KI-10 system. We encourage our collaborators to avoid prime-time use of the system when possible.

For these reasons, we strongly support the proposed augmentation of the SUMEX-AIM hardware. Any part of our computations which can be shifted to another machine will not only facilitate export of our software but will ease the load on the DEC-10s and make it easier to continue our research. Both will serve to make SUMEX more responsive and our productivity higher.

III. Research Plans

III.A. Project Goals and Plans

Current research efforts were described in highlight form in the first section Summary of Research Program. In this section we discuss in outline form the major goals of our current grant period (5/1/80 - 4/30/83).

Our goals 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 (13CMR);

B) proton magnetic resonance (1HMR);

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-AIM computer resource;

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.

III.B. Justification and Requirements for Continued SUMEX use

In previous sections we discussed the relationship between the DENDRAL Project and SUMEX-AIM, methods for using SUMEX-AIM for dissemination of our programs to a broad community of structural chemists and biochemists and a critique of resource management. In this section we wish to emphasize certain factors which were not discussed earlier and to show how our future directions and interests are closely related to the proposed continuation and augmentation of the SUMEX-AIM resource.

As resource-related research, DENDRAL is intimately tied to the SUMEX resource. Our involvement with SUMEX goes far beyond simple use of the facility. We use SUMEX as the focal point for a number of collaborative efforts, for export of our software and for the communication facilities essential to maintaining close contact with remote research groups working with us. We have already discussed in our critique the difficulties we have, in view of heavy SUMEX load, of maintaining both our research effort and the resource-sharing aspects of our project.

In view of these factors and because SUMEX is our sole source of computational facilities, we took certain steps in our renewal proposal to attempt to alleviate our situation. Specifically, we requested a computer for our own project, a DEC VAX 11/780, to be linked to SUMEX via ETHERNET. This computer was meant to help offload some of the computational burden DENDRAL places on SUMEX, to provide a facility for production use of our programs by our collaborators and to represent a model for the type of low-cost, scientific computer available in the future to many investigators who could then run our programs in their own laboratories.

Our request for the VAX was turned down with specific comments made that SUMEX facilities should be used to support development of new programs and to the extent possible, encourage preliminary production use of our programs by outside persons. In our opinion this view is somewhat shortsighted, because SUMEX is currently overloaded to the extent that even development is impeded. In addition, our current situation leaves no room for the computational burden created by some of our collaborators who need considerably more than "preliminary" access because they have no access to a computer suitable for running our programs.

For this reason, we strongly support the effort of SUMEX to acquire a VAX and other small machines in future years, for all the reasons mentioned above. Although we realize that such machines will have to be shared among the SUMEX-AIM community as a whole, the augmentation of the resource would go a significant way to meeting the computational requirements of our project and provide a variety of systems of potential use for future export of our programs.

III.C. Needs and Plans for Other Computing Resources

For several years now we have directed some attention toward alternative computing resources which could be used to support all "production" use of our programs, i.e., all applications designed to use the programs to solve real problems. Although this would have the severe disadvantage of separating our research effort from many of the applications, it has been our hope that emerging technology in networking would enable us to keep in reasonably close contact with another resource. Two resources have emerged as candidates for systems where our programs can be accessed and used in problem-solving. Unfortunately, neither has so far proven feasible for several reasons (mentioned below). At this time we cannot determine if the problems will be resolved. Until such time, we will remain completely dependent on SUMEX for all our computational needs.

One alternative resource is the NIH/EPA Chemical Information System. For more than three years we have been working with them to obtain sufficient contract money to provide a version of CONGEN integrated into that system. The concept and the funds were approved but a contract has never been issued due to administrative problems at the EPA. Although there have been some developments recently, we still have no firm idea on when such a contract will be issued. If this effort is successful, then we can encourage persons who desire access to our programs to consider using the NIH/EPA system.

A second alternative is the National Resource for Computation in Chemistry (NRCC). Until recently, the computational facilities at the NRCC have not been suitable for running interactive programs. Recently, however, the NRCC has obtained a VAX system and we will investigate whether or not the community as a whole will have access to that system. The NRCC is currently under review for continued funding. Obviously that review will have to be favorable for the NRCC to represent an alternative for access to our programs.

III.D. Recommendations for Future Resource and Community Development

We have discussed previously our recommendation for the hardware augmentation, particularly with regards to purchase of small machines to facilitate future export. We also have increasing need for more file storage on-line. This is a result of building large data bases as part of our research in spectral interpretation. For the time being we are working with experimental programs and small data bases. As time progresses, however, these data bases will grow rapidly as our group and a number of our collaborators add additional structures and associated spectral data.

Another capability which is of increasing importance to our own work is access to low-cost graphics systems. Our programs will develop increasing dependence on graphics for visualization of three-dimensional molecular structures. Scientists desiring access to our programs will need a graphics terminal for optimum use of our systems. Currently available vector displays are simply too expensive for the average investigator. The emerging technology of low-cost raster display systems offers a more

promising possibility. However, no currently available machine has the required capabilities for under $10,000, and this is an area where machines like the Alto hold more promise. SUMEX could perhaps initiate an effort to obtain a system which has the hardware necessary for frame-based display. Such a system allows rotation of three-dimensional objects in a way which permits visualization of the actual shape of the object.

9.1.4 MOLGEN Project

MOLGEN - A Computer Science Application to Molecular Biology

Profs. E. Feigenbaum, L. Kedes, and D. Brutlag, Dr. P. Friedland
Department of Computer Science
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The MOLGEN project has focused on research into the applications of symbolic computation and inference to the field of molecular biology. This has taken the specific form of systems which provide assistance to the experimental scientist in various tasks, the most important of which have been the design of complex experiment plans and the analysis of nucleic acid sequences. We plan to expand and improve these systems and build new ones to meet the rapidly growing needs of the domain of recombinant DNA technology. We do this with the view of including the widest possible national user community through the facilities available on the SUMEX-AIM computer resource.

It is only within the last few years that the domain of molecular biology has needed automated methods for experimental assistance. The advent of rapid DNA cloning and sequencing methods has had an explosive effect on the amount of data that can be most readily represented and analyzed by computer. Moreover we have already reached a point where progress in the analysis of the information in DNA sequences is being limited by the combinatorics of the various types of analytical comparison methods available. The application of judicious rules for the detection of profitable directions of analysis and for pruning those which obviously lack merit will have an autocatalytic effect on this field in the immediate future.

The MOLGEN project has continuing computer science goals of exploring issues of knowledge representation, problem-solving, and planning within a real and complex domain. The project operates in a framework of collaboration between the Heuristic Programming Project (HPP) in the Computer Science Department and various domain experts in the departments of Biochemistry, Medicine, and Genetics. It draws from the experience of several other projects in the HPP which deal with applications of artificial intelligence to medicine, organic chemistry, and engineering.

During the next three years of MOLGEN research we intend to begin a transition from being primarily a computer science research project to being an interdisciplinary project with a strong applications focus. The tools that we have already developed will be improved to the point where they make a significant contribution to both research and engineering in the domain of molecular biology.

B. Medical relevance and collaboration

The field of molecular biology is nearing the point where the results of current research will have immediate and important application to the pharmaceutical and chemical industries. Recombinant DNA technology has already demonstrated the possibility of harnessing bacteria to produce nearly limitless amounts of such drugs as insulin and somatostatin. Several companies (Genentech, Cetus, Biogen) have already formed to exploit the commercial potential of the burgeoning technology.

The programs being developed in the MOLGEN project have already proven useful and important to a considerable number of molecular biologists. Currently several dozen researchers in various laboratories at Stanford (Prof. Paul Berg's, Prof. Stanley Cohen's, Prof. Laurence Kedes', Prof. Douglas Brutlag's, Prof. Henry Kaplan's, and Prof. Douglas Wallace's) and many others throughout the country (University of Utah, Syracuse University, NIH, Johns Hopkins, Yale, Rockefeller University, and others) are using MOLGEN programs over the SUMEX-AIM facility. We have exported some of our programs to users outside the range of our computer network (University of Geneva, for example).

C. Highlights of Research Progress

Accomplishments

The current year has seen the completion of what might be considered the first phase of the MOLGEN project. This section will summarize the major accomplishments of that first phase.

Representation Research

The domain of molecular biology has proven a fruitful testbed in the development of a flexible software package, the Unit System, for symbolic representation of knowledge. The package is already in use by a variety of research projects both within the Heuristic Programming Project at Stanford and at other institutions. It provides for acquisition and storage of many different types of knowledge, ranging from simple declarative types like integers and strings to complex declarative types like nucleic acid restriction maps to procedural types like a rule language in a subset of English.

Planning Research

The problem of designing laboratory experiments in molecular biology has been fundamental to MOLGEN research. The work has been split into two major subparts, each resulting in a doctoral thesis in computer science. The two systems, developed by Peter Friedland and Mark Stefik, produce reasonable experiment designs on test problems suggested by laboratory scientists.

Friedland's system is based on the observation that human scientists rarely plan experiments from scratch. They start with an abstracted or "skeletal" plan which contains the entire design in outline form. The

major design task is in instantiating or detailing the steps by finding tools that will work best in the given problem environment. This system has roots in classic problem-solving work dating back to Polya, and also in the Scripts language understanding work of Schank and Abelson. It is heavily dependent upon large amounts of domain specific knowledge, especially upon good heuristics for choosing among alternatives for plan-step instantiation.

Stefik's system emphasizes the role that interactions between steps in a plan should have when the plan is being designed. It uses an approach called "constraint posting" to make the interactions between subproblems explicit. Constraints are dynamically formulated and propagated during hierarchical planning and are used to coordinate the solution of nearly independent subproblems. The system also formalizes the problem of control during planning (what to do next) within a structure called "meta-planning". See Appendix B for an annotated example of the system at work.

Knowledge Base Construction

With the experiment design research as an impetus and the Unit System as a tool, a large knowledge base has been constructed by several Stanford molecular biologists--Prof. Douglas Brutlag, Prof. Laurence Kedes, Dr. John Sninsky, and Rosalind Grymes. This knowledge base is near-expert in several areas (enzymatic methods, nucleic acid structures, detection methods) and contains pointers and references to almost all areas of modern molecular biology. Its design and construction will soon be taken over by a full-time molecular biologist.

Besides its use as a fundamental part of an experiment design system, the knowledge base is proving useful for applications in teaching, in automated nucleic acid sequence analysis (see below), and as an intelligent "encyclopedia" for providing information about technique selection in the laboratory.

Other Applications of Symbolic Computation to Molecular Biology

Along with the central research in representation and planning, considerable work has been devoted to the construction of tools that are immediately useful to molecular biologists. Most of these tools were developed at the request of the various domain scientists working on the MOLGEN project and are being used by several dozen scientists both at Stanford and elsewhere through the facilities of the SUMEX computer system.

Interactive tools for nucleic acid sequence analysis--a multi-purpose program for analysis of primary sequence data has been made interactive with full help facilities. The program has also been improved to correctly calculate the expected probability of symmetries and homologies, and to properly allow for GU and GT bonding. A series of smaller programs for similar tasks has also been made interactive on the SUMEX system.

Sequence analysis through the knowledge base--some of the representational tools developed during the process of knowledge base construction (see above), have proven useful for computer-assisted sequence

analysis. Facilities are available for building and displaying restriction maps and region information, and for writing rules which cause this information to be automatically updated as new enzymes or structures are added to the knowledge base.

A program for restriction mapping,the GA1 program constructs restriction maps using data from total and partial restriction enzyme digests.

A program was written which aids in enzyme selection for gene excision. The SAFE program takes amino acid sequence data and predicts those restriction enzymes which are guaranteed not to cut within the gene.

A ligase simulation program was written. It is based on a kinetic theory of ligation which helps scientists select time of reaction and concentrations of reaction components to produce single inserts into vectors.

Research in Progress

The remainder of the current grant period will be spent on the further development of the tools that have been constructed for experiment design and sequence analysis and on expansion and improvement of the knowledge base. This section details those research plans.

Experiment Design

Both Friedland's and Stefik's experiment design system have already achieved modest success in producing reasonable plans for a variety of synthetic and analytic problems in molecular biology. Friedland's system can provide technically competent designs for about twenty different types of analytical problems. Stefik's system provides more innovative planning for a single type of synthetic experiment.

We intend to begin to integrate the two systems; Stefik's system will serve as a "front-end" that supplies the skeletal plans that drive Friedland's system. The combination of the two methods should provide a synergistic effect that facilitates both efficiency and innovation.

A second area of improvement in experiment design lies in providing the design systems with a deeper "theory of the domain." We would like design decisions to be made on the basis of mechanism whenever possible; e.g. to denature a molecule pick the best hydrogen bond-breaker, rather than the best pre-stored denaturation method. The current first step in making this improvement is in giving the representation formalism the power to work with sequence and topology of molecules, as described below.

An added benefit of the work on sequence and topology is in giving the planning system the ability to carry out certain steps of experiment designs. Many problems involve one or more steps that can be solved by use of the sequence analysis tools described in the previous section. The design system can make use of these tools directly and sometimes find faster and better solutions than can be achieved in the laboratory.

For example, the sub-problem of finding the right restriction enzymes to excise a gene for cloning can be solved by laborious experimental effort or by a few seconds of automated comparison of the gene with the cutting sites of all of the available restriction enzymes.

Knowledge Base Construction

The current knowledge base contains information about some three hundred laboratory methods and thirty strategies (skeletal plans) for using those methods. It also contains the best currently available data on about forty common phages, plasmids, genes, and other known nucleic acid structures.

We have recently concentrated on providing rules that allow the knowledge base to be automatically updated as new techniques or structures are added (for example, automatically revising restriction maps when a new restriction endonuclease is described). We are also working on mechanisms for facilitating the description of restriction sites and functional regions within molecules. After we are satisfied that our representation method is adequate, rules that model the changing structure of nucleic acid structures during the course of an experiment will be added to the knowledge base.

The knowledge base work to date has all been accomplished with the limited time of several expert molecular biologists, particularly Professors Douglas Brutlag and Laurence Kedes. We have just completed a search for an expert to carry on the knowledge base improvement full time and have hired Dr. Rene' Bach for this role. He will begin work on the MOLGEN project sometime early this summer.

Sequence Analysis

The sequence analysis methods described in the previous section have proven useful to a varied group of users throughout the country over the SUMEX-AIM facility. We will continue to improve these powerful tools and plan to make them available to the scientific community at large on the SUMEX-AIM national resource. If this test is successful, it will demonstrate the need for a full-scale national facility for sequence storage and analysis, and also the ability of MOLGEN to fill that need.

D. Publications

Feitelson J., Stefik M.J., A Case Study of the Reasoning in a Genetics Experiment, Heuristic Programming Project Report HPP-77-18 (Working Paper) (May 1977)

Friedland P., Knowledge-Based Experiment Design in Molecular Genetics, Proceedings Sixth International Joint Conference on Artificial Intelligence, 285-287 (August 1979)

Friedland P., Knowledge-Based Experiment Design in Molecular Genetics, Ph.D. Thesis, Stanford CS Report CS79-760 (December 1979)