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

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

2.1.1.3 Export status

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

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

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

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

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

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

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

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

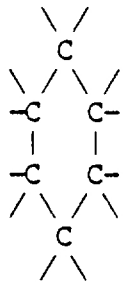
Several other computer systems and languages have been explored for suitability for CONGEN. So far all have suffered from language deficiencies which do not allow the heavy recursion required for CONGEN's basic algorithms (e.g. FORTRAN), from lack of transportability (the BLISS and C languages), or from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

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

2.2 CONGEN Developments

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

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



A

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

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

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

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

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

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

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

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

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

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

Step-by-step input;

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

Condensed input;

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

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

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

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

2.3 RESOURCE SHARING

2.3.1 CONGEN Workshops

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

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

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

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

2.3.2 Conclusions from the Workshop

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

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

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

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

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

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

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

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

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

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

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

2.4 Stereochemistry

2.4.1 SAIL Program

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

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

2.4.2 BCPL program

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

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

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

2.4.3 European speaking trip

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

2.5 Structure checking functions for CONGEN

2.5.1 Introduction

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

The main structure checking functions currently available through STRUCC are:

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

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

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

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

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

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

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

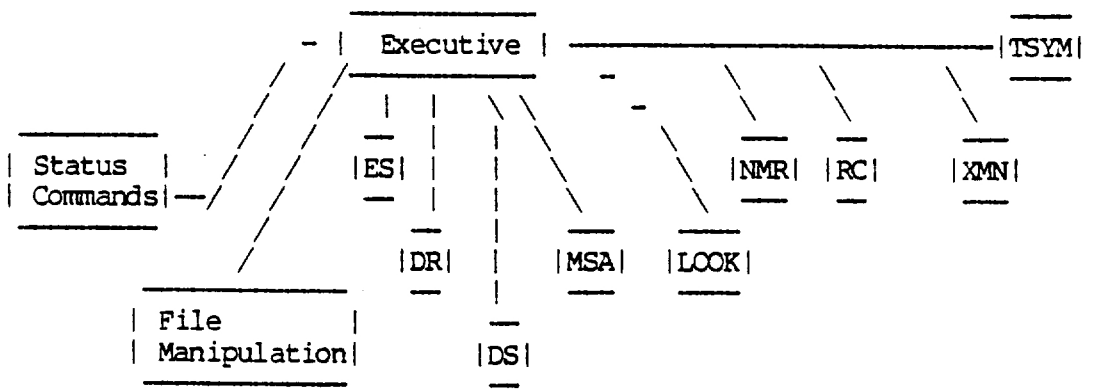
6) NMRFLT: The NMRFLT functions represent a first attempt at developing a system for predicting proton resonance spectra of candidate structures, and for using differences between predicted and observed spectra as a basis for pruning the structure list.

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

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

2.5.2 The Form of the STRUCC Program

The following diagram indicates schematically the overall form of STRUCC:



1) Status commands:

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

ii CLEAR: Restarts the program.

iii CM?: Gives the structure composition.

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

v CT?: Gives the current number of structures.

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

vii EXIT: Ends the program.

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

2) File Manipulation:

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

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

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

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

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

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

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

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

6) MSA: Mass spectral analysis.

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

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

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

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

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

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

2.5.3 STRUCC's HELP System

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

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

2.5.4 DEFINE-STRUCTURES

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

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

2.5.5 MSA, The Mass Spectral Analysis Functions.

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

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

i Parameters controlling the fragmentation process

ii A minimum plausibility value for ions to be listed

iii The minimum mass of interest.

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

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

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

i) The index numbers of the structures to be processed

ii) The observed spectrum

iii) Fragmentation control parameters.

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

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

i) The observed spectrum.

ii) Fragmentation control parameters.

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

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

b) The ion's mass

c) The ion's observed intensity

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

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

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

4) EXAMINE: In this mode, the program identifies all structures for which the observed ions are predicted. The information is converted into a form that can be used by EXAMINE. The observed ions can then be used as EXAMINE-selection keys, just like substructural features; so, one can select structures with
>C8H6Ol AND C6H1ØN1O3 AND C4H8N1O2
The number of ions that can be rationalized in terms of a given structure is used as a score for that structure. This score is available in EXAMINE. So, as well as checking for structures that can explain particular ions, it is possible to request those which can explain a given number of the observed ions. In EXAMINE mode, MSA requires the same data as when in RANKING mode.

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

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

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

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

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

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

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

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

8) if H-transfers or neutral losses are specified

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

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

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

2.5.6 LOOK

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

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

2.5.7 The Proton NMR Functions

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

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

The formulae used are:

$$Delta_{CH2} = Ø.2 + C1 + C2$$

$$Delta_{CH} = Ø.2 + C1 + C2 + C3$$

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

where the Ci and Zi values are supposedly additive constants.

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

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

2.5.8 BCPL versions of STRUCC

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

2.6 Meta-DENDRAL

2.6.1 META-DENDRAL PROGRESS

2.6.1.1 INTSUM

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

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

2.6.1.2 RULEGEN

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

2.6.1.3 EDITSTRUC INTERFACE

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

2.6.1.4 PREDICTION and RANKING

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

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

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

2.6.1.5 HUMAN ENGINEERING

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

2.6.2 RESULTS

The practical value and capability of new programs are best evaluated by applying them to real, non-trivial problems. In our case, we have chosen the biologically important marine sterol compounds. Their mass spectra are predominant in the structure elucidation of new compounds in spite of the fact that relatively few of the fragmentation mechanisms are known. Often very similar spectra are recorded due to the great similarity of common skeletons. Our study involves the comparison of predicted spectra of known structures with the observed spectra of unknown compounds. We want to compare the usefulness of different methods of forming the rules used for spectrum prediction. We distinguish 3 methods: 1) Half-order theory (can be supplemented by functional group rules). 2) Class-specific rules (selected by the chemist) 3) Computer-generated rules Our results were obtained using nine selected 4-demethylsterols (six isomers of composition C29H48O, two C28H46O and one C27H44O). Each spectrum of the nine selected marine sterols was considered to be the observed spectrum and ranked against 23 candidate structures (the 23 candidates contained 17 different C7 - C11 sidechains and three 4-demethylsterol skeletons). For the half-order theory an overall average performance of (2.4 Ø.9) was obtained. The first number gives the number of candidates ranked better than the correct one, the second represents the number of candidates ranked equally with the correct one. In this case the average value is not very representative, as its value is strongly reduced by a compound which was ranked in 17th place. This compound, the 23-demethylgorgosterol, contains a cyclopropane in the sidechain for which no special fragmentation processes are considered in the simple half-

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

2.7 REACT and MAXSUB Programs

2.7.1 REACT

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

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

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

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

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

2.7.2 MAXSUB Program

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

2.8 High Resolution GC/MS System

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

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

1) High Resolution analyses:

a) DENDRAL-related	134
b) Outside collaborators	45

2) High Resolution GC/MS

a) DENDRAL-related	86
b) Outside collaborators	13

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

2.9 References

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

(1) T.H. Varkony, R.E. Carhart, and D.H. Smith, "Computer-Assisted Structure Elucidation. Modelling Chemical Reaction Sequences Used in Molecular Structure Problems," in "Computer-Assisted Organic Synthesis," W.T. Wipke, Ed., American Chemical Society, Washington, D.C., 1977, p. 188.

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

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

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

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