

1976 - 1977
ANNUAL REPORT

RESOURCE-RELATED RESEARCH
COMPUTERS AND CHEMISTRY

Grant No. RR-00612

BIOTECHNOLOGY RESOURCES PROGRAM
OF THE
NATIONAL INSTITUTES OF HEALTH

February, 1977

COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY

SECTION 1

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE PUBLIC HEALTH SERVICE APPLICATION FOR CONTINUATION GRANT	REVIEW GROUP	TYPE	PROGRAM	GRANT NUMBER (insert on all pages)
	SSS	5	R24	RR-00612
	TOTAL PROJECT PERIOD			
	From: 5/1/74		Through: 4/30/77	
REQUESTED BUDGET PERIOD				
From: 8/1/76		Through: 4/30/77		

1. TITLE
Resource-Related Research Computers and Chemistry

2A. PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR
(Name and Address, Street, City, State, Zip Code)
DJERASSI, CARL
STANFORD UNIVERSITY
DEPT. OF CHEMISTRY
STANFORD, CALIF. 94305

4. APPLICANT ORGANIZATION (Name and Address-Street, City, State, Zip Code)

STANFORD UNIVERSITY

2B. DEGREE
PH.D.

2C. SOCIAL SECURITY NO.

5. PHS ACCOUNT NUMBER
1941156365A1

2D. DEPARTMENT, SERVICE, LABORATORY OR EQUIPMENT
CHEMISTRY

6. TITLE AND ADDRESS OF OFFICIAL IN BUSINESS OFFICE OF APPLICANT ORGANIZATION

DEPUTY V. P. FOR BUSINESS & FINANCE
STANFORD UNIVERSITY
STANFORD, CALIF. 94305

2E. MAJOR SUBDIVISION
SCH. OF HUMANITIES AND SCIENCES

3. ORGANIZATIONAL COMPONENT TO RECEIVE CREDIT FOR INSTITUTIONAL GRANT PURPOSES
20 OTHER

7. RESEARCH INVOLVING HUMAN SUBJECTS (See Instructions)

No Yes APPROVED

Date

8. INVENTIONS (See Instructions)

No

Yes-not previously reported

Yes previously reported

9. PERFORMANCE SITE(S)

STANFORD UNIVERSITY
DEPARTMENT OF CHEMISTRY
COMPUTER SCIENCE DEPARTMENT
DEPARTMTNE OF GENETICS
STANFORD, CALIF. 94305

TELEPHONE INFORMATION

11A PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR (Item 2a)	Area code	Tele. No. & Ext.
	415	497-2783
11B Name of business official (Item 6)		
K. D. Creighton	415	497-2251
11C Name and title of administrative official (Item 15b)		
D'Ann B. Downey, Sponsored Projects Dir.	415	497-2883

10. DIRECT COSTS REQUESTED FOR BUDGET PERIOD

\$129,931.00

12A. CONGRESSIONAL DISTRICT OF APPLICANT ORGANIZATION SHOWN IN ITEM 4

TWELFTH

12B. COUNTY OF APPLICANT ORGANIZATION SHOWN IN ITEM 4

SANTA CLARA

13. DO NOT USE THIS SPACE

14. CERTIFICATION AND ACCEPTANCE. We, the undersigned, certify that the statements herein are true and complete to the best of our knowledge and accept, as to any grant awarded, the obligation to comply with Public Health Service terms and conditions in effect at the time of the award.

15. SIGNATURES

(Signatures required on original copy only. Use ink. "Per" signatures not acceptable.)

15A. PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR

DATE

15B. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION

DATE

SECTION II

SECTION II—BUDGET (USUALLY 12 MONTHS)

FROM
May, 1977

THROUGH
April, 1978

GRANT NUMBER
RR-00612

A. ITEMIZE DIRECT COSTS REQUESTED FOR NEXT BUDGET PERIOD

PERSONNEL		TIME OR EFFORT %/HRS (c)	SALARY REQUESTED (d)	FRINGE BENEFITS (See Instructions) (e)	TOTAL (f)
NAME (Last, First, Initial) (a)	TITLE OF POSITION (b)				
	PRINCIPAL INVESTIGATOR				
See Proposal for Continued Research 1977-80 Submitted May, 1976					
Subtotals			\$	\$	
(Indicate cost of each item listed below)					\$
TOTAL (Columns (d) and (e))					\$
CONSULTANT COSTS (See Instructions)					\$
EQUIPMENT					\$
SUPPLIES					\$
TRAVEL	DOMESTIC				\$
	FOREIGN				\$
PATIENT COSTS (See instructions)					\$
ALTERATIONS AND RENOVATIONS					\$
OTHER EXPENSES (Itemize)					\$
TOTAL DIRECT COST (Enter on Page 1, Item 10)					\$

INDIRECT COST (See Instructions)

% S&W*
% TDC*
*If this is a special rate (e.g. off-site), explain.

Date of DHEW Agreement:

() Not Requested
() Under negotiation with

SECTION II—BUDGET (Continued)

Grant Number

RR-00612

B. Supplemental information regarding ITEMS in the proposed budget for the next period which require explanation or justification. (See instructions)

See Proposal for Continued Research

1977-80

submitted May, 1976

SECTION III

SECTION III—FISCAL DATA FOR CURRENT BUDGET PERIOD (USUALLY 12 MONTHS)	FROM	THROUGH	GRANT NUMBER
	8/1/76	4/30/77	RR-00612

The following pertains to your CURRENT PHS budget. Do not include cost sharing funds. This information in conjunction with that provided on Page 2 will be used in determining the amount of support for the NEXT budget period.

A. BUDGET CATEGORIES		CURRENT BUDGET (As approved by awarding unit) (1)	ACTUAL EXPENDITURES THRU 1/31/77 (Insert Date) (2)	ESTIMATED ADDITIONAL EXPENDITURES AND OBLIGATIONS FOR REMAINDER OF CURRENT BUDGET PERIOD (3)	TOTAL ESTIMATED EXPENDITURES AND OBLIGATIONS (Col. 2 plus Col. 3) (4)	ESTIMATED UNOBLIGATED BALANCE (Subtract Col. 4 from Col. 1) (5)
Personnel (Salaries)		100,761	59,784	36,593	96,377	4,384
Fringe Benefits		18,943	11,144	6,953	18,097	846
Consultant Costs		-	-	-	-	-
Equipment		-	-	-	-	-
Supplies		-	50	10	60	(60)
TRAVEL	Domestic	1,400	1,509	1,741	3,250	(1,850)
	Foreign					
Patient Costs						
Alterations and Renovations						
*Other		8,827	7,355	4,585	11,940	(3,113)
Total Direct Costs		129,931	79,842	49,882	129,724	207
Indirect Costs (If included in award)						
TOTALS →		\$ 129,931	\$ 79,842	\$ 49,882	\$ 129,724	\$ 207

Use space below to:

- B. List all items of equipment purchased or expected to be purchased during this budget period which have a unit cost of \$1000 or more.
 C. Explain any significant balance or deficit shown in any category of Column 5.
 D. List all other research support for Principal Investigator by source, project title, and annual amount.

B. Equipment Purchased

Terminal Rental	3,587	3,957	1,800		
Telephone	2,100	2,340	1,080		
Publications and Technical services	2,900	984	1,600		
Postage	240	74	105	11,940	(3,113)

D. Research Support for Principal Investigator

NIH AM04257	Mass Spectrometry in Organic Biochemistry	\$ 67,700
NIH GM06840	Marine chemistry with special Emphasis on Steroids	101,490
NIH GM20276	Applications of Magnetic Circular Dichroism	40,834

SECTION IV

APPLICANT: REPEAT GRANT NUMBER SHOWN ON PAGE 1 →		GRANT NUMBER	
SECTION IV—SUMMARY PROGRESS REPORT		RR-00612	
PRINCIPAL INVESTIGATOR OR PROGRAM DIRECTOR (Last, First, Initial)		PERIOD COVERED BY THIS REPORT	
NAME OF ORGANIZATION		FROM	THROUGH
		8/1/76	4/30/77

TITLE (Repeat title shown in Item 1 on first page)

1. List publications: (a) published and not previously reported; (b) in press. Provide five reprints if not previously submitted.
2. List all additions and deletions in professional personnel and any changes in effort.
3. Progress Report. (See Instructions)

Progress Report Follows

Resource Related Research - Computers & Chemistry

Stanford University
NIH/BRP Grant RR-00612

Carl Djerassi, Principal Investigator
(Social Security No.)

Research Highlights (1976-77)

1. One of the DENDRAL programs, named CONGEN, continued to be used by working scientists on problems of biomedical importance. CONGEN is a complex symbol-manipulation program that helps working scientists develop hypotheses about the molecular structure of organic chemical compounds. The scientist provides structural inferences, the program provides the result of combining the scientist's partial hypotheses in all plausible ways. Several scientists across the nation have used the CONGEN program to help them elucidate molecular structures of naturally-occurring compounds. We anticipate continued use of this program, and similar ones, as scientists realize how helpful they can be for structure elucidation.

2. The Meta-DENDRAL program successfully formulated rules of mass spectrometry that were new to the science. These rules, along with a discussion of the methodology, were published in the scientific literature [Jnl Am Chem Soc, 98:6168 (1976)]. The

program was tested to see if it could rediscover the rules of mass spectrometry for two classes of chemical compounds that were already well understood (amines and estrogenic steroids). Then it was applied to three classes of compounds whose behavior in a mass spectrometer was not as well known (mono-, di-, and tri-ketoandrostanes). The program produced three sets of rules that explained much of the significant data for these classes. The time for manual rule formation for these data was estimated to be several months.

Table of Contents

Section	Page
Subsection	
1. OVERVIEW OF RESEARCH ACTIVITIES	1
2. EXPERIMENT PLANNER	5
3. Applications of REACT to Structure Elucidation Problems	9
4. CONGEN Developments	15
4.1 Intelligent use of constraining substructural information	15
4.2 New tools for post-pruning CONGEN structures.	19
5. USE OF CONGEN BY OTHER SCIENTISTS	24
5.1 Chemists Using Exported Programs	24
5.2 Remote Users of SUMEX	26
5.3 Chemists Communicating by Mail	26
5.4 Chemical Problems Posed to CONGEN	26
6. Stereochemistry in CONGEN	30
6.1 Algorithm	30
6.2 Programming Progress	31
7. The GC/HRMS DATA SYSTEM	32
7.1 Improvements to the Data System	32

7.2	Changes in the Operating System	34
7.3	New Developments	35
8.	META DENDRAL	36
9.	C13 NMR SPECTROMETRY	38
10.	BUDGET	40
11.	RECENT PUBLICATIONS OF THE HEURISTIC PROGRAMMING PROJECT	41

Resource Related Research - Computers & Chemistry

ANNUAL REPORT
August 1, 1976 - April 30, 1977

Stanford University
NIH/BRP Grant RR-00612

Carl Djerassi, Principal Investigator
(Social Security No. [REDACTED])

1 OVERVIEW OF RESEARCH ACTIVITIES

The past year's activities in computer applications to chemical problems have continued the progression of new research, followed by applications and export to a wider community of scientists. The simplest way to detail this work is to place it within the framework of the larger problem of elucidation of unknown molecular structures. Our research, development and future plans focus on both the question of structure elucidation in general and the problem of providing computer assistance to scientists engaged in specific aspects of this important activity.

A simplified representation of major milestones in solving unknown biomolecular structures by manual methods is presented in Figure 1.

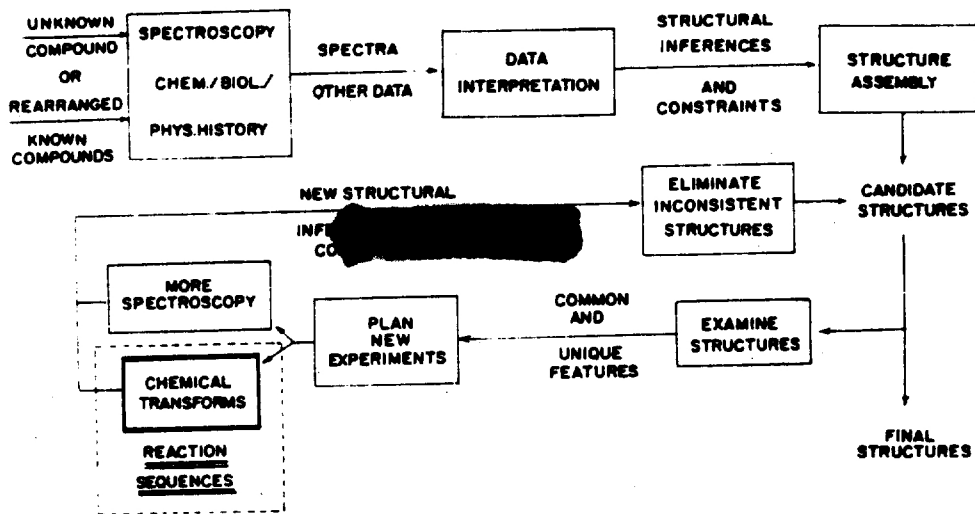


Figure 1. Important steps in manual solution of structures of unknown chemical compounds.

These steps, indicated as separate boxes, may be performed explicitly or implicitly. There are considerably more complex relationships among the boxes of Fig. 1 than are indicated when structures are actually solved. Nevertheless, the Figure provides a good introduction to both our recent work and our future directions. We describe briefly each of the milestones in the following paragraphs. More detailed discussions of each topic follow in subsequent sections.

The first step in identification of an unknown structure is to separate it from other components in a potentially complex mixture and to isolate it in reasonably pure form. These steps are performed by scientists, frequently with the assistance of various instruments. Although our research is not directed toward any part of this separation and isolation procedure (except insofar as these procedures also yield data which are subject to computer-assisted interpretation), information about the chemical and physical characteristics of the compound may be crucial to further efforts to determine its structure.

Depending on the quantity of sample available and its characteristics, various spectroscopic and additional chemical data are then collected on the unknown. A mass spectrum is frequently obtained, e.g., from a combined gas chromatograph/mass spectrometer (GC/MS) system. An important part of our recent proposal to the NIH is directed toward automation of combined GC/MS systems operated at high mass spectrometer resolving powers. Data on elemental compositions and relative

ion abundances are then available in computer-readable form for further analysis (see MSRANK). The chemist possess an armamentarium of spectroscopic techniques which can be brought to bear on a structure. One advantage of our work is that any data so obtained can be used to help solve the structure as long as it can be expressed, manually or by computer, in substructural statements about the unknown.

The next important phase in structure elucidation is interpretation of the available data (Fig. 1) in terms of structural features of the molecule. These interpretations may be in terms of known structural units ("superatoms", polyatomic aggregates of atoms in known configurations), or in terms of structural units, ring sizes, proton or carbon distributions. The latter set of features represents constraints on the kinds of structures which are possible. Our efforts in the area of computer-assisted data interpretation are focussed on mass spectral and carbon-13 nuclear magnetic resonance (¹³CMR) data. We are developing general approaches to automated analysis of these data in terms of structural features of unknowns.

Our recent efforts are summarized in Figure 2, and discussed in detail subsequently. We have been concerned with use of these data from two points of view, planning and prediction (Fig. 2). During planning, experimental data are examined in order to extract specific structural information to be used in assembling candidate structures. In prediction each candidate structure is tested to determine how closely its predicted spectrum agrees with the observed spectrum. The candidates can be ranked accordingly. The Meta-DENDRAL research is directed toward determination of rules of spectroscopic data which can be used either for planning or prediction (see below).

DATA INTERPRETATION"PLANNING"

EXTRACTION OF STRUCTURAL
INFORMATION DIRECTLY FROM
SPECTROSCOPIC DATA.

1. MASS SPECTRA - MDGGEN
2. ¹³CNMR

PREDICTION

USE OF SPECTROSCOPIC
DATA TO RANK
CANDIDATE STRUCTURES.

1. MSPRUNE, MSPRED
2. ¹³CNMR



FORMATION OF RULES TO BE
USED FOR BOTH PLANNING
AND PREDICTION.

Figure 2. Relationship between use of rules in either planning or prediction. Both approaches are used in utilizing data for structure elucidation.

Given possible structural fragments of the complete molecule and constraints on how these fragments may be assembled into complete molecules, a process of structural assembly follows (Fig. 1). There has been no proven algorithm for solving this problem prior to earlier work supported by the current grant. Traditionally, this process has been left to manual, pencil and paper work. Our CONGEN program, which was designed to solve this problem, is farthest advanced of programs designed to assist in various aspects of structure elucidation. It performs the structural assembly process, under constraints, and allows the scientist using the program to examine structural candidates and remove those deemed implausible (Fig. 1). A large portion of our recent and future work is directed toward improving the CONGEN program and building other facilities around it (see later sections). We have demonstrated the utility of CONGEN in structural studies, and subsequent sections discuss our recent developments and applications of CONGEN as well as our interactions with other scientists desiring access to our programs.

Given a set of structural candidates, the experimenter examines them to determine what experiments might be performed to focus on the correct structure by stepwise rejection of alternative hypotheses. When there are only a small number of possibilities under consideration, manual methods suffice. But CONGEN provides the capability for exhaustive enumeration of structural possibilities at a point in a structural problem when there may be many hundreds of possibilities. It is very difficult to examine these structures and plan experiments by hand. We have begun exploring ways to provide computer assistance to this important aspect of structure elucidation. We refer to this research area as the Experiment Planner, discussed in more detail below.

When new experiments have been planned the researcher carries them out and uses the results as additional constraints on the structural candidates (Fig. 1). New experiments may include collecting of additional spectroscopic data or performing a sequence of chemical reactions on the unknown. The latter experiments may be chosen to convert the unknown into a related compound which possesses physical or chemical properties more amenable to analysis. During the past year we have developed a program to assist scientists in carrying out representations of chemical reactions in the computer and eliminating undesired structural candidates based on constraints exercised on the products of the reaction. This work is described in two subsequent sections. One section describes use of the program, which we call REACT, to explore structural possibilities exactly as outlined above. A later section describes recent progress in increasing the power of REACT.

2 EXPERIMENT PLANNER

We have begun preliminary considerations of design and implementation of an experiment planner. This program will assist chemists in designing the most effective set of experiments to perform to solve the structure. Although the experiment planner will be a future activity of our group, we are developing and using other structure manipulation functions which will provide groundwork for future developments.

One important aspect of experiment planning is the ability to examine in some way the set of candidate structures. Although

many can be drawn for visual review, drawing is impractical when dozens or hundreds of structures are involved. To assist persons using CONGEN in reviewing their structures we have developed a function auxiliary to CONGEN which we call SURVEY.

SURVEY

FUNCTION: AID IN PERCEPTION OF ANY OF A
PRE-SPECIFIED SET OF STRUCTURAL
FEATURES IN A GROUP OF
STRUCTURAL CANDIDATES.

- E.G. A) FUNCTIONAL GROUPS
B) TERPENOID SKELETONS
C) AMINO ACID SKELETONS

Figure 3. Function of the SURVEY program and examples of recent application areas.

The function of SURVEY is summarized in Figure 3. SURVEY simply acts as a reminder to the scientist of the presence or absence of certain structures or structural features. During the past year we have used SURVEY extensively. For example, we have used it to detect implausible functional groups in a set of candidate structures, using a file of substructures representing a wide variety of functionalities. In many problems, implausible functional groups are forgotten and CONGEN is never constrained to remove them. Another example of use of SURVEY is in conjunction with collaborative work with persons in the Department of Genetics. In analysis of serum or urinary metabolites in patients of high risk of metabolic disorder, we have had occasion to use CONGEN in exploration of unknown structures [Report HPP-77-11]. Some of these structures could formally be conjugates of amino acids with organic acids. If so, such structures will possess backbones of naturally-occurring amino acids. SURVEY was used to provide a summary of which structural candidates possessed such amino acid skeletons.

We have recently used SURVEY in a related application involving the structure of "polyalthenol", discussed by LeBoeuf,

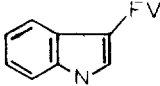
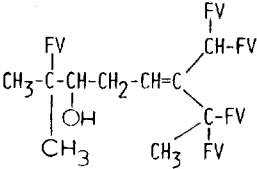

et al. (Figure 4). Superatoms and constraints supplied to CONGEN to derive structural candidates are summarized in Fig. 4.

We summarize in Figure 5 the structural possibilities which resulted. There are five structures possessing a bicyclo[2.1.1] system, and six which possess a bicyclo[4.3.1] system (Fig. 5, top). These structures are energetically less favorable. For example, several possess a double bond at a bridgehead atom, which violates Bredt's Rule. There remain, however, 11 structures which are not formally excluded by data presented by LeBoeuf, et al. Because these workers based their structural assignment on biogenetic grounds, we used SURVEY and REACT to test their hypothesis. We have, in computer-accessible libraries, known terpenoid ring systems which can be used within SURVEY to test sets of structures for known skeletons. None of the 22 structural candidates possesses a previously known skeleton. Because the authors postulated a relationship to a known skeleton via a single methyl shift, we used REACT to exercise a single methyl shift in all possible ways on each of the 22 candidates. SURVEY was then used to test the results for the presence of known terpenoid systems, and the drimane skeleton, the postulated precursor of polyathenol, was the only known skeleton which resulted. This does not prove the hypothesis of LeBoeuf, et al., but certainly helps strengthen it.

SURVEY is, however, only the barest beginning of an experiment planner, even though it has proven useful. We plan to build from this beginning toward a much more powerful system.

M. LeBOEUF, M. HAMONNIÈRE, A. CAVÉ, H. GOTTLIEB, N. KUNESCH,
AND E. WENKERT, TET. LETT., 3559 (1976).

"POLYALTHENOL" $C_{23}H_{31}NO$

SUPERATOMS	ARBITRARY NAME	NUMBER
	IN	1
	BI	1
CH ₃ -FV	ME	1
FV-CH ₂ -FV	CH2	3
	CH	1

CONSTRAINTS

- 1) ALL FREE VALENCES BONDED TO NON-HYDROGEN ATOMS
- 2) GOODLIST

(EVENTUALLY	IN-CH ₂ -BI	1 TO ANY
	IN-CH ₂ -CH ₀ →0)	
(EVENTUALLY	ME-(BI CH)	1 TO ANY
	CH ₃ -CH, EXACTLY 1)	
- 3) GOODRINGS

2	EXACTLY 5
---	-----------
- 4) BADRINGS

3	
---	--

Figure 4. Superatoms and constraints supplied to CONGEN in investigations of plausible structural alternatives to the proposed structure of Polyalthenol.

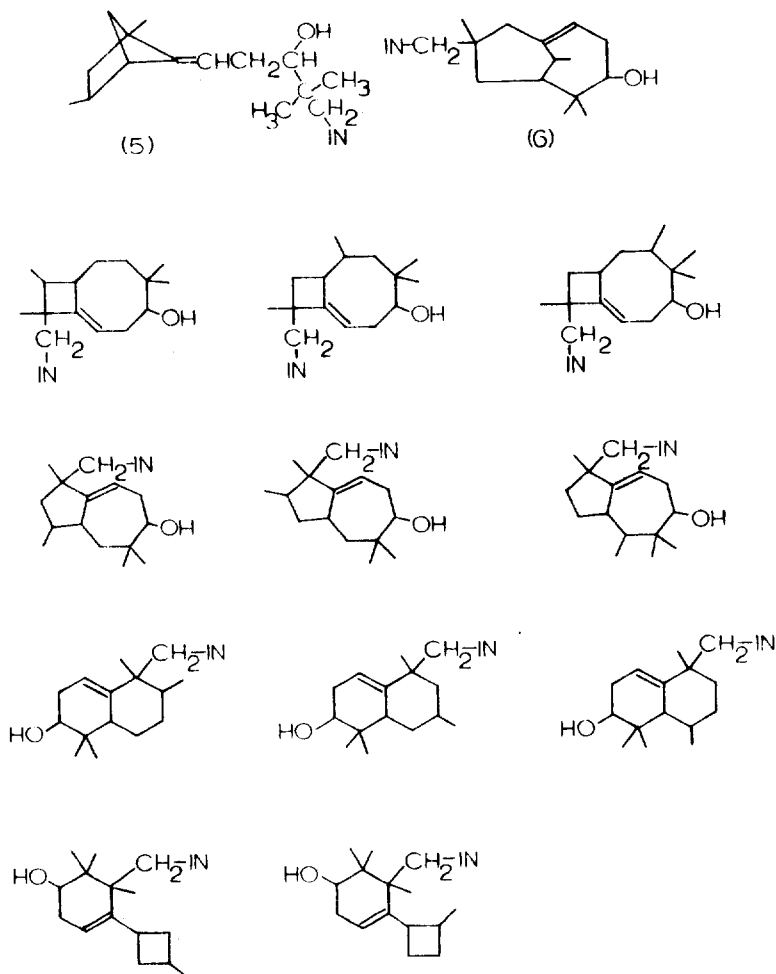


Figure 5. Structural candidates for Polyalthenol based on data given in Figure 4.

3 Applications of REACT to Structure Elucidation Problems

We have recently described our initial efforts toward representation of chemical reactions and their use in structure elucidation problems [Report HPP-76-5]. These efforts provided the framework for carrying out reactions within the computer which emulate actual laboratory reactions performed on a unknown. Constraints on the numbers and identities of the products are used to constrain the reaction products and, implicitly, the starting materials. Based on the results of that work we drew up a set of steps to be carried out to provide a truly useful tool for the chemist. Although the current program can be used in

applications to real problems it has some fundamental limitations which we have been working to solve. The developments we have undertaken to improve REACT are summarized in Figure 6.

REACTION CHEMISTRY DEVELOPMENTS

1. SEPARATION FROM CONGEN - COMMUNICATION VIA FILES OF STRUCTURES.
2. ADDING CONSTRAINTS - SITE - AND TRANSFORM - SPECIFIC.
3. CONTROL STRUCTURE - RAMIFICATION
 - A. ESTABLISH RELATIONSHIPS AMONG PRODUCTS AND REACTANTS
 - B. DEAL PROPERLY WITH RANGES OF NUMBERS OF PRODUCTS
4. INTERACTION - DEVELOP MANIPULATION COMMANDS WHICH PARALLEL LABORATORY OPERATIONS, E.G., SEPARATE INTO FLASKS, TEST CONTENTS OF VARIOUS FLASKS, INCOMPLETE SEPARATIONS, ETC.
5. REPRESENTATION OF REACTIONS
6. PROSPECTIVE DETECTION OF DUPLICATE PRODUCTS BASED ON SYMMETRY PROPERTIES OF: A) STARTING MATERIAL; AND B) TRANSFORMATION.

Figure 6. Current and future direction for improvement and extension of REACT, a program for exploration of applications of reaction chemistry to structure elucidation problems.

We first undertook to separate REACT from CONGEN, for two reasons. One reason was due to program size. Many functions of CONGEN are not needed in REACT and become unnecessary when only REACT is being exercised. The procedures of structure generation (CONGEN) and REACT are sequential and a separate program introduces no problems. A second reason was the different uses of certain CONGEN functions in REACT. For example, the ways in which the graph matcher is used are different between the two programs, necessitating keeping two different versions around with the programs together. The separation has been accomplished. The current version of REACT is now a separate program. It communicates structural information with CONGEN via files. All interactive portions are consistent with the structural manipulation functions of CONGEN so that learning the structural language of CONGEN is sufficient to use either program.

We have also added new constraint types to the reaction to expand greatly the ways in which reactions can be defined and constrained. An example of new extensions to reaction definitions illustrates some of the new features (Figures 7-10). The reaction defined here is one which will perform a dehydration of an alcohol; the site of the reaction is defined in Fig. 7.

```
:EDITREACT
NAME:DEHYDRATION
(NEW REACTION)
```

```
*SITE
>CHAIN 3
>ATNAME 1 0
>HRANGE 1 1 1 3 1 3
>ADRAW
```

DEHYDRATION: (HRANGES NOT INDICATED)

O-C-C

>DONE

```
*TRANSFORM
>UNJOIN 1 2
>JOIN 2 3
>DELATS 1
>ADRAW
```

DEHYDRATION: (HRANGES NOT INDICATED)

C=C

>DONE

Figure 7. Definition of reaction site and chemical transform in REACT.

The transform is defined as cleavage and loss of the oxygen resulting in formation of a double bond between the two carbon atoms of the original site (Fig. 7). In this particular dehydration the chemist wished to specify a site-specific constraint. It was known that a tertiary butyl group was part of the structure, and the dehydration will be prevented if that group is in close proximity to the reaction site (i.e., in a position alpha to the carbinol carbon).

```

*DEFINE-CONSTRAINTS
: ?
PLEASE ENTER ONE OF:
GRIPE          BUGOUT          GENERAL(G)     SITESPECIFIC(S)
TRANSFORMSPECIFIC(T)        DONE          HALT

: SITESPECIFIC
NAME: HINDERED
(NEW CONSTRAINT)
(WARNING: THE FINAL CONSTRAINTS MUST HAVE AT LEAST ONE ATOM OF THE
SITE)
> NDRAW

HINDERED: (HRANGES NOT INDICATED)
NON-C ATOMS: 1 0

1-2-3

> BRANCH 3 2 4 1 4 1
> ADRAW

HINDERED: (HRANGES NOT INDICATED)

      C
      |
O-C-C-C-C
      |
      C

> DONE

```

Figure 8. Definition of a site-specific constraint to be applied to the reaction DEHYDRATION.

The definition of this constraint is given in Figure 8. Subsequently, this constraint ("HINDERED") is placed on BADLIST for constraints specific to the site as shown in Fig. 9. The completed definition of the reaction is summarized in Figure 10.

*CONSTRAINTS

:?

PLEASE ENTER ONE OF:

GRIPE

BUGOUT

ST FOR CONSTRAINTS ON STARTING MATERIAL

S FOR SITESPECIFIC CONSTRAINTS

T FOR TRANSFORMSPECIFIC CONSTRAINTS

PR FOR CONSTRAINTS ON PRODUCTS

DONE

HALT

:S

>BADLIST

BADLIST CONSTRAINTS

CONSTRAINT NAME:HINDERED

CONSTRAINT NAME:

>DONE:DONE

Figure 9. Specification of constraint named HINDERED as a BADLIST constraint for the reaction.

```

*SHOW
SITE:
NAME=DEHYDRATION
ATOM# TYPE ARTYPE NEIGHBORS HRANGE
  1   O  NON-AR   ?       1-1
  2   C  NON-AR   1 3     1-2
  3   C  NON-AR   2       1-2

DEHYDRATION: (HRANGES NOT INDICATED)
NON-C ATOMS: 1  0

1-2-3

TRANSFORM:
  UNJOIN 1 2
  JOIN 2 3
  DELATS 1

DEHYDRATION: (HRANGES NOT INDICATED)

2=3

CONSTRAINTS:
CONSTRAINTS ON STARTING MATERIAL:
NO CONSTRAINTS
SITE-SPECIFIC CONSTRAINTS:
-----
BADLIST CONSTRAINTS
  NAME
  HINDERED
-----
TRANSFORM-SPECIFIC CONSTRAINTS:
NO CONSTRAINTS
CONSTRAINTS ON PRODUCTS:

NO CONSTRAINTS
*DONE
(DEHYDRATION DEFINED)
(DEHYDRATION ADDED TO THE REACTION LIST)

```

Figure 10. Summary of the completed definition of the DEHYDRATION reaction.

The remaining items summarized in Figure 6 are currently under development. We are redesigning the control structure so that the scientist using the program can use intuitive concepts as commands, such as separation. To carry this out important parts of the current mechanism have to be redesigned. Although the current program can be used effectively, its non-intuitive approach to dealing with reactions yielding multiple products and subsequent separation (within the computer) and analysis of each product presents a barrier to use by a wider community. We are continuing to develop our capabilities for representing reactions to ensure that the user of REACT has a complete descriptive language with which to specify reactions. We continue to study ways to avoid duplication in carrying out reactions. We know how to implement certain of the symmetry-related constraints and will do so shortly.

4 CONGEN Developments

The problem solving paradigm that has emerged from DENDRAL work is the so-called "plan-generate-test" paradigm. It is based on heuristic search of a space of possible hypotheses with planning before generation of hypotheses and testing of each generated candidate.

The generator for DENDRAL, named CONGEN, is a general-purpose graph generator which produces a list of all possible graphs containing specified numbers of nodes of various types. The most important features of the generator are that the list of graphs is guaranteed to be complete and non-redundant and, equally important, that the list need not be exhaustively generated. The generator can be constrained to produce only graphs that meet specified criteria that are inferred from the initial problem data.

During the past year, CONGEN has developed along two major lines: 1) tools have been developed which will allow more efficient and "intelligent" use of substructural information supplied by the chemist; and 2) data from chemical reactions and from observed mass spectra can be used to eliminate unlikely structural candidates from a set produced by a CONGEN generation. These extensions will be discussed below.

4.1 Intelligent use of constraining substructural information

There is sometimes a significant conceptual gap between the intuitive chemical phrasing of a CONGEN problem and the phrasing which is most efficient, in both computer time and storage requirements, for the program. CONGEN provides a rich language for stating structure elucidation problems in precise substructural terms. However, there are usually many ways of defining a given problem and different definitions can place widely different demands upon the program. We have a continuing interest in reducing this conceptual gap by in making CONGEN responsible for rephrasing a problem in the most efficient way, thus freeing the chemist to concentrate upon the chemical, rather than the algorithmic, aspects of a given case.

One distinction which is frequently puzzling to new CONGEN users is the one between superatoms and GOODLIST items. A superatom is a polyatomic "building block" which CONGEN joins with other superatoms and single atoms to form full structures. GOODLIST items are substructures which are required to be present in those full structures, but they are not incorporated directly into the initial phrasing of a problem as are superatoms. Rather, their presence or absence is tested by a graph-matching

routine after the structures are produced. Frequently, a great many structures produced by the structure generator are discarded by this final test and a significant amount of the program's time can be spent "shooting blanks". The concepts behind these two types of constraints - that specified substructural features must be present - are similar, but their implementations differ substantially in efficiency.

GOODLIST items cannot simply be transferred to the superatom list, though, because GOODLIST items are allowed to share atoms and bonds with other GOODLIST items or with superatoms. For example, if two substructures which are benzene rings are placed on GOODLIST, then a naphthalene derivative will be an acceptable structure even though the two occurrences of the ring have two atoms and one aromatic bond in common. Because of the building-block nature of superatoms, they may be joined to one another by additional bonds in CONGEN, but never "merged" (i.e., overlapped). Thus the price of efficiency is a more restricted interpretation of structural possibilities for superatoms.

We have developed a new procedure which captures the best of both situations. In order to incorporate a GOODLIST substructure into the problem at the earliest stage, it is necessary to find all unique ways that the given substructure can be created using parts of the existing building blocks (atoms and superatoms). This produces a set of new CONGEN problems with more or larger superatoms, each of which is easier to solve than the original one because the GOODLIST item is built-in and needs not be tested. Figure 11 shows schematically some of the ways this construction might occur: a) by bonding together two (or more) existing superatoms to create one larger one; b) by bonding additional atoms to a superatom to create a larger one; and c) by constructing a copy of the substructure from single atoms, creating a new superatom.

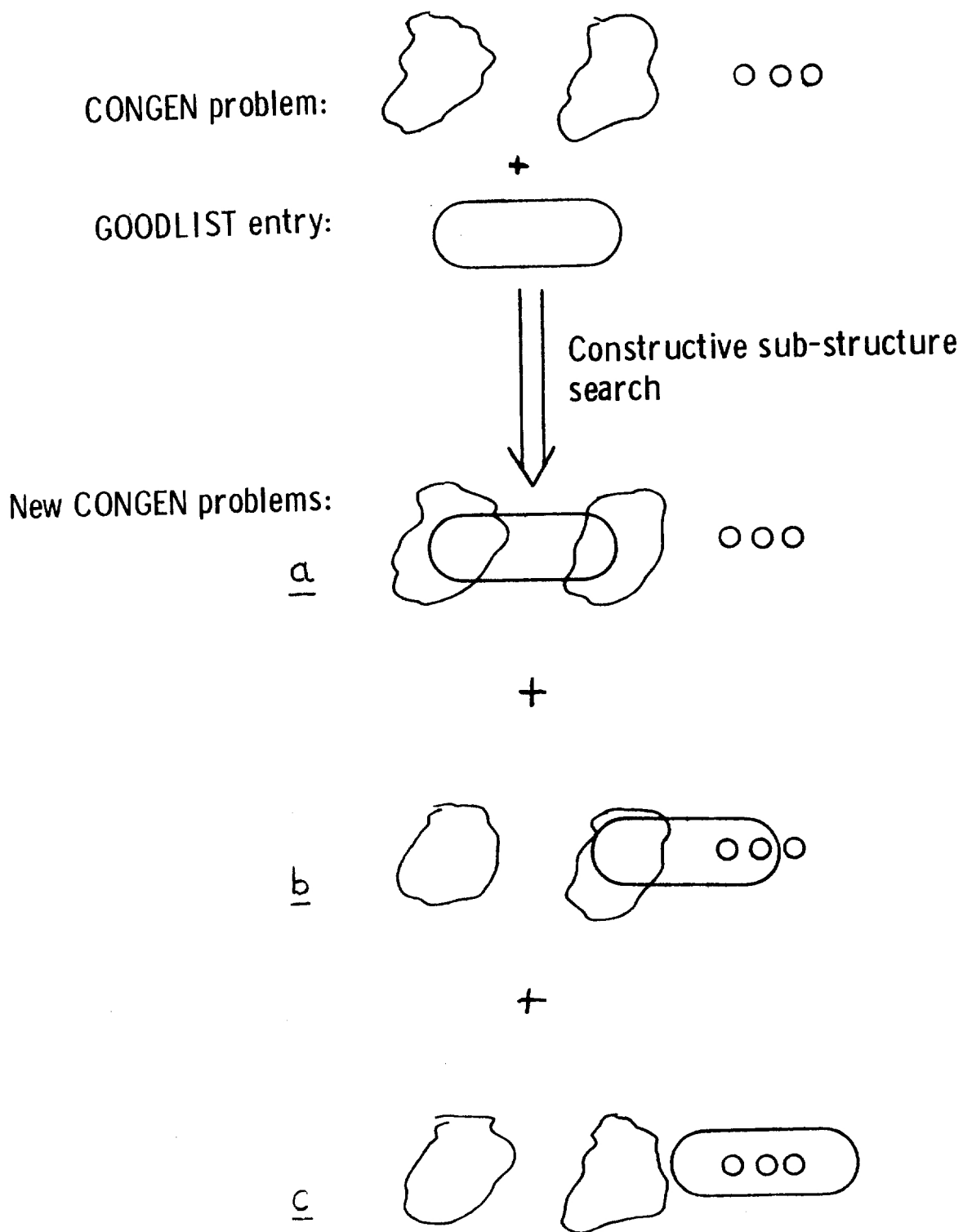
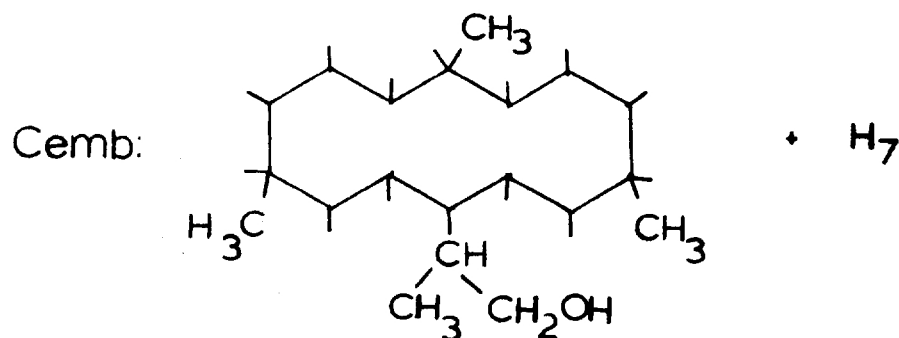


Figure 11. Example of breaking one GOODLIST substructure into several subproblems for CONGEN, each with different superatoms.

The algorithm is derived from the CONGEN graph-matching routine with the additional feature that as it searches for the substructure it is allowed to create new bonds (up to the limit of available new bonds in the original CONGEN problem) whenever they are necessary for the search to proceed. During the search, full account is taken of the topological symmetry of the superatoms in the original problem so that fittings which are redundant with respect to these symmetries are avoided. The substructure itself may possess some symmetry as well, but this is currently not considered.

Figure 12 summarizes a CONGEN problem which was attempted but which could not be completed because of the unintelligent use of GOODLIST. The problem amounts to finding all ways of allocating three new bonds to the free valences (the bonds with unspecified termini) in the superatom CEMB such that the three indicated substructures are present in the final molecules. There are perhaps 10,000 unique allocations of those three new bonds, but only 7 pass the GOODLIST tests. Using GOODLIST as a post-test only, CONGEN would generate all 10,000 and discard nearly all of them, a process which would have been so lengthy that it was never completed. The constructive graph-matching routine approaches the problem in a much more efficient and chemically intuitive way: 1) there are only three places in which the first GOODLIST item can be constructed; 2) for each of these, there are four ways of constructing the second; and 3) for each of these, there are 0, 1 or 2 ways of incorporating the third. It quickly arrives at the correct set of solutions.



GOODLIST:

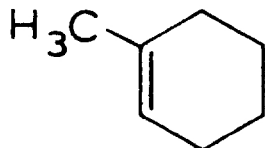
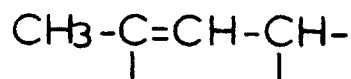
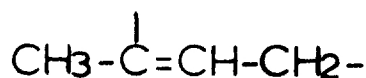


Figure 12. Example showing the inefficiency of specifying a constraint as a GOODLIST item instead of analyzing its implications for constructing allowable chemical graphs.

Most CONGEN problems contain one or more GOODLIST items which can be processed in this way, and when the constructive graph-matcher is fully integrated into CONGEN, it will make a substantial difference in its ability to use this structural information effectively.

4.2 New tools for post-pruning CONGEN structures.

From an algorithmic standpoint, CONGEN is successful if it can, in a reasonable amount of time and without exhausting storage resources, produce a list of candidate structures satisfying the chemist's constraints. However, this list is often quite large, perhaps several hundred structures, and from a chemical standpoint the problem may be far from complete. It

remains for the chemist to discriminate among the candidates, eventually reducing the possibilities to just one structure. A SURVEY function is available for classifying the list into groups of chemically related structures using either pre-defined or user-defined libraries of substructural features, and this process can help the chemist perceive groups which might easily be ruled out by additional experiments. Also, the graph-matching (pruning) mechanism of CONGEN allows him to express, in terms of substructural tests on the candidates, new data which he gathers on the unknown. These are both important aids in dealing with a list of candidates, but are restricted to tests which can easily be phrased purely in terms of structural features of the candidates themselves.

There are two informative sources of data which cannot always be phrased in this way: 1) structural features observed in products of the unknown when it undergoes simple chemical reactions; and 2) empirical spectroscopic measurements on the unknown which cannot be interpreted unambiguously in precise structural terms. During the past year, we have made progress in utilizing such information. The program REACT addresses the first problem while MSRANK concerns the second, in the context of mass spectrometric observations.

4.2.1 REACT

This program [see Report HPP-76-5] has two basic goals: 1) to provide the chemist with a computerized language for defining graph transformations and applying them to structures, thus simulating chemical reactions; and 2) to automatically keep track of the interrelationships between structures in a complex sequence of reactions so that whenever structural claims are made ruling out structures at one level, the implications in terms of structures at other levels can be traced. During the last year some progress has been made toward both of these goals.

EDITREACT, the reaction-editing language, has been extended to allow the user to define subgraph constraints which apply relative to a potential reaction site rather than to the molecule as a whole. For example, in the present version of REACT, we can say either that a hydroxyl group (OH), if present anywhere in the reactant molecule, would inhibit the reaction, or that such inhibition would take place only if the OH group is adjacent to the reaction site. Such site-specific constraints, applied either before or after the transformation (i.e., reaction) has been carried out on the site, are critical to the detailed description of real chemical reactions. The inclusion of this facility in REACT substantially increases its usefulness in real-world chemical problems.

The bookkeeping problem has undergone a complete

reconceptualization in the past year, the purpose being to mimic more closely the actual steps taken by a chemist in the laboratory. In the initial implementation, a set of products arising from the application of a given reaction to a given starting structure could be subjected to a multi-level classification which grouped the products based upon user-defined substructural constraints. Each of these classes had an associated minimum and maximum number, representing the numbers of products which were allowed to be members of the class. Any starting materials whose products could not satisfy these conditions were removed from the list of candidates. Structures in any class could be further reacted, their products classified, and so on. This treatment of bookkeeping was sufficient for stating many chemical problems. For example, suppose a chemist knew that a particular reaction on an unknown compound yielded two carbonyl compounds (i.e., containing C=O), at least one of which was an ester (-O-C=O). He could define a product class CARBONYL using the C=O substructure with a minimum and maximum of two products. He could then define a sub-class of CARBONYL called ESTERS using the substructure -O-C=O with a minimum of one and a maximum of two products. The program would automatically use this information to eliminate candidate starting structures which could not give the indicated product distribution with the given reaction.

There are chemical problems, though, for which the above scheme is too rigid. For example, suppose a reaction gives several products, two of which are isolated and labelled P1 and P2. Suppose that only a small amount of P1 is available so only mass spectroscopic measurements are practical. Suppose also that a deuterium-exchange experiment shows that P1 has two exchangeable protons (say, either N-H or O-H). P2 shows a strong carbonyl absorption in the IR. P1 might also contain a carbonyl group, but that was never determined, and neither was the number of exchangeable protons in P2, which could be two. No matter how one attempts to use the above-described classification system, one cannot express this information accurately.

In the new approach, for which the algorithmic design has been completed, one is allowed to express data in a much more natural sequence which parallels the experimental steps. The first experimental step after a reaction is usually the separation and purification of products. An analogous step is to be included in REACT, in which the separation amounts to the setting up of a specified number of labelled "flasks" (analogous to the labels P1 and P2 in the above example) each of which is ultimately to contain a specified number (usually 1) of the products. As experimental data are gathered on each real product, corresponding substructure constraints are attached to the corresponding flask in the program. As each such assertion is made, the bookkeeping mechanism verifies that, for a set of reaction products from a given starting material, there is at

least one way of distributing them among the flasks such that each product satisfies the constraints for its flask. If this test is ever violated, the starting material is removed as a candidate structure. Flasks containing more than one product may be further separated into "subflasks" to any level, and the contents of any flask may be made to undergo further reactions. This capability, the reacting of flask contents, is analogous to common laboratory procedures in which incomplete separations of products are encountered. Dealing with such situations adds considerable complexity to the bookkeeping mechanism, because the contents of a flask may be ambiguous to the program when the reaction is applied. REACT must keep track of all possible structures which might, based on the current flask constraints, occupy the reacting flask. If such a reaction fails (because the products did not satisfy the constraints specified for them), REACT does not eliminate the starting structure entirely, but notes that the structure may not occupy that flask in future flask-allocation tests.

4.2.2 MSRANK

This program is an outgrowth of MSPRUNE described in last year's annual report. It is a combination of a predictor which uses a very simple theory of mass spectrometry to predict the spectra of candidate structures, and an evaluation function which compares the predictions with the observed spectrum of the unknown, assigning a goodness-of-fit score to each candidate. The candidates are then sorted based upon how well they match the observations. The basic concept here is not a new one to the DENDRAL project [see, for example, Buchanan, et al. in Machine Intelligence 4 (Meltzer & Michie, eds., Edinburgh Univ. Press, 1969)], but there are some new aspects to the problem when viewed in the overall CONGEN context.

Because of the wide variety of structural types which can be produced by CONGEN, it is necessary for MSRANK to use a very general model of mass spectrometry. The best predictive theories of mass spectrometry are limited to families of closely related structures (i.e., class specific theories), and the Meta-DENDRAL program is designed to help in discovering such theories. There are very few general principles upon which to draw in predicting mass spectra, though, so MSRANK is limited to only the most approximate kinds of evaluation functions. One principle which we noticed being used by practicing mass spectrometrists was: of two candidate structures for an unknown, the most likely structure is the one which explains the observations most "simply" - i.e., with the fewest complex explanations involving many bond cleavages and the transfer of many hydrogen atoms. The evaluation function used by MSRANK is based on a quantitation of this principle.

In predicting a spectrum, MSRANK explores all possible cleavages of the molecule within some very general user-defined constraints concerning the number of bonds broken and the number of steps in a process, the proximity of pairs of cleaved bonds (i.e., whether or not two adjacent bonds can break in a given process) and the multiplicity or aromaticity of each cleaved bond. Within these general limits, the user also supplies numerical plausibilities from 0 to 1 on the various kinds of breaks which are allowed to occur. For example, he might give unit plausibility to 1-bond cleavages, .8 to 2-bond processes and .6 to 3-bond processes. Aromatic-bond, multiple-bond and adjacent-bond cleavages, if allowed, are given separate plausibilities, as are the allowed neutral transfers. MSRANK combines these values multiplicatively in evaluating the overall plausibility of a specific mass spectral process, and that value is associated with the corresponding predicted mass point. If two different processes predict the same mass point, the higher plausibility value is retained. The result is a predicted spectrum with numbers attached to each peak, interpreted roughly as the "reasonableness" or "simplicity of prediction" measure.

We expect such a theory to be overly complete in the sense that, when applied to the correct structure for an unknown, it will doubtless predict many plausible peaks for which there is no observation. This simply reflects the fact that the "break everything" approach to mass spectrometry is a considerable oversimplification. Thus the evaluation function does not penalize for predicted but unobserved peaks. What we do expect, though, is that a large number of the observed peaks, particularly the intense ones, will have plausible explanations with respect to the correct structure. Thus a "reward" is given to every observed peak which is predicted, the amount being proportional to the plausibility of the prediction and (at the user's option) to the intensity and/or mass value of the observed peak. The sum of rewards for all observed peaks then constitutes the overall score for the candidate which gave rise to the predicted spectrum.

MSRANK is quite new and we have not yet had sufficient experience with it to evaluate its overall usefulness. By using only unit plausibilities for selected characteristics of the mass-spectral cleavages, we are able to duplicate earlier results obtained with the predictor/comparator functions applied to mono- and di-ketoandrostanes. These tests serve to check the accuracy of the MSRANK program. We are now doing a systematic study of various classes of compounds by ranking the spectrum of a known structure against a CONGEN-generated list of structures which contains the correct one among several which are closely related.

5 USE OF CONGEN BY OTHER SCIENTISTS

The number of persons experimenting with CONGEN has grown as a result of both the continuing practice of issuing an "invitation for program trial use" at the conclusion of publications, as well as continuing personal contact between Dendral project members and potential program users. Three categories of users make up this group:

5.1 Chemists Using Exported Programs

The part of CONGEN responsible for teletype output of chemical structures (the DRAW program) is coded in Fortran. Since the paper describing this program appeared in print [R. Carhart, JACS, 16:82, 1976], we have exported the program to half a dozen sites, ranging from Japan, across North America, to England. Similarly, the entire CONGEN program, is largely coded in Interlisp and SAIL, and has been exported to a collaborator in England who is very interested in the methods and programming techniques employed in coding the program. Another program which we have exported for use by other chemists is the PDP-11 CLEANUP program which was described in ANALYTICAL CHEMISTRY [48:1368, 1976]. This program "cleans up" new GC/MS data to eliminate noise peaks and to separate the data associated with components in the mixture.

In each case, the requestors were provided with an initial choice of format options from which they could select the one most suitable for their computer installation. They were asked to send a 2400 foot reel of magnetic tape appropriate to the selected format option. The programs were written on the tape and returned to them along with a brief written explanation of program organization. Accurate records are kept of who has received the programs, so that omissions and errors can be corrected by mail at a later date, if ever necessary.

1. Dr. James F. Elder, Dow Chemical U.S.A., Midland, Michigan.
2. Dr. Robert M. Supnik, Massachusetts Computer Associates, Inc., Wakefield, Massachusetts.
3. Mr. Dan Pearce, Orange County Sheriff-Coroner Department, Santa Ana, California 92702
4. Dr. H. J. Stoklosa, Central Research & Development Department, E. I. du Pont de Nemours & Company, Wilmington, Delaware.
5. Dr. Douglas W. Kuehl, Environmental Research Laboratory-Duluth, Duluth, Minnesota.

6. Dr. Richard A. Graham, Food Sciences Laboratory, U. S. Army Natick Laboratories, Natick, Massachusetts.
7. Dr. Walter M. Shackelford, United States Environmental Protection Agency, Environmental Research Laboratory, Athens, Georgia.
8. Dr. Richard Gans, Chemical Research Division, American Cyanamid Company, Bound Brook, New Jersey.
9. Dr. John C. Marshall, Department of Chemistry, the University of North Carolina, Chapel Hill, North Carolina.
10. Dr. Graham S. King, Department of Chemical Pathology, Queen Charlotte's Hospital for Women, London, England.
11. Dr. J. Wyatt, Chemistry Division, Naval Research Laboratory, Washington, D. C..
12. Dr. Gareth Templeman, Research and Development Laboratories, The Pillsbury Company, Minneapolis, Minnesota.
13. Dr. J. B. Justice, Department of Chemistry, Emory University, Atlanta, Georgia.
14. Dr. Thomas Knudsen, Northrop Services, Environmental Sciences Group, Research Triangle Park, North Carolina.
15. Dr. Ingolf Meineke, Fachbereich Chemie, Philipps Universitaet, Lahnberge, West Germany.
16. Dr. M.A. Shaw, Unilever Research, Port Sunlight Laboratory, Wirral, Merseyside, England.
17. Dr. Ernst Weber, Varian MAT, Bremen, West Germany.
18. Paul V. Fennessey, Department of Pediatrics, University of Colorado Medical Center, Denver, Colorado.
19. R. G. A. R. Maclagan, Department of Chemistry, University of Canterbury, Christchurch, New Zealand.
20. James E. Oberholtzer, Arthur D. Little, Inc., Cambridge, Massachusetts.
21. F. Street, AEI Scientific Apparatus Limited, Manchester, England.

5.2 Remote Users of SUMEX

Due to the fact that the SUMEX computer is available via both the TYMNET and ARPANET communication networks, it is possible for scientists in many parts of the world to directly access the Dendral programs on SUMEX. Primary usage is centered on CONGEN, although INTSUM is beginning also to gain a following. Although access points to SUMEX are widespread, they frequently are not diverse enough to accommodate the dispersed group of scientists who have expressed an interest in using one of the Dendral programs. For example, Dr. Joseph Baker of the Roche Institute of Marine Pharmacology in Dee Why, Australia, is looking at the possibility of accessing SUMEX by using International Direct Distance Dialing (IDDD).

5.3 Chemists Communicating by Mail

Many Scientists interested in using DENDRAL programs in their own work are not located near a network access point. Users of this type choose to use the mail to send details of their structure elucidation problem to a Dendral Project collaborator at Stanford.

5.4 Chemical Problems Posed to CONGEN

Following is a list of CONGEN users, and a brief summary of their program interests during the past year.

1. Dr. Roger Hahn, Syracuse University. While at Stanford he used CONGEN to help solve the structures of photoproducts by obtaining all possibilities under available constraints and designing NMR experiments to differentiate the possibilities. This work will be published soon.
2. Dr. William Epstein, University of Utah. During a demonstration of CONGEN, he posed a problem to verify that the structural possibilities he determined for an unknown were in fact all possibilities. The structure of methyl santolate has been published (see Epstein, et al., J.C.S. Chem. Commun., 590 (1975)).
3. Dr. Clair Cheer, University of Rhode Island. While on sabbatical at Stanford, Dr. Cheer has worked on a number of structure elucidation problems using CONGEN including Briareine D and [+]-Palustrol (Cheer et al., Tetrahedron Letters, 1807 (1976)). Work is

5.2 Remote Users of SUMEX

Due to the fact that the SUMEX computer is available via both the TYMNET and ARPANET communication networks, it is possible for scientists in many parts of the world to directly access the Dendral programs on SUMEX. Primary usage is centered on CONGEN, although INTSUM is beginning also to gain a following. Although access points to SUMEX are widespread, they frequently are not diverse enough to accommodate the dispersed group of scientists who have expressed an interest in using one of the Dendral programs. For example, Dr. Joseph Baker of the Roche Institute of Marine Pharmacology in Dee Why, Australia, is looking at the possibility of accessing SUMEX by using International Direct Distance Dialing (IDDD).

5.3 Chemists Communicating by Mail

Many Scientists interested in using DENDRAL programs in their own work are not located near a network access point. Users of this type choose to use the mail to send details of their structure elucidation problem to a Dendral Project collaborator at Stanford.

5.4 Chemical Problems Posed to CONGEN

Following is a list of CONGEN users, and a brief summary of their program interests during the past year.

1. Dr. Roger Hahn, Syracuse University. While at Stanford he used CONGEN to help solve the structures of photoproducts by obtaining all possibilities under available constraints and designing NMR experiments to differentiate the possibilities. This work will be published soon.
2. Dr. William Epstein, University of Utah. During a demonstration of CONGEN, he posed a problem to verify that the structural possibilities he determined for an unknown were in fact all possibilities. The structure of methyl santoninate has been published (see Epstein, et al., J.C.S. Chem. Commun., 590 (1975)).
3. Dr. Clair Cheer, University of Rhode Island. While on sabbatical at Stanford, Dr. Cheer has worked on a number of structure elucidation problems using CONGEN including Briareine D and [+]-Palustrol (Cheer et al., Tetrahedron Letters, 1807 (1976)). Work is

continuing on the structure of another marine natural product, presumably a cembrenolide, for which there are currently seven possibilities.

4. Dr. Jerrold Karliner, Ciba-Geigy Corporation. Dr. Karliner has solved several structural problems using CONGEN, including material with flame retardant properties, an impurity in a production sample and nitrogen heterocycles being investigated for pharmacological activity. CONGEN enabled reduction of the number of possibilities to the point where subsequent experiments led to unambiguous structural assignment.
5. Dr. Gino Marco, Ciba-Geigy Corporation. He has used CONGEN to help solve structures of conjugates of pesticides with sugars and amino acids.
6. Dr. Milton Levenberg, Abbott Laboratories. He has worked on the structure of a compound with mild antibiotic activity, isolated from a fermentation broth. There are currently ten structural possibilities, reduced to that number from the 33 initially determined using CONGEN by additional experimental data.
7. Dr. David Pensak, DuPont. He is currently learning to use CONGEN and plans to evaluate its utility for structural problems of some of his coworkers.
8. Dr. Douglas Dorman, Eli-Lilly. He is using CONGEN to assist in structure elucidation of metabolites of microorganisms shown to have pharmacological activity. He has worked on five such problems, including a current one where the developing MSPRUNE capabilities are being used.
9. Dr. L. Minale, Napoli, Italy. We have worked with him by sending him structural alternatives for proposed structures for some marine natural products (Pallescensins, Tetrahedron Letters, 1417 (1975)) and cyclic diethers from the lipid fraction of a thermophilic bacterium (J. C. S. Chem. Commun., 543 (1974)).
10. Dr. K. Nakanishi, Columbia University. We have worked with him by sending him structural possibilities for termite defense compounds (structure finally solved by X-ray crystallography). This trial plus a live demonstration to one of his students has resulted in efforts toward continued collaboration on other insect defense secretions and

exploration of the possibility of his direct access to SUMEX.

11. Dr. L. Dunham, Zoecon Corporation. We have collaborated with him on the use of INTSUM for mass spectral fragmentation studies of insect juvenile hormones.
12. Dr. A. G. Gonzales, Tenerife, Spain. We have recently sent him structural alternatives for constituents of *Laurencia Perforata* (Tetrahedron Letters, 2499 (1975)), and expect to continue discussions on the structures of these compounds.
13. Dr. T. Irie, Sapporo Japan. We have recently sent him structural alternatives to published structures on constituents of *Laurencia Glandulifera* (Tetrahedron Letters, 821 (1974)) and expect to continue discussions on this problem.
14. Dr. C. J. Persoons, Delft. We have corresponded with him on structural alternatives for cockroach sex pheromones (Periplanone-B (Tetrahedron Letters, 2055 (1976))), and he has agreed to further collaboration on new problems.
15. Dr. F. Schmitz, University of Oklahoma. We explored for him structural alternatives for an unknown diterpenoid hydrocarbon. We obtained 25 possibilities, of which only four obeyed the isoprene rule.
16. Dr. J. Baker, Roche Institute of Marine Pharmacology, Australia. We plan collaboration with Dr. Baker on the sterol fractions of various marine organisms and are exploring ways for him to access CONGEN.
17. Dr. E. VanTamelon, Stanford University. We have used the developing reaction features of CONGEN to explore structural possibilities for both chemical and biogenetic cyclization products of squalene-oxide congeners. We have suggested alternatives to proposed structures and helped to design experiments to differentiate them.
18. Dr. J. C. Braekman, Brussels. Dr. Braekman visited Stanford as a part of continuing collaboration in marine chemistry with Dr. Tursch's group. While at Stanford he explored use of CONGEN for use in current problems in marine natural products, and worked on the problems of Drs. Irie and Gonzales (see above).

He is currently exploring access to CONGEN from Brussels, via TYMNET.

Some problems have arisen as a result of the Dendral commitment to working with outside chemist users. The primary area of difficulty arises from the fact that the Dendral project, as one of the many projects which use the SUMEX facility, is allocated a certain portion of system resources. Therefore, support of an extensive body of outside users means that resources to support these users must be diverted from the research goals of the project.

In encouraging new users, Dendral must be careful to state that access to Dendral programs might have to be restricted in the future if system loading becomes extensive. Understandably then, some scientists are reluctant to invest time in learning to use a complicated, although potentially useful program which they may well only be able to use on a temporary basis. One solution to this problem is to make the available programs as efficient as possible, and/or to make it possible to distribute copies of the program to other sites.

Use of CONGEN by working scientists has turned up one major area in which additional information to the user was thought to be necessary. CONGEN users unanimously indicated their desire for a method what percentage of the whole problem was solved at any moment, i.e., total number of possible structures is represented by the number already generated. In a prototype system we have implemented the Cntrl-I and Cntrl-S user information interrupts, to show how far CONGEN has progressed. If, for example, someone who has generated 357 structures is told that this indicates that they have generated 1 percent of the total possible structures, they immediately know that they do not want to finish generating all the structures. Even if there were enough space, 40,000 structures would be far more than they would want to see.

We implemented another user-oriented facility for an invited paper presented at the 172nd American Chemical Society meeting, in August of 1976. Special features were added for a character-oriented, screen-addressable CRT terminals to give users an informative visual interface to CONGEN, an otherwise complex. The dynamic field of view provided by this type of terminal was used to advantage to give the chemist-user a continuous, graphic summary of both the information he has supplied to the program and the dynamic use of that information by the program.

6 Stereochemistry in CONGEN

We have started the complex task of giving CONGEN the capability of recognizing stereochemical features of molecules and using stereochemical information in structure determination. The ability to recognize stereochemical features would allow, for example, the generation of all stereoisomers of a given topological structure with or without constraints. The ability to use stereochemical information would allow the determination of constraints on stereoisomer (and topological isomer) generation caused by, for example, partial knowledge of relative or absolute stereochemistry of structural fragments, knowledge of overall molecular chirality (or lack of), absolute and relative stereochemistry from circular dichroism measurements, and so forth. Thus far, only the topological information (constitution) has been recognized and used by CONGEN.

The first stage of this development is to produce a program which generates all the stereoisomers of a given topological structure. This program will be placed at the end of the existing CONGEN program. The present report describes the development of the theory and algorithm for stereoisomer generation and the progress on the programming of this algorithm.

6.1 Algorithm

The carbon stereoisomers of a given topological structure are in correspondence with the double cosets:

$$\text{TSG}[A4] / \text{TSG}[S4] / \text{CSG}$$

in which:

1. $\text{TSG}[A4]$ is the wreath product of the Topological Symmetry Group and the alternating group $A4$. This group expresses the invariance of a carbon stereoisomer to all even permutations of the ligands connected to any carbon stereocenter.
2. $\text{TSG}[S4]$ is the wreath product of TSG and the symmetric group $S4$. This group expresses the invariance of the connectivity of a topological structure to all permutations of ligands connected to any carbon center.
3. CSG is the Configurational Symmetry Group and is isomorphic to the TSG represented on the two-valued configurations of the carbon stereocenters.

The cosets of $\text{TSG}[A4]$ in $\text{TSG}[S4]$ correspond to the 2^m maximum possible stereoisomers where m is the number of carbon

stereocenters. The effect of the group CSG on these cosets is to collect the possible stereoisomers into equivalence classes of distinct stereoisomers. Intuitively this corresponds to the mental process of considering all possible stereoisomers of a topological structure and collecting those equivalent by symmetry.

The algorithm to generate stereoisomers from a CONGEN topological structure must perform three transformations:

1. The connection table (CT) corresponding to the CONGEN topological structure must be modified to include only those carbon centers which need be considered as stereocenters. That is, methylenes, methyls, carbons with gem-dimethyls etc., do not exhibit configurational stereochemistry. A prefilter must act on the CT and return a Stereocenter Connection Table (SCT).

2. The TSG which comes from CONGEN must be modified to give the CSG described above.

3. Given the SCT and CSG, the possible distinct stereoisomers must be generated. This involves an implementation of the theory presented in the previous paragraphs. Further details of this algorithm are given in the next section.

6.2 Programming Progress

All programming is being done in the SAIL language.

1. The development of a program to perform the prefilter function on the connection table is currently in progress. The CT will first be scanned to eliminate methylenes and methyls and then iteratively scanned to find identical achiral substituents on common carbons (gem-dimethyl, gem-diethyl, etc.).

2. A program to obtain the configurational symmetry group (CSG) from the topological symmetry group (TSG) has been written. The elements of TSG are allowed to act on the connection table and the parity of the permutations on each stereocenter is determined. The permutation with these parity designations is the desired element of CSG.

3. A program which, when given the SCT and CSG, will generate all distinct stereoisomers has been written. Special use is made of the fact that all elements of the CSG will be hyperoctahedral group elements. That is, CSG will be a subgroup of the wreath product $S_n[S_2]$, called the hyperoctahedral group, where n is the number of stereocenters. The order 2 group, S_2 , is represented by the two-valued configuration of each carbon stereocenter. This two-valued nature of each stereocenter's

configuration is easily represented by a single two-valued bit which makes a very compact machine representation. The program has the capability of representing the hyperoctahedral group by bit permutation and reversals. This will accommodate any conceivable symmetry and any stereochemistry resulting from carbon (or analogous element) configurations.

As an example, consider the problem of the number of stereoisomers of inositol, $(\text{CH}(\text{OH}))_6$. The CSG can be obtained from the TSG as described above and when input with the stereochemical connection table to this segment of the program, the desired 9 isomers are found and output as canonical structures based on the original atom numbering. (This will probably not be the final choice for a canonical stereostructure.) The interfacing of these segments of the stereoisomer generator and the interfacing with the existing CONGEN program is also in progress.

7 The GC/HRMS DATA SYSTEM

7.1 Improvements to the Data System

The introduction of the gas chromatograph (GC) into the high resolution mass spectrometry (HRMS) system produced a number of problems in data reduction that are not present without the GC. The primary problem is the increase in the number of mass peaks in a spectrum from the column bleed of the GC. This makes the problem of separating calibration and reference peaks from the true sample peaks a much more difficult problem. A number of improvements have been introduced to the software to solve this problem.

The instrument is calibrated by injecting a sample of perfluorokerosene (PFK) and running REFRUN. This collects a spectrum which can be calibrated by looking for various characteristic peaks in the spectrum. The masses of certain peaks are stored on a file. Once these calibration peaks have been identified, the masses can be used to interpolate and find the mass of all other peaks in the spectrum. The results of a satisfactory reference run is stored on a file, as well as being listed on a line printer.

The spectrum of the sample is taken by running SAMRUN, which collects a spectrum of the sample and PFK. The main problem now is finding the peaks from PFK, and using them to calculate the masses of the peaks from the sample. The first ten calibration peaks are located by applying a template, or pattern matching algorithm to the data. This template assumes that characteristics of the mass spec will change only systematically with time. This has proven to be a very successful and sensitive method of locating calibration peaks. Once the initial ten peaks are located, the program scans the data by taking four calibration peaks and, using a model of the scan, projecting for a fifth. Once this is located the masses of the peaks in between the calibration peaks are interpolated, and a decision is made on whether a given peak was in the reference run, or is truly a sample peak. The four calibration masses are shifted so that the calibration peak just projected becomes one of the four, and the process is repeated until masses have been assigned to all of the peaks.

Problems occur when, during projection, either no peak or more than one peak is found as a calibration peak. If no peaks are found, the mass is counted as missing, and the next calibration mass is searched for. Since the calibration peaks are chosen as being among the most prominent peaks in the spectrum, the problem in this case is usually not that the peak is absent. The more common problem is that there are so many data peaks from the GC that more than one peak shows up as a candidate for the calibration peak. If the program chooses the wrong peak as the calibration peak, the crawl through the data quickly goes bad. Various schemes have been tried to minimize this problem. Originally the first peak in the window was chosen, since PFK has a very large negative mass defect. This produced occasional problems, however. Next a more sophisticated approach was tried. If the projection produced multiple candidates for the calibration peak, the two peaks closest to the projection were selected, and another projection was done from each of those. The one giving rise to the least total projection error was selected. We found one batch of data, however where it happened that at one section of the spectrum, two incorrect peaks produced a total error less than the two correct peaks. Neither of these algorithms use any information from the reference run, so an attempt was made to fold in the information from the reference run in the case of an ambiguity.

The spectrum is taken in an exponential downscan, i.e. high mass to low mass on an exponential curve. The only two parameters of this curve that can change are the time offset of the curve and the time constant of the exponential. The template mentioned earlier assumes that either of these parameters can change and attempts to find a set of peaks in the sample run that map most accurately into the reference run. This mechanism works well only in low masses, however, since in the higher masses the

curve is more gaussian than exponential. The template can be written for this, but the amount of space and time required for it made it appear impractical for the system. On examination of data it became obvious that the time constant of the exponential changed very little, if at all, from the reference run to the sample run. This means that a very good approximation of where a given calibration peak should appear can be obtained by merely adding in the time shift from the reference run. The final algorithm that resulted goes through the following steps: 1) the fifth calibration peak is projected. 2) if there is more than one candidate, then a projection is done on the two closest calibration peaks. 3) if both of these peaks project to another peak (there is still ambiguity, in other words), the peak which is closest to the time in the reference run based an exponentially weighted time shift from the previous calibration peaks is chosen. This has proven to be fast and reliable on the data tested so far, including data that had produced incorrect results from the previous two methods.

7.2 Changes in the Operating System

The current operating system for the PDP-11, DOS 9, has produced a number of problems. Poor keyboard interaction, generally slow response time, and extremely slow system programs, while surmountable, are factors that make the system difficult to use. We decided to look at the feasibility of changing to either RSX-11M or RT-11. RSX-11 proved to be too big and much more flexible than needed. RT-11 however had several advantages over DOS 9. The keyboard interaction is easier to use and more suited to a real time environment. The IO queuing structure is much simpler and faster, although the file structure is not as flexible as DOS 9. In addition the system itself is much faster, and there is a noticeable improvement in time of just loading programs. Based on this, a decision was made to switch from DOS to RT-11.

The conversion of programs from DOS to RT-11 has proven to be much more work than originally expected. The main problems have been incompatibilities between the two versions of FORTRAN and the different linkage editors. Since all the programs in the high resolution system are overlaid, this second factor has proven to a major problem, since some of the logic in the program must be reworked to make the overlay correct. The conversion effort has been aided by several factors, however. The speed of the system and system programs is often several times faster than similar programs under DOS. As an example, to link the REFRUN portion of the High Resolution system takes about 30 minutes under DOS, whereas the same program takes about 5 minutes under RT-11. The FORTRAN compiler and MACRO assembler are also faster.

The conversion, and software development in general, has been greatly improved by the addition of a teletype line from the SUMEX PDP-10 to the PDP-11. Programs have been written to transfer files between the two systems. This has had the effect of switching literally all of the editing to SUMEX because of the superior editor. The ease and speed of the file transfer makes it practical to make even minor modifications of a program on the the 10, and then transfer the edited version to the PDP-11. This process of using SUMEX to develop software will continue with the release of MAINSAIL, a machine independent language. MAINSAIL is a dialect of SAIL, which is a dialect of ALGOL 60. It has undergone many design changes since its original inception, but has been released in a limited version for the PDP-10. The main value of MAINSAIL is that programs written on one machine will be directly transportable to another with no modification. This allows us to write, test and debug software systems on SUMEX, which leaves only the the machine dependent portion of the system (for example the actual real time data acquisition) to be worked out on the PDP-11. This not only gives the programmer better tools (such as superior editors) but also frees up the PDP-11 for production work.

7.3 New Developments

In addition to upgrading old versions of the high resolution system, work is being done on creating a low resolution system for the MAT 711. The ultimate aim is collect data that can be run through CLEANUP, a program that resolves multiple spectra under a single GC peak, and cleans up the final spectra. The problem with the current system is that we cannot scan fast enough to provide CLEANUP the data it needs. The high resolution system requires resolution good enough to separate sample peaks from the reference peaks. If the scan is sped up past a certain point, SAMRUN can no longer separate the peaks, and therefore cannot calibrate the run. At the same time, CLEANUP requires at least 7 spectra across a GC peak be taken to insure resolution of multiple spectra. The fundamental problem then is that an alternate method of calibrating the mass spectrum, without using known calibration peaks, must be found before scan speeds required by CLEANUP can be achieved. The most direct solution to this is to directly measure the magnetic field strength of the instrument, and using it to calculate the mass that is being observed. To do this we inserted a hall probe between the poles of the magnet, and connected it to the data acquisition system on the PDP-11/20.

The main problems with the hall probe are as follows: 1) to make sure that the ion reading and the hall probe reading are simultaneous 2) to insure that the correct hall reading can be assigned to the correct ion reading 3) to determine the

reproducibility of hall readings versus mass being observed in both dynamic (scanning) and static situations and 4) to decide if the probe has the speed and accuracy to calibrate the instrument. The first two problems are a matter of hardware. The configuration of the original data collection system is as follows: the ion detector goes to an A/D converter, which is connected to a DMA. The DMA is on an 11/20, which has a data collection system, SAQMON, running. This performs various low level filtering and buffering operations. The DMA is actually a low level processor which counts the number of samples taken, stores them into successive memory locations, and interrupts the central processor when a block of data has been collected. The timing of the sample collection is controlled by a quartz crystal clock. On each timing pulse, a signal is sent to the A/D on the ion detector to convert that value to a digital number. To accommodate the hall probe, the DMA was modified so that on the timing pulse, the start signal is sent simultaneously to both the A/D on the ion detector and the A/D on the hall probe. The DMA then services both of the A/D's, and stores the readings in successive memory locations. The net result is that when the DMA interrupts the central processor, the block of data is a set of pairs of readings, an ion reading and the hall reading for that time. This solves both of the first two problems, since we now have the ion reading and the hall reading connected both in time and location.

The second two problems, testing the reliability and reproducibility of the hall probe, requires new software. We are currently modifying portions of the calibration mechanism of the high resolution system to calculate masses for a large number of hall readings.

8 META DENDRAL

The success of any reasoning program is strongly dependent on the amount of domain-specific knowledge it contains. This is now almost universally accepted within AI, partly because of DENDRAL's success. Because of the difficulty of extracting specific knowledge from experts to put into the program, many years ago we began to explore the problems of efficiently transferring knowledge into a program. We have looked at two alternatives to "hand-crafting" each new knowledge base: interactive knowledge transfer programs and automatic theory

formation programs. In this enterprise the separation of domain-specific knowledge from the computer programs themselves has been a critical component of our success.

One of the stumbling blocks with the interactive knowledge transfer programs is that for some domains there are no experts with enough specific knowledge to make a high performance problem solving program. We were looking for ways to avoid forcing an expert to focus on original data in order to codify the rules explaining those data because that is such a time-consuming process. Therefore we began working on an automatic rule formation program (called Meta-DENDRAL) that examines the original data itself in order to discover the inference rules for that part of the domain.

The problem solving paradigm for Meta-DENDRAL is also the plan-generate-test paradigm used in Heuristic DENDRAL. In this case one part of the program (RULEGEN) generates plausible rules within syntactic and semantic constraints and within desired limits of evidential support. The model used to guide the generation of rules is particularly important since the space of rules is enormous. The planning part of the program (INTSUM) collects and summarizes the evidential support. The testing part (RULEMOD) looks for counterexamples to rules and makes modifications to the rules in order to increase their generality and simplicity and to decrease the total number of rules.

Meta-DENDRAL successfully formulated rules of mass spectrometry that were new to the science. These rules, along with a discussion of the methodology, were published in the scientific literature [Report HPP-76-4]. The program was tested to see if it could rediscover the rules of mass spectrometry for two classes of chemical compounds that were already well understood (amines and estrogenic steroids). Then it was applied to three classes of compounds whose mass spectrometry was not as well known (mono-, di-, and tri-ketoandrostanes). The program produced three sets of rules that explained much of the significant data for these classes. The time for manual rule formation for these data was estimated to be several months.

Progress was made on generalizing the Meta-DENDRAL program, and rules for a new domain were successfully discovered by the program. A scientific paper on this application was submitted for publication [Report HPP-77-4]. The new application was learning rules for interpreting signals from C13-NMR spectroscopy. The instrument produces data points in a bar graph in response to the resonance of each carbon-13 nucleus in the sample. The rules describe an environment of a C13 atom and predict a resonating frequency range for every atom that matches the description. The Meta-DENDRAL program needed some modification because the rules are predicting ranges of data points, and not precise processes, as for the mass spectrometry version.

The RULEGEN component of Meta-DENDRAL was demonstrated to work with its heuristic search paradigm. Guidance from a model of mass spectrometry is an important feature of RULEGEN. Also, the program uses problem data for pruning possible rules (and all more specific rules formed from those). The amount of data examined during the search is very large and the space of rules is immense, so the search needs to be rather coarse in order to produce plausible, but not necessarily optimal, rules.

The RULEMOD program for "fine-tuning" Meta-DENDRAL's newly-discovered rules was finished. This program provides a number of important subtasks, including merging similar rules, making rules more specific or more general, and filtering out the weakest rules. RULEMOD checks for counterexamples to rules and uses this information in all of the named tasks. Because of the expense of computing counterexamples to possible rules, this computation is delayed until Meta-DENDRAL has a set of plausible rules, rather than computing counterexamples on each possible rule examined in the search of the rule space.

A report was written on the AI methodology underlying Meta-DENDRAL. The major idea developed in this report is that knowledge of the domain can be used effectively to guide a learning program. The major difference between Meta-DENDRAL and statistical learning programs is that Meta-DENDRAL uses a strong model of mass spectrometry, including any assumptions the user cares to make about the domain, to guide the formation of explanatory rules.

9 C13 NMR SPECTROMETRY

¹³C NMR was selected as a new application area for the rule formation program, Meta-DENDRAL. The algorithms used for mass spectrometry rule formation were extended to ¹³C NMR and used to obtain a set of rules for These two classes and acyclic amines. These two classes were chosen since compounds in these classes are known to show a strong correlation between structural environment and shift. Thus, the programs could be tested knowing that the underlying basis for the form of the rule was valid.

The form of the rule is
substructure ---> shift range.

A sample rule generated is
C-C*-C-X- ----> 19.85<= (delta sub C)<=21.3.

The asterisk in the substructure description denotes the atom for which the shift is predicted. Only topological descriptors were used to construct the substructures. The addition of stereochemical terms is a topic of current work.

It was necessary to change RULEGEN so that the left-hand sides of rules were expanded outward from a carbon atom rather than from a bond. The right-hand side of the rule is associated with a range rather than a precise mass as in the mass spectrometry program. This modification also required changes in the rule search procedure. The user sets two parameters which guide the rule search. These parameters are MINIMUM-EXAMPLES which requires each rule to explain a given number of peaks in the training set and MAXIMUM-RANGE which defines the acceptable shift range for a rule. These parameters regulate the degree of specificity or generality of the rules.

From the set of rules generated a subset is selected corresponding to the "best" set which still covers all the training set data. The best rule is selected by calculating

$$(\text{number of peaks predicted}/(\text{range} ** 2)),$$

Data which are predicted by the best rule are removed and the next best rule is found for the remaining data using the criterion given above. This process is repeated until all data are explained.

In order to test the informational content of the rules generated a second program was written which applied the rules to a list of candidate molecules and ranked the molecules. First, all possible structural isomers for a given empirical formula were generated using CONGEN. The rules were applied to each of the possible isomers and spectra were predicted. The predicted spectra were compared to that of a known spectrum from a compound with the same empirical formula. The structural isomers were ranked according a comparison score to determine how well the correct compound was distinguished from its isomers, on the basis of the predictive rules.

The details of the generation of rules and the use of rules for structure selection can be found in a paper recently submitted for publication [Report HPP-77-4]

The ¹³C NMR rule formation program was applied to a set of paraffins and acyclic amines. The program generated 138 rules to cover 435 data peaks. The rules generated were applied in a structure selection test for the structural isomers of C₉H₂₀ and C₆H₁₅N. No structures with these empirical formulas were

included in the training set. Twenty-four C₉H₂₀ and eleven C₆H₁₅N ¹³C NMR spectra were available to act as unknowns in the structure selection test. The results of the structure ranking applied to these spectra are shown below.

EMPIRICAL FORMULA	NUMBER OF CANDIDATE ISOMERS	NUMBER OF CANDIDATES RANKING			
		1st	2nd.....6th.....	9th	
C ₉ H ₂₀	35	20/24	3/24	1/24	
C ₆ H ₁₅ N	39	8/11	2/11	1/11	

The performance of the rules in discriminating among similar structures not included in the training set data demonstrated the content of the rules.

10 BUDGET

Budget Information relevant to future funding was submitted with the renewal proposal to the BRP.

11 RECENT PUBLICATIONS OF THE HEURISTIC PROGRAMMING PROJECT

(Only publications related to computers in chemistry are shown.)

- HPP-76-1 D.H. Smith, J.P. Konopelski and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIX. Computer Generation of Ion Structures", *Organic Mass Spectrometry*, 11: 86, (1976).
- HPP-76-2 Raymond E. Carhart and Dennis H. Smith, "Applications of Artificial Intelligence for Chemical Inference XX. Intelligent Use of Constraints in Computer-Assisted Structure Elucidation", *Computers In Chemistry* (in press).
- HPP-76-3 C.J. Cheer, D.H. Smith, C. Djerassi B. Tursch, J.C. Braekman and D. Dalozé, "Applications of Artificial Intelligence for Chemical Inference XXI. Chemical Studies of Marine Interbrates - XVII. The Computer-Assisted Identification of [+]-Palustrol in the Marine Organism *Cespitularia* sp., aff. *subviridis*", *Tetrahedron*, 32:1807, Pergamon Press, (1976).
- HPP-76-4 B.G. Buchanan, D.H. Smith, W.C. White, R.J. Gritter, E.A. Feigenbaum, J. Lederberg, and Carl Djerassi, "Application of Artificial Intelligence for Chemical Inference XXII. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program", *Journal of the American Chemical Society*, 98: 6168 (1976).
- HPP-76-5 T.H. Varkony, R.E. Carhart and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference XXIII. 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., in press.
- HPP-76-6 D.H. Smith and R.E. Carhart "Applications of Artificial Intelligence for Chemical Inference XXIV. Structural Isomerism of Mono and Sesquiterpenoid Skeletons 1,2-", *Tetrahedron*, 32:2513, Pergamon Press (May 1976).
- HPP-76-10 Bruce G. Buchanan and Dennis Smith, "Computer Assisted Chemical Reasoning", in *Proceedings of the III International Conference on Computers in Chemical Research, Education and Technology*, Plenum Publishing, (1976).

- HPP-77-4 T.M. Mitchell and G.M. Schwenzer, "Applications of Artificial Intelligence for Chemical Inference. XXV. A Computer Program For Automated Empirical ^{13}C NMR Rule Formation", (Submitted to JACS, January 1977).
- HPP-77-6 STAN-CS-77-597 Bruce G. Buchanan and Tom Mitchell. "Model-Directed Learning of Production Rules", Submitted to the Proceedings for the Workshop on Pattern-Directed Inference Systems in Hawaii, (February, 1977).
- HPP-77-11 Dennis H. Smith and Raymond E. Carhart, "Structure Elucidation Based on Computer Analysis of High and Low Resolution Mass Spectral Data". Proceedings of the Symposium on Chemical Applications of High Performance Spectrometry. University of Nebraska, Lincoln, (in press).