Table of Contents

List of Figures

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).

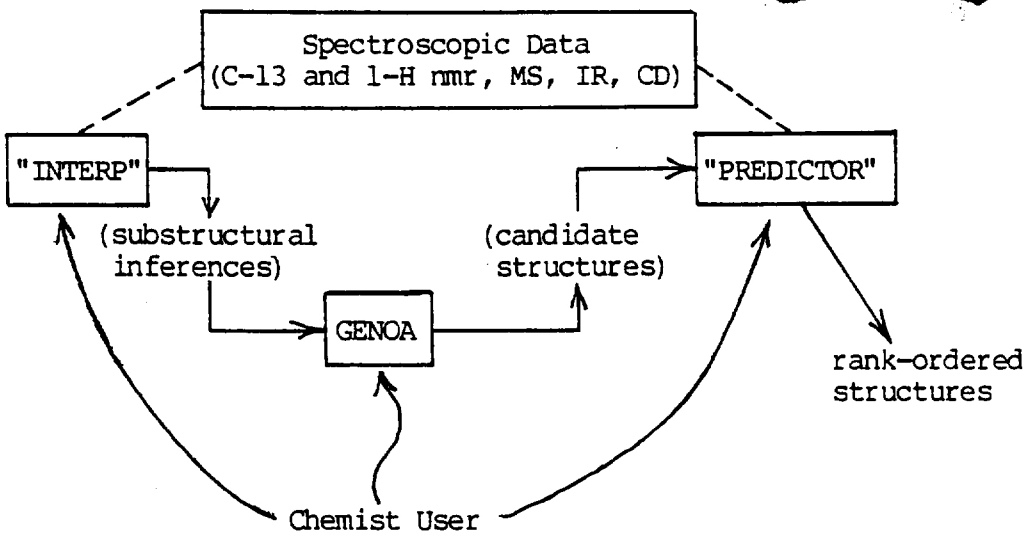


Figure 1. Block Diagram of the SASES system.

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.

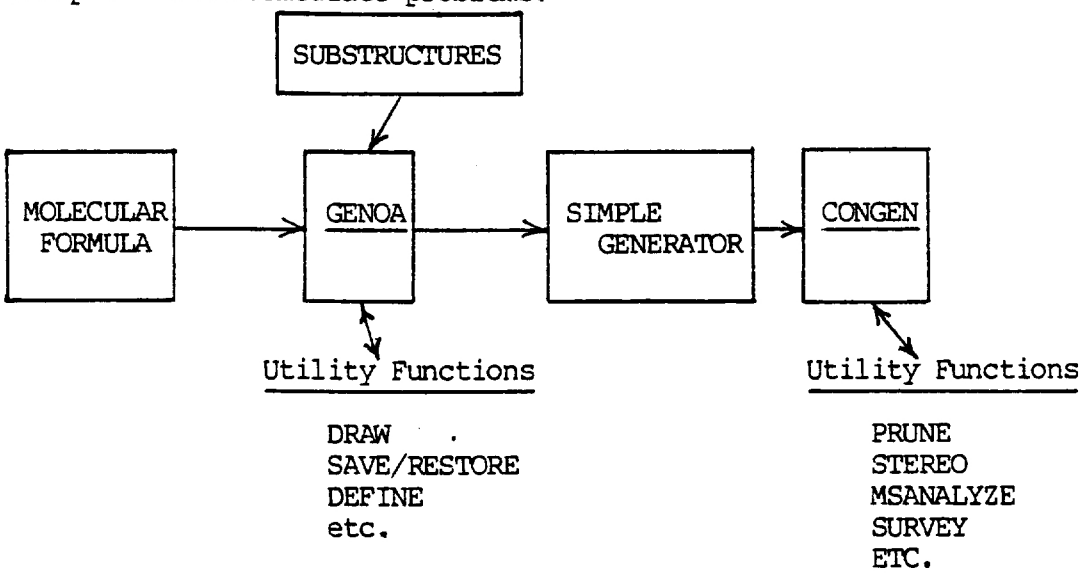


Figure 2. Current Status of GENOA and its Interface to the CONGEN Program.

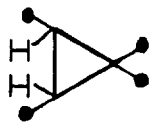
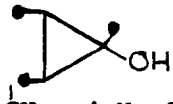
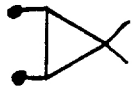
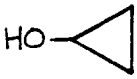
This example (Figure 3) is shown to illustrate several points about the use of GENOA in a structural problem. One can, as was done in this problem using CONGEN [17a], wait until many of the data are gathered, determine non-overlapping substructures and use constraints to test for those substructures 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.

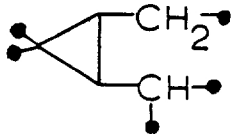
[redacted]

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) $(CH_3)_2$-CH-

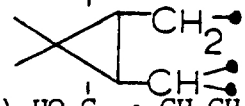
2 PROBLEMS, #1,#3, PREVIOUS STEP.

9. (1)

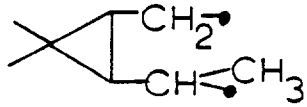


2 PROBLEMS;

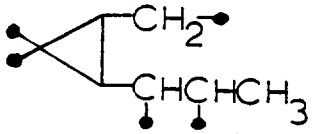
1) HO-C- ; 2 X CH_3CH-; 3 X-CH_2- ;



2) HO-C- ; CH_3CH-; -CH-; 3 X -CH_2- ;

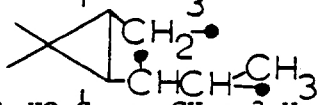


10. (1)

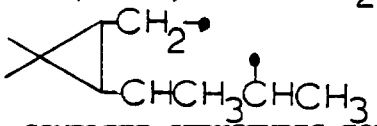


2 PROBLEMS:

1) HO-C- ; CH_3CH-; 3 X -CH_2- ;



2) HO-C- ; -CH-; 3 X -CH_2- ;



11. GENERATE STRUCTURES:

(Ø) C=C; (1) 

(Ø) C≡C

81 COMPLETE STRUCTURES FOR FURTHER EXAMINATION, FOR EXAMPLE:

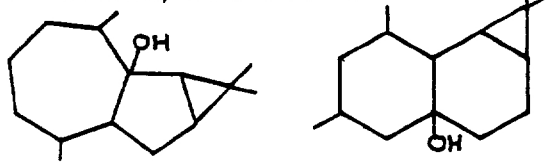


Figure 3. Illustration of the Use of GENOA in the Step-By-Step Specification of Overlapping Structural Information for the Structure of Palustrol [17A].

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