III. Research Plan

1 Introduction

1.1 Objectives

Our principal objectives are

A. Developing an integrated approach to computer-assisted elucidation of biomolecular structures; and

B. Applying the techniques of computer-assisted structure elucidation to a wide range of biomolecular structural problems.

To those ends, we will endeavor:

To extend the current DENDRAL programs and write new programs

To explore other aspects of the structure elucidation problem.

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

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

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

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

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

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

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

1.2 Background and Rationale

1.2.1 The Structure Elucidation Problem

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

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

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

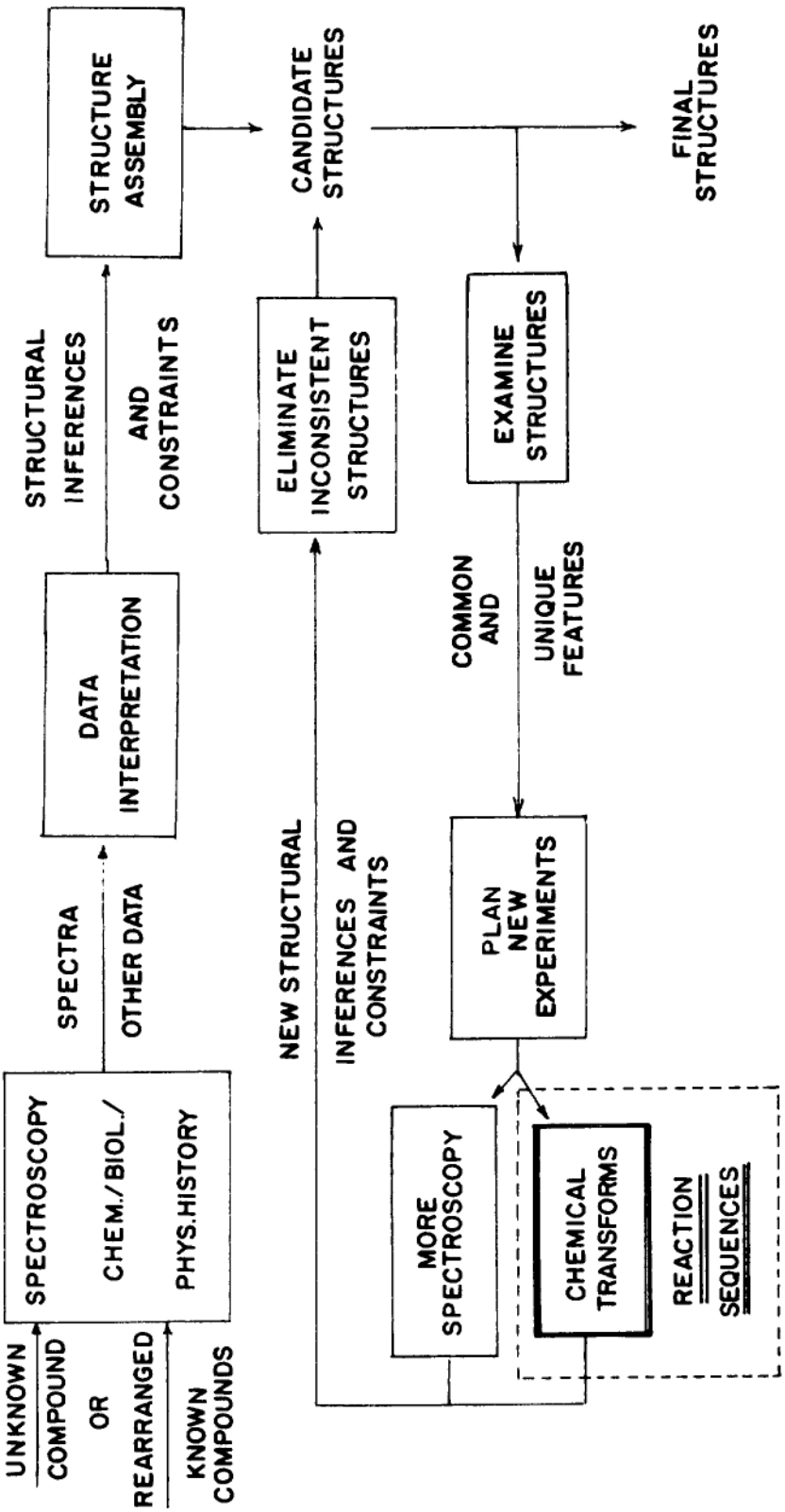

UNKNOWN COMPOUND
OR
REARRANGED
KNOWN COMPOUNDS
SPECTROSCOPY
CHEM./BIOL./
PHYS.HISTORY
SPECTRA
OTHER DATA
DATA INTERPRETATION
STRUCTURAL INFERENCES
AND CONSTRAINTS
STRUCTURE ASSEMBLY
CANDIDATE STRUCTURES
ELIMINATE INCONSISTENT STRUCTURES
NEW STRUCTURAL
INFERENCES AND CONSTRAINTS
EXAMINE STRUCTURES
COMMON AND
UNIQUE FEATURES
FINAL STRUCTURES
PLAN NEW EXPERIMENTS
MORE SPECTROSCOPY
CHEMICAL TRANSFORMS
REACTION SEQUENCES

Figure 1

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

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

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

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

1.2.2 Historical Background

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

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

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

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

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

1.3 Existing Capabilities

1.3.1 CONGEN

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

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

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

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

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

1.3.2 Meta-DENDRAL

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

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

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

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

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

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

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

1.3.3 Gas Chromatography/High Resolution Mass Spectrometry Resource

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

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

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

1.3.4 Related Computer Programs

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

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

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

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

1.3.5 Collaborative Research Environment

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

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

1.4 Relationship to Mass Spectrometry and AIM-SUMEX Resources

1.4.1 Mass Spectrometry Resource

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

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

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

1.4.2 AIM-SUMEX Resource

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

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

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

2 Specific Aims

2.1 Add More "Intelligence" to Existing Programs

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

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

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

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

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

2.2 Develop New Computer Programs that Assist in Biomolecular Structure Elucidation

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

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

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

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

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

2.3 Develop New Programs that Aid in Rule Formation

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

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

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

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

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

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

2.5 Increase the Availability of the Structure Elucidation Techniques

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

2.6 Maintain and Improve the GC/HRMS System

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

3 Methods

3.1 Extra Intelligence in Existing Programs

3.1.1 Constraints Interpreter for CONGEN

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

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

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

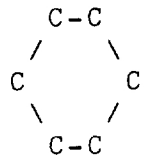
Input Translation. An input translator will determine the implications of the input data and find a new internal definition

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

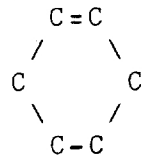
In other cases, substructures appear on GOODLIST either because they are too imprecise to be superatoms (i.e., they may contain atoms or bonds multiplicities which can take on a range of values) or because they may overlap other superatoms (superatoms are required by CONGEN to be atom-disjoint fragments).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

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

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



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



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

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

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

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

3.1.2 Intelligent Help System

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

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

We have solved interface problems until now in a piecemeal

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

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

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

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

3.2 New Programs for Structure Elucidation

3.2.1 Experiment Planning Program

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

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

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

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

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

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

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

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

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

Our proposals for this organization are as follows. Considering the substructural organization of the knowledge base (which provides the keys which can be searched for in intercomparison of structures) we assume a hierarchy of structural descriptors, from broad descriptions to specific items. Broad descriptors include one category for functional groups, one category for proton distributions (e.g., from 1H NMR data), one for carbon distributions (e.g., from C13 NMR data), one for ring size and type distributions, and so forth. Each category will be further subdivided as appropriate. For example, the functional group category can be subdivided according to heteroatom, local functional environments for each heteroatom, and "extended" environments which include the functionality and more remote structural features. As each descriptor becomes more specific and an experiment exists which can provide some information about the descriptor, the experiment will be included as part of the information associated with the substructure. Associated with each experiment will also be qualifiers on sample requirements, interfering functionalities, and preconditions for the experiment (e.g., solubility, etc.). Of course, the experiments will become more specific also. For example, an initial suggestion for an experiment might be to obtain a 1H NMR spectrum if one has not been obtained. The next suggestions would depend on how the structures differed in those characteristics which are normally easy to determine from a 1H NMR spectrum, e.g., number of methyl groups, vinyl and aromatic protons, etc. At the most detailed level, specific proton decoupling experiments would be proposed if the candidate structures differed in appropriate ways.

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

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

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

3.2.2 Reaction Chemistry Program

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

3.2.2.1 Use of REACT in Structure Elucidation

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

a) test for a specific functional group;

b) simplify the problem by decomposing the unknown into smaller, more easily characterized molecules;

c) modify the skeleton or functional groups to define more accurately their respective environments or make the unknown more amenable to analysis (e.g., increase its volatility); or

d) unambiguously relate the unknown to a previously characterized compound.

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

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

substantially increase its usefulness in real-world chemical problems.

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

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

3.2.2.2 Importance of REACT for Relatiing Graph Theory to Chemistry

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

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

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

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

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

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

C1=C2 ---> C1-C2 (a)

C1=C2 ---> C1-C2-OH (b)

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

CH2=CH-CH3

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

CH3-CH2-CH3

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

HO-CH2-CH2-CH3 and CH3-CH-CH3
|
OH

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

3.2.3 General Mass Spectrum Analysis Program for Unknowns

PLANNER, which is currently the only program we have for