Applications of REACT to Structure Elucidation Problems

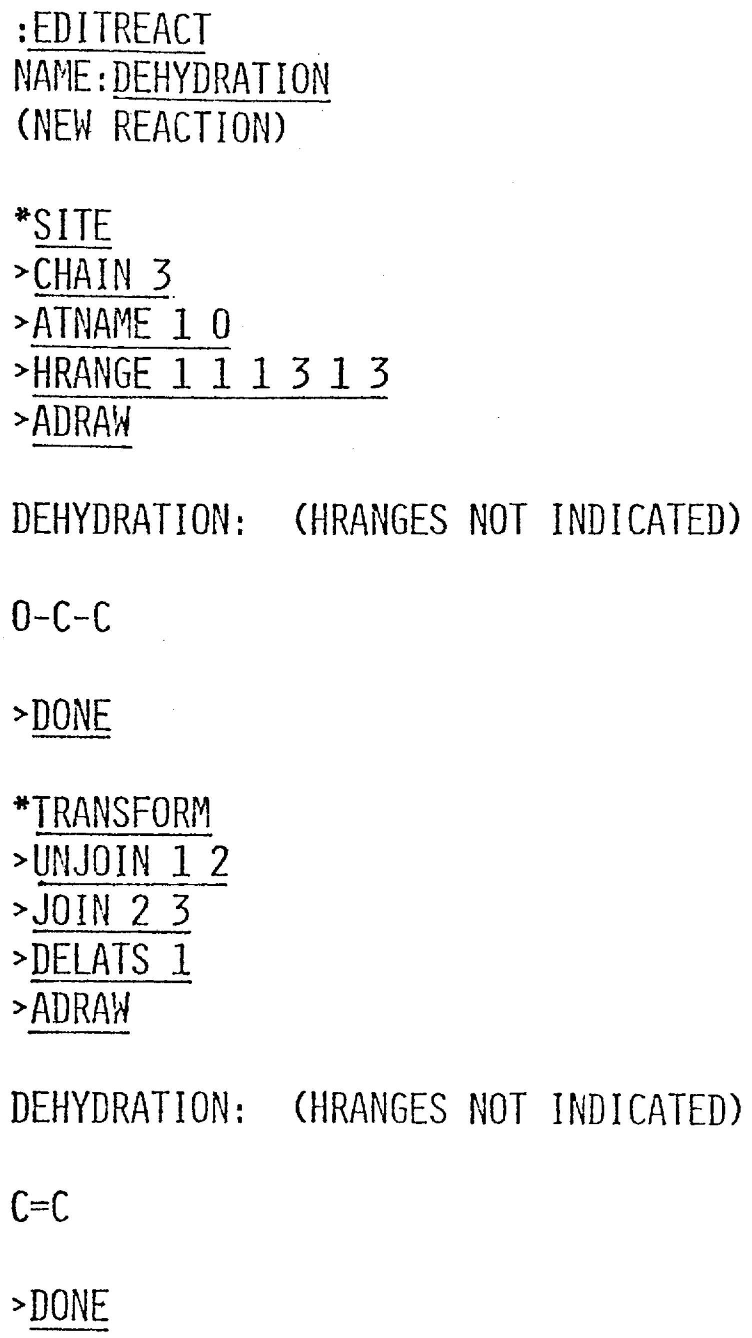
We have recently described our initial efforts toward representation of chemical reactions and their use in structure elucidation problems [Report HPP-76-5]. These efforts provided the framework for carrying out reactions within the computer which emulate actual laboratory reactions performed on a unknown. 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.

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.

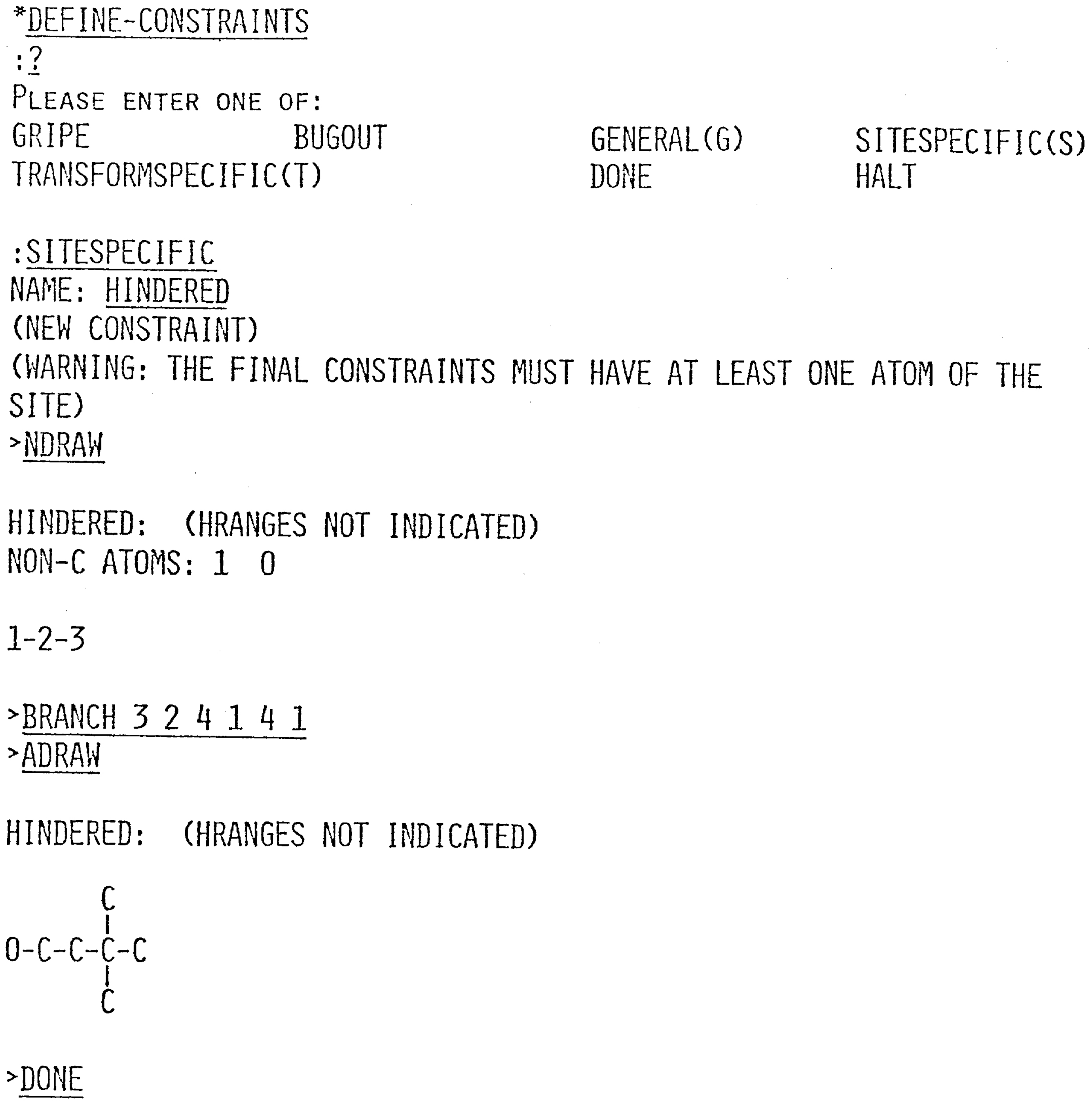
The transform is defined as cleavage and loss of the oxygen resulting in formation of a double bond between the two carbon atoms of the original site (Fig. 7). In this particular dehydration the chemist wished to specify a site-specific constraint. It was known that a tertiary butyl group was part of 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).

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.



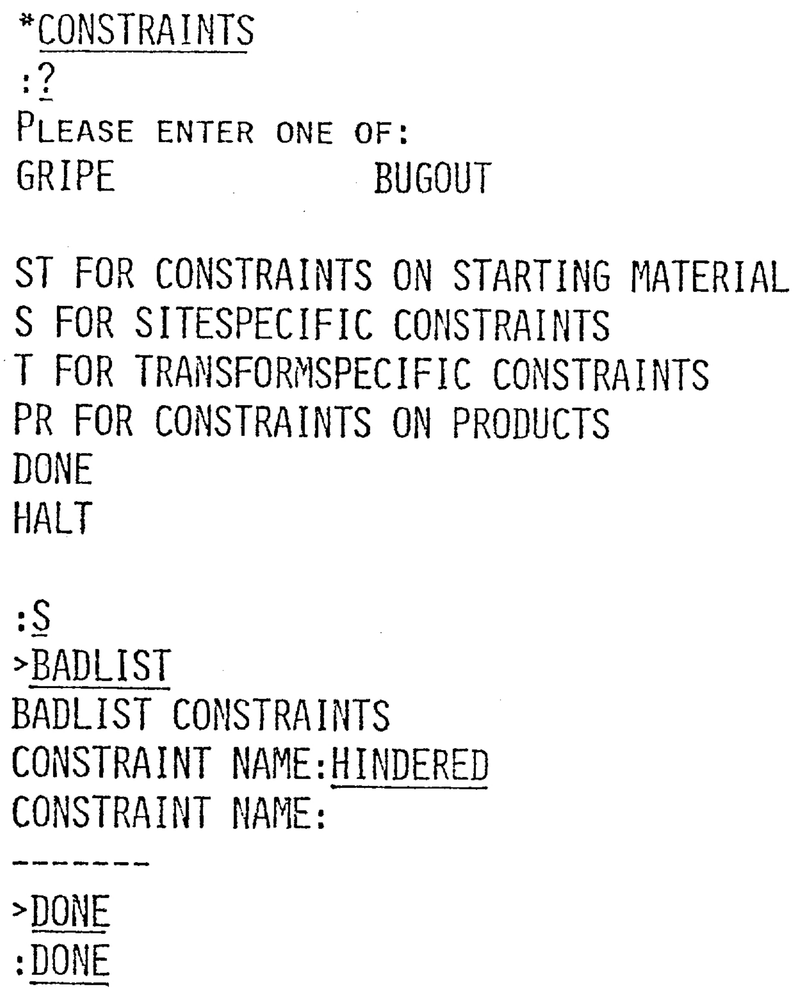
```
:EDITREACT
NAME:DEHYDRATION
(NEW REACTION)

*SITE
>CHAIN 3
>ATNAME 1 O
>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.



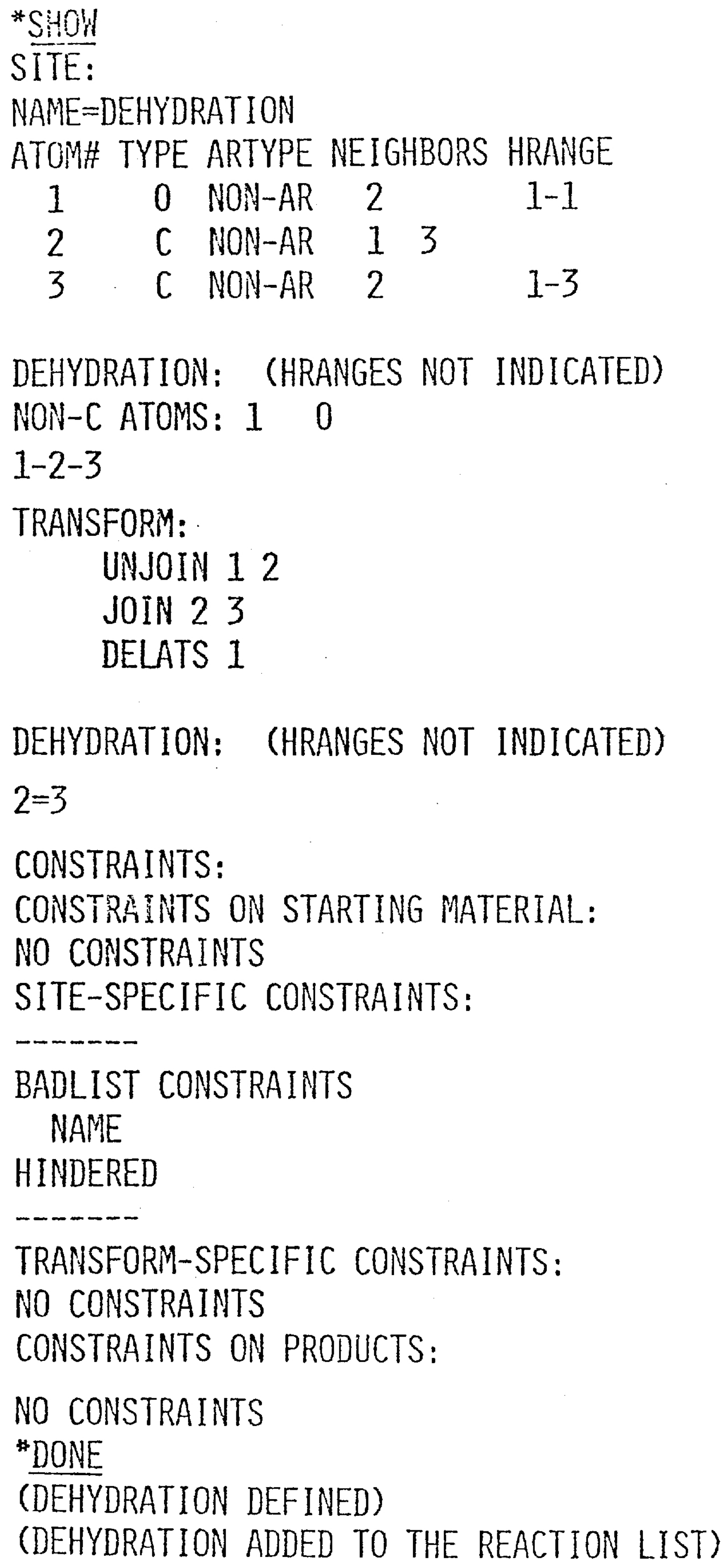
```
*DEFINE-CONSTRAINTS
:?
PLEASE ENTER ONE OF:
GRIPE             BUGOUT              GENERAL(G)       SITESPECIFIC(S)
TRANSFORMSPECIFIC(T)                  DONE             HALT

: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  O

1-2-3

>BRANCH 3 2 4 1 4 1
>ADRAW

HINDERED:  (HRANGES NOT INDICATED)

      C
      I
O-C-C-C-C
      I
      C

>DONE
```

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



```
*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     0  NON-AR   2         1-1
  2     C  NON-AR   1  3
  3     C  NON-AR   2         1-3

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.

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.

1) Intelligent use of substructural information as constraints

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 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.

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.

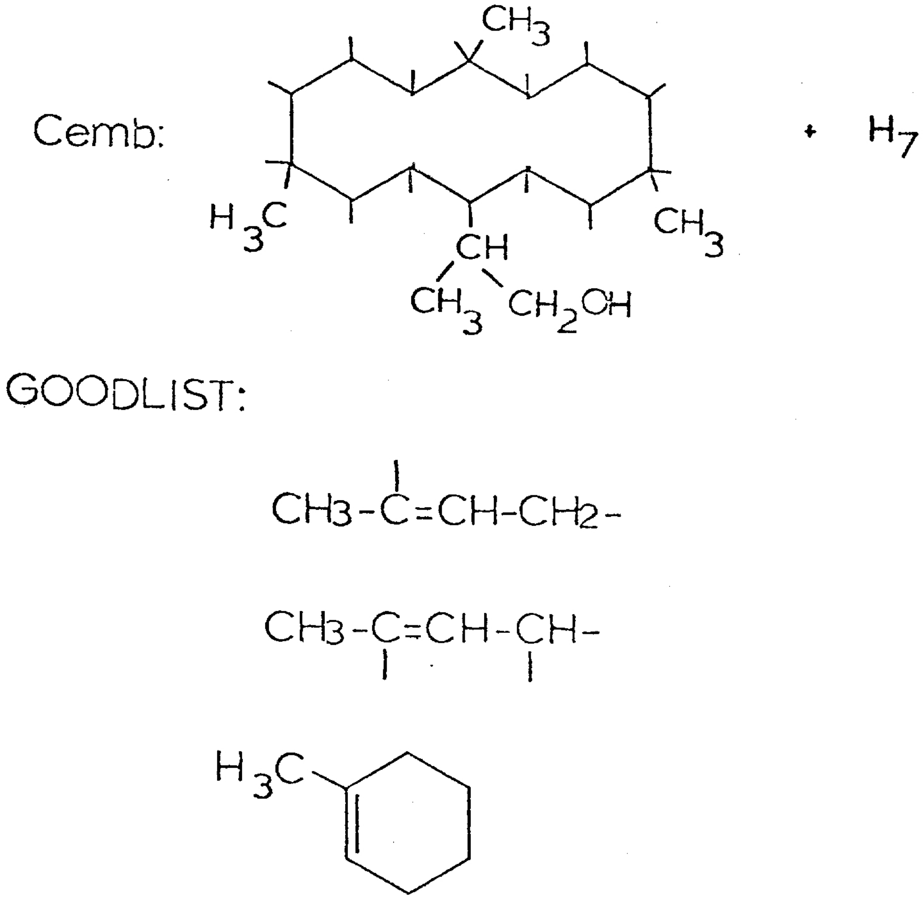


Figure 11. Example of breaking one GOODLIST substructure into several subproblems for CONGEN, each with different superatoms.

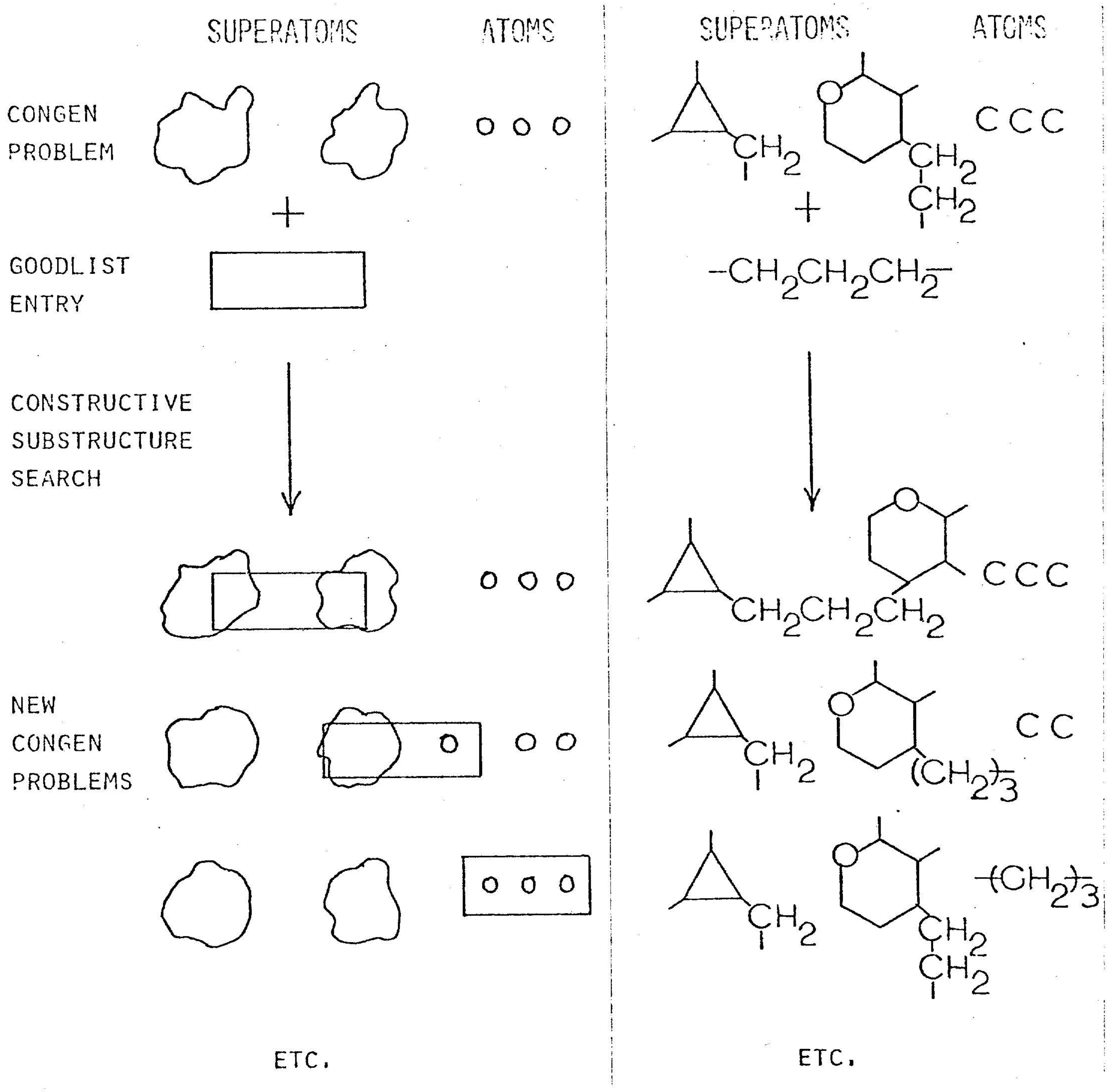


Figure 12. Example showing the inefficiency of specifying a constraint as a GOODLIST item instead of analyzing its implications for constructing allowable chemical graphs.

2) New tools for post-pruning CONGEN structures.

From an algorithmic standpoint, CONGEN is successful if it can, in a reasonable amount of 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.

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 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 exchangable 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 exchangable 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.

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.

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/comparitor 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.

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.

The GC/HRMS DATA SYSTEM

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 controled 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.

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.

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

(number of peaks predicted/(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. Firsts, 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 ^{13}C 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 C_9H_{20} and $C_6H_{15}N$. No structures with these empirical formulas were included in the training set. Twenty-four C_9H_{20} and eleven $C_6H_{15}N$ ^{13}C 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_9H_{20}	35	20/24	3/24		1/24
$C_6H_{15}N$	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.

FUNDING STATUS

Renewal of funding for three years was just received for NIH Grant RR-00612 from the Biotechnology Resources Program (May, 1977 - April, 1980). The award for 1977-78 is approximately $193,000. In addition, support for the basic artificial intelligence research on which this work is grounded is provided by the Advanced Research Projects Agency of the Department of Defense (ARPA Contract DAHC-15-73-C-0435). A new two-year contract was just negotiated for the period July, 1977 - June, 1979.

RECENT PUBLICATIONS

(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.

Daloze, "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 13C NMR Rule Formation", (Submitted to JACS, January 1977).

HPP-77-6 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). (STAN-CS-77-597)

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).

II. INTERACTION WITH THE SUMEX-AIM RESOURCE

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:

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.

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).

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.

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 santolinate 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 pheremones (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. VanTamelen, 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.

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 of determining 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.

INTERACTION WITH OTHER SUMEX-AIM PROJECTS

We have had numerous discussions with Prof. Todd Wipke's research group in meetings of our combined groups. Because the problems of manipulating chemical graphs are much the same for both groups, frequent discussions are mutually advantageous.

Almost daily contact with other Stanford-based projects provides new ideas and programming assistance. In particular, there is considerable interaction with members of the MYCIN, MOLGEN and Protein Crystallography projects. Many of our experiment planning ideas have come from discussions with the MOLGEN group. Our ideas about explaining a program's reasoning are derived from the success of MYCIN's explanation package. And our ideas about integrating multiple sources of knowledge in data interpretation have been enhanced through discussions with the Protein Crystallography group. The large number of excellent INTERLISP programmers in all these groups provides a pool of programming expertise that we draw on frequently also.

We are collaborating with Dr. Robert Lindsay on a monograph about the DENDRAL programs, with most of our interaction and all our text preparation taking place over the SUMEX system. We have also discussed helping Dr. Lindsay with a knowledge-based reasoning program to help pathologists at the University of Michigan.

CRITIQUE OF RESOURCE SERVICES

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.

The interactive computing environment provided by the SUMEX-AIM resource and the power of the INTERLISP language give us the capability of building and debugging complex programs rapidly. These are the best tools currently available for AI research. Because these tools are available and they are almost always available on command, our researchers are working at the frontier of applied artificial intelligence. The SUMEX staff does an outstanding job of keeping the computer and peripheral devices running reliably: without this professional support we would not be able to build, enlarge, and test programs as complex as the DENDRAL programs.

The large number of persons who use the resource is our single biggest source of frustration. Several of the DENDRAL programmers work frequently from midnight to 8:00 a.m. just to avoid computing during the day. Although this minimizes their interaction with the rest of the research group, it allows them to work on large, cycle-intensive programs without competing for resources during "prime-time" hours.

III. USE OF SUMEX DURING THE FOLLOW-ON GRANT PERIOD (8/78-7/83)

LONG-RANGE GOALS

Our primary goal is to build reliable, useful tools for biomolecular structure characterization and make them available for widespread use. The CONGEN program is farthest along in this respect. We will extend its scope and add features to make it easier to use, while working on the problems of increasing its availability. By building onto CONGEN we will develop a broader set of tools with capabilities for helping biomedical scientists in many ways. By increasing the generality of Meta-DENDRAL we intend to provide tools for model-directed learning from empirical data that will complement purely statistical tools.

At the same time we are building tools we are also exploring basic AI issues of knowledge representation, use, and acquisition in complex reasoning programs. These are fundamental issues for knowledge-based programs, such as those currently running on SUMEX.

JUSTIFICATION FOR CONTINUED USE OF SUMEX

The research goals and methods of the DENDRAL project fit well within the stated AIM criteria. We are building knowledge-based programs, and extending the art of applying AI to medicine to the benefit of both working biomedical scientists and other groups building similar tools.

We need the SUMEX-AIM resource for our work because of its excellent environment for symbolic computing. The interactive computing facilities and the features of the INTERLISP language on SUMEX give us a several-fold increase in productivity over our previous batch computing environment using LISP-360.