

that the user need not be concerned with any differences between topological and stereochemical constraints.

C.3 Development of Automated Approaches to the Exploitation of Spectroscopic Data.

Research will be undertaken into the development both of methods for deriving structural constraints for GENOA by automated spectral interpretation, and of methods for evaluating generated candidate structures by comparative analysis of predicted and observed spectral properties.

Spectral interpretation, to derive structural constraints, is appropriate only for those techniques in which, at least to a first approximation, spectral features can be correlated with fairly localized structural environments. Some spectrum/structure correlations are relatively weak; these can only be taken as suggestions, and not as absolute constraints. An example of such weak correlations would be any association of a particular group with an IR absorption below 1500cm^{-1} (for that region of the IR spectrum is usually dominated by combination and overtone bands due to vibrations of the entire molecule that render reliable interpretation difficult). Generally, inferences based on such weak spectrum/structure correlations should not be employed as constraints prior to structure generation but can be exploited when ranking structures subsequent to generation. In our proposed work on inferring structural constraints from spectral data, we shall be concentrating mainly on magnetic resonance techniques in which spectrum/substructure correlations are usually well defined.

Once structures have been generated, it is possible to exploit spectral data that relates more to the complete molecular structure than to any isolated subpart. We have already developed functions for predicting mass spectra using models, "theories", of how a given molecular structure might fragment [23,67]. Functions that predict magnetic resonance spectra, using a model of a complete molecular conformation, will in general be more accurate than those based purely on localized substructures, and consequently, will permit finer discriminations to be made between different candidate structures. Other post-testing of generated structures can exploit the suggestive evidence of spectrum/structure correlations that yield evidence for or against particular structural features but are not absolutely definitive tests.

C.3.a Carbon-13 Magnetic Resonance.

The chemical shift of a given carbon atom is sensitive to features of its local environment up to, and sometimes including, delta-neighbors. For some molecules, with rigid conformations, topologically more remote neighbors may also produce significant shifts through steric crowding. Although there is a great deal of information in ^{13}C MR data, frequently structure elucidation studies use the ^{13}C MR results just to determine gross features of the carbons in the structure such as their hybridization, degree of substitution and possible bonds to electronegative atoms. We intend to explore more constructive uses of CMR including both the inference of substructural parts prior to structure generation, and spectrum prediction and evaluation for complete, generated structures.

C.3.a.i Carbon-13 Spectrum Interpretation.

One aspect of the Meta-DENDRAL project in recent years was the development, by Mitchell and Schwenger, of a system for the structural interpretation of ^{13}C MR spectra [16]. The Mitchell/Schwenger system utilized rules that correlated particular resonance values with substructures; each rule involved a main prediction defining a fairly precise range in which a particular carbon in the substructure should resonate and a number of secondary and support predictions defining ranges in which the resonances should occur for the other constituent atoms of the substructure. An example of the type of rule used is shown in Figure 4. The rule is interpreted to mean that if a resonance is observed in the range 44.7-44.9, and if appropriate secondary/support predictions are satisfied, then the given substructure can be taken as a possible explanation for the resonance. These rules were abstracted from a set of spectra of standard alkanes and alkylamines and could be used for the identification of additional alkanes and amines not present in the training set.

| 44.7 ppm < delta(3) < 44.9 ppm => | | | | 1 5-4-3-2-7 8 |
|-----------------------------------|----------|------------------------|------------------------|-------------------------------|
| Node | atomtype | secondary prediction | support prediction | |
| 1 | C | | 27.1 < delta(1) < 34.9 | |
| 2 | C | 29.7 < delta(2) < 35.6 | 30.7 < delta(2) < 33.4 | |
| 3 | CH2 | | | |
| 4 | CH2 | 17.9 < delta(4) < 56.9 | 17.9 < delta(4) < 27.6 | |
| 5 | C | | 15.4 < delta(5) < 24.3 | |
| 7 | C | | 27.1 < delta(7) < 34.9 | |
| 8 | C | | 27.1 < delta(8) < 34.9 | |

Figure 4. Meta-DENDRAL C-13 Spectrum Interpretation Rule

Structure elucidation was accomplished by determining the rules, and consequently the substructures, that could be associated with each individual resonance and then, in a complex search scheme, determining allowed combinations of these separate substructures. The Mitchell/Schwenger system involved a heuristically guided, depth first search in which substructures were combined and expanded by identifying their allowed partial overlaps. The method has been illustrated through its analysis of a fully-decoupled spectrum of 3-3-dimethylhexane [16].

Methods of spectrum interpretation using data bases of ^{13}C MR spectra have been reported by both Bremser and Jezl [68,69,70,71]. These two systems allow for a number of search options; the data base can be searched in a conventional manner to identify reference compounds with spectra similar to that for an unknown, or individual resonances may be entered and the data base searched to retrieve all substructures showing similar shifts, or the data base may be searched to find the range of shifts associated with some given substructure (i.e. spectrum prediction). Bremser has illustrated how results of a standard file search may be interpreted

manually to yield information on substructural components of a compound not in fact present in his reference file.

In these two systems, the data characterizing each reference compound consists of atom-centered codes, (describing the topological environment of each carbon with an assigned shift), and associated shifts. Originally, both systems defined atom-centered codes that represented a tree-structure grown through the molecule from the central atom; tags were used in the Jezl system to indicate ring-membership and similar properties. These codes were ambiguous in that quite distinct, cyclic sub-structures could yield the same code. The Jezl scheme also suffered from the disadvantage that the codes were derived through relatively complex rules, handling many special cases, and the coding process could not be automated. The Bremser "HOSE" code was more simply defined as a Hierarchically Ordered description of an atom's Spherical Environment; generation of the code could be completely automatic. Bremser's original code had problems of ambiguity similar to those of the Jezl code; Bremser has subsequently modified his coding scheme but problems, such as truncation after a certain number of characters, remain.

We have started to develop a system for ^{13}C MR spectrum interpretation that is an extension of the second-type of search option in the Bremser and Jezl schemes. Our program takes, as input, the individual resonances of all carbon atoms in the unknown structure and searches a reference file to identify all atom-centered codes associated with similar resonances. Unlike the Jezl and Bremser systems, these retrieved codes are not the program's output, but are in fact the input to a more detailed stage of analysis in which the results for individual resonances are combined to yield much larger substructures. In essence, the subsequent processing steps correspond to the structure building procedures of the Mitchell/Schwenzer system; however, in this context it is possible to avoid some of their complexities relating to graph matching of possibly partially overlapping structures.

The current implementation of these algorithms is incomplete. We have systems for generating atom-centered codes and creating reference libraries of codes and shifts, and we have preliminary versions of functions for exploiting such reference libraries. The atom-centered code that we employ is a complete, canonical description of an atom's topological environment (out to a distance of three bonds). The coding system will be extended to include stereochemical information once such data has been incorporated into our standard structure definitions.

The following example is derived from data on isoterpinolene Figure 5:

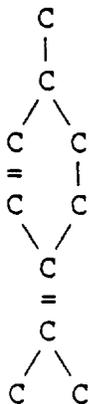


Figure 5. Isoterpinolene — example structure for C-13 interpretation functions

The program requires the shift, multiplicity and allowed error in shift for every carbon in the structure.

```

| C-resonance : 21.4 4 0.5
| C-resonance : 30.6 2 0.5
| C-resonance : 133.1 2 0.5
| C-resonance : 125.0 2 0.5
| C-resonance : 127.7 1 1.0
| C-resonance : 125.7 1 1.0
| C-resonance : 25.7 3 1.0
| C-resonance : 31.7 3 1.0
| C-resonance : 20.5 4 0.5
| C-resonance : 19.6 4 0.5
| Cl3 reference file : crsnrc

```

Figure 6. Spectral data input for Isoterpinolene

The program searches the file of substructures and associated shifts, retrieving those substructures with appropriate shifts and multiplicities. In this first pass through the reference file, each shift is considered individually.

```

| 3565 references read, 280 references written.
|
| allowed atom codes are :
|   -CH3  -CH2- >CH-  -CH=  >C=   -C*H=  >C*=
|
| LINE TYPE
| 1  -CH3
| 2  >CH-
| 3  -CH=  -C*H=
| 4  -CH=  -C*H=
| 5  >C=   >C*=
| 6  >C=   >C*=
| 7  -CH2-
| 8  -CH2-
| 9  -CH3
| 10 -CH3

```

(The symbol C* indicates an aromatic atom type). In a subsequent checks, through the reduced reference file, the program utilizes existing partial results regarding possible alpha-neighbors for each atom. Such checks are able to derive the following information:

allowed atom codes are :
 -CH3 -CH2- >CH- -CH= >C=

LINE TYPE

1 -CH3
 2 >CH-
 3 -CH=
 4 -CH=
 5 >C=
 6 >C=
 7 -CH2-
 8 -CH2-
 9 -CH3
 10 -CH3

Current partial adjacency matrix:

| | Hetero | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|--------|---|---|---|---|---|---|---|---|---|----|
| 1 | . | . | ? | . | . | ? | ? | . | . | . | . |
| 2 | . | ? | . | * | . | . | . | ? | ? | ? | ? |
| 3 | . | . | * | . | ? | ? | ? | . | . | . | . |
| 4 | . | . | . | ? | . | ? | ? | ? | ? | . | . |
| 5 | . | ? | . | ? | ? | . | * | ? | ? | ? | ? |
| 6 | . | ? | . | ? | ? | * | . | . | . | ? | ? |
| 7 | . | . | ? | . | ? | ? | . | . | ? | . | . |
| 8 | . | . | ? | . | ? | ? | . | ? | . | . | . |
| 9 | . | . | ? | . | . | ? | ? | . | . | . | . |
| 10 | . | . | ? | . | . | ? | ? | . | . | . | . |

Figure 7. Adjacency matrix derived for Isoterpinolene using Alpha-neighbors

(In the adjacency matrix, the "."s indicate no bond possible, "?"s indicate possibility of a bond and "*"s show where the program finds bonds to have been definitely established). The program has been able to determine that the none of the methyls can bond to the methylene groups, has established that the sole methine carbon must bond to one of the alkene units and identified the other alkene part. These restrictions could be input to GENOA (see Section C.2) and would considerably reduce the size of the structure generation problem. However, in this example, it was possible to derive further constraints by utilizing data on beta-neighbors provided by the retrieved codes. Using beta-neighbors leads to the following final connectivity matrix of Figure 8. The only uncertainty remaining concerns which of the methyl groups is attached to the methine and which are substituted onto the alkene. Because the methyl groups exhibit different shifts, they are distinct within this program. As far as they are defined, these results do correspond to the correct structure.

Current partial adjacency matrix:

| | Hetero | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|--------|---|---|---|---|---|---|---|---|---|----|
| 1 | . | . | ? | . | . | . | ? | . | . | . | . |
| 2 | . | ? | . | * | . | . | . | . | * | ? | ? |
| 3 | . | . | * | . | * | . | . | . | . | . | . |
| 4 | . | . | . | * | . | * | . | . | . | . | . |
| 5 | . | . | . | . | * | . | * | * | . | . | . |
| 6 | . | ? | . | . | . | * | . | . | . | ? | ? |
| 7 | . | . | . | . | . | * | . | . | * | . | . |
| 8 | . | . | * | . | . | . | . | * | . | . | . |
| 9 | . | . | ? | . | . | . | ? | . | . | . | . |
| 10 | . | . | ? | . | . | . | ? | . | . | . | . |

Figure 8. Adjacency matrix derived for Isoterpinolene using Beta-neighbors

The current program has several limitations (apart from the fact that the structure codes are purely topological with no stereochemical data). The major limitation is in the reference file which is derived from data on rather less than 500 compounds (standard alkanes, alkenes, terpenes, diterpenes and a few other natural products such as coumarins). Very frequently, the file does not contain any substructure equivalent, out to beta neighbors, to some feature in a new unknown compound. In such cases, attempts to utilize beta-neighbors will lead to a situation where no complete structure may be generated. However, matching alpha-neighbors does still provide a few usable constraints. We propose to obtain from other sources computer readable files of extensive collections of ^{13}C MR spectra. One such file is available through the NIH/EPA Chemical Information System. However, although most of the spectra are assigned, structures are not associated with the spectra. We are exploring ways to obtain the structures (in computer readable form) and integrate them with the spectra themselves. Bremser [70] has available a large collection of assigned spectra with associated structures. The charge is, however, DM 1.75 per spectrum (approximately 14,000 spectra are available). If we are unable to obtain the required data from the NIH/EPA system we will be forced to submit a supplementary application for funds to purchase Bremser's collection.

The current program is not interfaced to the rest of the CONGEN system, and some of the constraints that it can identify cannot be expressed in terms of the substructures that may be input to the current GENOA program. We propose to build those interfaces and to develop ways to translate constraints derived from the ^{13}C analysis into a form usable by GENOA. This work will initially relate only to the topological representations of structure. We will at the same time, however, begin a design of the required stereochemical extensions to both the interpretive aspects of the program and the interfaces to the structure generators.

C.3.a.ii Prediction of C-13 Spectra.

Several approaches to the computerized prediction of ^{13}C MR data have been reported. The CARBON-13 program of Munk, et al. [39] is concerned solely with the prediction of the number of distinct resonances that might be expected in the ^{13}C MR spectrum. Their prediction method is based on identifying topologically equivalent atoms with a few, ad hoc, extensions to identify anisochronous methyl groups in those diastereotopically related geminal methyls which are a commonly encountered feature of organic molecules. Our algorithms for determining the true stereochemical symmetry group would permit the correct enumeration of such resonance counts.

Other programs reported in the literature have been concerned with the

prediction of approximate resonance shifts for each carbon nucleus. Some attempts have been made to utilize additivity rules for ^{13}C MR spectrum prediction. Thus, Clerc has a program for estimating ^{13}C MR shifts of singly bonded carbons in acyclic structures containing combinations of non-cyclic functional groups [72]. Predictions for more complex cyclized structures generally require parameterized schemes devised especially for sets of related structures. While topological descriptors may suffice for certain applications, geometrical descriptors are also of value. Some advances in the computerized derivation of ^{13}C MR prediction functions, based on both topological and geometrical descriptors, have recently been reported [73].

Spectrum/substructure correlation rules of course provide another method of spectrum prediction. The spectrum prediction system of Mitchell [15] employs a slightly unusual representation of such a correlation table. In this system, 138 rules of the form "substructure \rightarrow shift range" were abstracted from a set of alkane and amine spectra. Spectrum prediction was achieved by graph matching each substructure to the given structure, and assigning to each carbon a shift range representing the common range of shifts predicted by each applicable rule. The interactive NIH ^{13}C MR system [74] can work in a somewhat similar manner; it differs in that the range of shifts, and details of their distribution, are derived from data base as required, rather than being abstracted into a "prediction rule". The substructure of interest is defined and then the program retrieves the shifts of all instances of that substructure in the current data base. (This has the advantage that expansion of the data base does not require re-analysis to derive new prediction rules). The search of the potentially large data base is made efficient by the use of fragment codes, bit screens etc; the final steps do, like Mitchell's system, require some graph matching.

Bremser has noted the use of his library of atom-centered codes and related shifts for spectral prediction; a paper with more details is due to be published. Essentially the same Mitchell/NIH correlation methods are used; atom-centered codes are generated for each atom in the given structure and the data base is searched to retrieve shifts associated with these codes. The graph-matching processes are here replaced by the string comparisons on the atom-centered codes. The Jezl/Dalrymple RACES system may be used for spectrum prediction in exactly similar fashion.

Currently, we have a similar spectrum prediction capability. We are choosing to concentrate at present on deriving substructural constraints from these spectrum/structure correlations, rather than explore spectrum prediction and ranking schemes. If the methods of interpreting ^{13}C MR data in terms of substructures are successful (and used to derive constraints for a structure generator), then spectrum prediction based on this type of model would not serve as a basis for further discrimination between structures. We will consider the use of more elaborate approaches to spectrum prediction that might be applicable once the CONGEN programs have been extended to provide complete conformational, stereochemical representations of structures. The conformations could be used as input to a force-field structure modeler and the resulting configurations analyzed to derive both topological and steric contributions to chemical shifts (as in the work of Smith and Jurs noted earlier).

C.3.b Proton magnetic resonance.

The chemical shift of a proton is, like a carbon nucleus, determined largely by features of its local environment. However, the factors determining the shift tend to be somewhat shorter range than those for ^{13}C MR. Not all influences on a resonance can be correlated with the immediate topological neighbors;

steric/conformation effects can result in protons experiencing strong shielding/deshielding due to topologically remote, but sterically close, pi-systems. Consequently, as with ^{13}C MR, much preliminary information can be derived, prior to structure generation, by interpreting ^1H MR in terms of local environments but sensitive discrimination between related structures may require spectrum prediction based on accurate models of molecular conformations.

Some attempts have been made to derive structural information from the coupling pattern exhibited in a proton spectrum either independently of [75], or in association with chemical shift data [76]. These methods assume well resolved, first order spectra. It is rare for compounds of bio chemical interest to show such simple spectral characteristics, and these interpretation methods are mainly of educational rather than practical interest.

Sasaki [41] uses conventional correlation chart approaches to derive structural data from the ^1H MR spectrum. He employs a table of about 200 standard fragments, each defined by the number of protons that should resonate in particular spectral regions. Analysis of a spectrum, by referencing this table, identifies maxima and minima for the numbers of groups of different types. For moderately large structures, the results from this type of ^1H MR interpretation are generally somewhat ambiguous; usually, many different combinations of substructures can be found that would satisfy these weak spectral constraints. However, even this simple interpretive process does capture the kind of negative evidence that chemists frequently fail to provide to CONGEN and other structure generating programs; the absence of required resonances will lead to the prohibition of the generation of certain substructures — a prohibition that the chemist will typically only apply after seeing the generation of invalid structures.

We have made some preliminary experiments, within the INTERLISP version of CONGEN, on the use of additivity rules for the prediction of proton resonance spectra. This ^1H MR system consisted of functions for predicting the resonance values of protons attached to alkyl and alkene carbons, and functions allowing the user to define constraints on the number and type of protons in particular regions of the spectrum. The proton shifts were predicted by conventional additivity formulae using incremental shifts due to alpha-functional groups with some corrections for beta-groups and membership of small rings. The performance of this spectrum-prediction/structure-pruning system was erratic when applied to typical CONGEN problems. In some cases, more than 80% of generated candidate structures could be successfully eliminated using relatively weak constraints. More frequently, within the accuracy of the model, there were no significant differences in the spectra predicted for different candidate structures and pruning of the structure list was not possible. In a few cases, there were sufficiently large differences between the observed and predicted spectra of the true structure that pruning on the basis of what might have been assumed to be reasonable constraints would in fact have eliminated the correct structure. Prediction performance appeared to be poorest for protons on di- and tri-substituted alkenes; but, the scope and limitations of the additivity rules could not be clearly defined. Consequently, this proton magnetic resonance system was not made routinely available to CONGEN users.

Our current intention is to apply to proton NMR the same methods being investigated for ^{13}C MR. Priority is being accorded to ^{13}C MR in part because of the lack of suitable, machine-readable collections of assigned proton spectra.

C.3.c IR.

Spectral-structure correlations based on infra-red data have been exploited

in a number of other structure elucidation programs. Thus, both Sasaki [41,42] and Gribov [50] employ what are essentially correlation tables, while Munk has a system for inferring substructures by either "Pattern Recognition" or "Artificial Intelligence" IR-classification functions [37].

However, these IR interpretation schemes are limited. For example, Sasaki's system attempts merely to determine minima and maxima constraints on the number of oxygens in carbonyl, hydroxy and ether groups. Other groups with absorptions above 1500cm^{-1} might be characterizable by similar correlation rules. However, reliable identification is frequently not possible. Problems of interpretation result from extreme variability in intensity of "characteristic" absorption bands (e.g. the absence of a nitrile absorption in many alpha-hydroxy-nitriles), and from the fact that many of these absorptions lie in the same $1500\text{--}1700\text{cm}^{-1}$ region. Spectra below 1500cm^{-1} are frequently dominated by combination and overtone bands characteristic of the molecule as a whole rather than any isolated subparts; while such spectral data are valuable in file search oriented methods of identifying structures, they are of limited diagnostic value offering suggestive rather than definitive evidence. The same qualification — suggestive rather than definitive evidence — has to be applied to attempts to derive additional structural information from the particular position of absorption of a better characterized group such as a carbonyl. A carbonyl absorption around 1770cm^{-1} is suggestive of a five-membered lactone ring, but presumption of such a substructure, and its use as a constraint on a structure generator, would be ill-advised (similar absorption also characterize vinyl esters and several other substructures).

Since the identification of principal functionality has normally been completed long before a structure-elucidation problem is given to CONGEN, there appears to be relatively limited use for this type of IR-interpretive scheme in association with CONGEN/GENOA.

Gribov has worked on the evaluation of candidate structures using predicted IR spectra. The spectrum prediction is based on conventional approaches for analyzing the vibrational frequencies of a molecule using parameterized bond strengths etc. As noted by Gribov, this approach is really only applicable to fairly small molecules. Conceivably, such an approach might be tried with the larger molecules analyzed by CONGEN.

The only use for IR currently envisaged is in a very simple ranking scheme. In the LISP version of CONGEN, functions were available that allowed the user to associate positive and negative scores with particular substructures and to let the program rank candidate molecules by combining scores associated with their constituent substructures. It is possible to implement a scheme that would derive scores for different substructures through the suggestive, but not definitive, IR-interpretation rules noted earlier. This would allow the association of plausibility values for each structure which might then be used with the simple ranking functions.

C.3.d Mass Spectrometry.

Interpretive processing of mass spectral data by means of spectrum/structure correlations is of limited utility. These correlations, e.g. ($M-44 \Leftrightarrow$ anhydride), generally only provide weak evidence for, or against, the presence of functional groups more readily identified by other spectral/chemical techniques. This type of mass spectral processing is available as a minor component in Gribov's system, and may yet be incorporated in Sasaki's programs. However, it is not considered to be of value in the type of structure elucidation problem generally given to CONGEN.

The Mass Distribution Graph method does constitute a more general model for mass spectrum interpretation [67]. However, the applicability of the method to large structures is limited by the combinatorial nature of the algorithms. We have no plans currently to develop further this approach; in spite of the elegance of its conceptual base, a practical, useful program seems extremely difficult to develop.

Some of the more sophisticated mass spectral file search systems have the capability of identifying constituent substructures of molecules not actually present in the reference file [28]. McLafferty's group is exploring the possibility of using CONGEN in association with their STIRS file-based mass spectral interpretation system.

While methods for interpretation of mass spectra are still of limited applicability, our algorithms for mass spectrum prediction and structure ranking now seem well proven [67,23]. For mass spectral predictions, the use of a topological model of a chemical structure, as produced by current CONGEN, is generally quite satisfactory. The mass spectral processing algorithms can employ models, "theories", of fragmentation processes of varying degrees of specificity as may be appropriate to a particular application.

C.3.e Circular Dichroism and Magnetic Circular Dichroism

Circular Dichroism (CD) and Magnetic Circular Dichroism (MCD) are spectral techniques particularly useful in structural/stereochemical studies. Because of their strong dependence on stereochemistry, these techniques have heretofore not found much use in CONGEN and this represents a serious deficiency. With the addition of conformation information to CONGEN it will now be possible to incorporate structural and stereochemical information from these sources into computer-assisted structure elucidation using CONGEN and GENOA.

The interpretation and prediction of CD spectra based on substructural hypotheses is well established. Examples are the familiar octant rule originally derived for ketones [77], Brewster's rules for paraffinic hydrocarbons [78], and extensions to sector rules such as those of Kirk and Klyne [79]. In all cases the substructure considered must include both configurational and conformational information. The general method is to associate with each substructure an intensity (either positive or negative) which contributes to the overall observed intensity of the spectrum. The assumption is often made of additivity, but not always. Interpretation of CD spectra makes primary use of the observed sign and intensity of the spectrum. First of all, the fact that a CD spectrum was observed at all says that the molecule is chiral, which is a very useful constraint for CONGEN's structure generation. Rationalization of the observed intensity is best made with respect to possible structures, i.e., after a CONGEN generation of possible chiral structures and their conformations. The observed spectrum will not usually lead to structural hypotheses by itself since the (usually single) intensity can be caused by any number of substituents around the chromophore involved. The real use of the method is to rule out possible structures with specified conformation.

Useful, but somewhat coarse, predictions of CD and MCD spectra can be made with the use of a number of models proposed which correlate spectral intensity with substituents. Among these are the octant rule [80] and the "zig-zag" model which are most useful for the CD spectra of saturated ketones. A simple model has also been proposed for MCD spectra of saturated ketones [81]. These models generally are used for predicting spectra of structures having idealized geometries with substituents in nearby or well-defined locations with respect to the chromophore. These methods can therefore be applied to structures output from the proposed

conformation generator. We propose to develop a program which will allow us to discriminate between candidate conformations using these models and experimentally obtained CD or MCD spectra.

A more precise prediction of CD spectra requires a refined geometry and a more detailed energy calculation. If a flexible structure is in an equilibrium between two or more conformations, relative energies are needed to determine populations before computation of spectra. To make use of these methods, an interface between the output of CONGEN and the input to these already existing energy programs will be required. This work will be the second stage of this effort and will be guided by results of the first stage using the simpler models.

While CD spectroscopy in the ultraviolet region is long established as a useful tool for structure elucidation, recent work on MCD (Magnetic Circular Dichroism), VUVCD (Vacuum Ultraviolet Circular Dichroism), and VCD (Vibrational Circular Dichroism) indicate considerable promise for structure elucidation. While the methods of interpretation for these spectra differ, the end result is usually a correlation of substructure (including stereochemistry) with observed intensity and sign. We feel our proposed system for CD spectra will incorporate, with modest modification, substructural information from these newer methods as they become more common in biomedical structure elucidation.

C.3.f Combined Spectral Interpretation.

As we mentioned previously, structural information from different physical techniques is often complementary. Any program seeking to perform automated analysis of data from different techniques must eventually be capable of examining the data and resulting inferences from each technique in light of what has been learned from other techniques. We illustrated a very simple example of this approach in the example on interpretation of ¹³CMR presented above. The assignment of specific substructures to observed resonances implies additional assignments, which can be made on the basis of further data analysis or logically, to cite the trivial example that assignment of one carbon to a C=C functionality means that another carbon must also be assigned to the double bond. We propose to generalize this method, perhaps using an adjacency matrix representation for the growing structure, with a program which acquires inferences from each technique and automatically searches data from other techniques for confirmation or denial of the inferred substructure. This approach will be complicated by the fact that some of the inferences will be tentative. We will evaluate schemes for assigning certainty factors to each inference to explore the possibility of generating structures with associated plausibilities (determined by appropriate combination of the certainty factors for each component part).

There is another useful application of methods for considering the collection of available data. As noted earlier, chemists using CONGEN frequently fail to specify all the structural constraints that, in fact, they know. For the most part, it is negative constraints that are forgotten. Because such constraints are neglected, some problems appear to be too large and many others are solved extremely inefficiently with the chemist generating and then pruning away large classes of structures. If some convenient method of entering spectral data can be devised, then relatively unsophisticated spectral interpretation techniques could be used to derive such negative constraints, effectively prohibiting the generation of particular substructures.

C.4 Conformation Generator for CONGEN.

A great deal of the information input to a structure elucidation problem is conformational in nature. Examples of such information are vicinal and long range couplings in NMR, steric shifts in ^{13}C MR such as the differentiation between axial and equatorial substituents, Circular Dichroism spectra, etc. At present, this information can be used indirectly in CONGEN only as it pertains to the constitution (bond-connectivity) or configuration (chiral centers and double bonds) of a proposed structure.

To eliminate this deficiency and enhance the value and scope of CONGEN, we propose to provide CONGEN with an exhaustive and irredundant generator of conformations based on a chosen discrete set of possible torsional states (i.e., positions of rotation) around rotatable bonds. The input to such a generator would be a CONGEN stereoisomer (specified constitution and configuration) and the chosen set of torsional states (which might be a default set contained in the program). The output would be the possible conformations based on these possible torsional states. Only torsional angles will be considered as variables for the conformation generation, the bond lengths and bond angles will be considered fixed.

While the conformation generator will indeed be exhaustive and irredundant, special attention will be given to common conformational situations such as six-membered rings. This will be done to improve the efficiency of the program for common problems while retaining the desired assurance for more complex problems (see Sec. C.4.b.i).

A conformation generator will provide the necessary link from CONGEN to existing computer methods which require input structures with coordinates. This link is shown schematically in Figure 9. Examples of such methods are molecular mechanics energy calculations, quantum mechanical energy calculations, graphics display programs, and finer grid (torsional angles) conformation generation. A conformation generator would also allow CONGEN to be applied in conformational analysis, another form of structure elucidation. We propose to develop this link to existing methods and applications.

We feel the proposed investment of resources into developing this conformation generator will be repaid many times over because:

- 1) The added versatility it will provide CONGEN to deal with all facets of conformational information in structure elucidation;

- 2) It will allow development of links to existing atomic coordinate based methods which permit many potential new biomedical applications. (See Sections C.4.C, F).

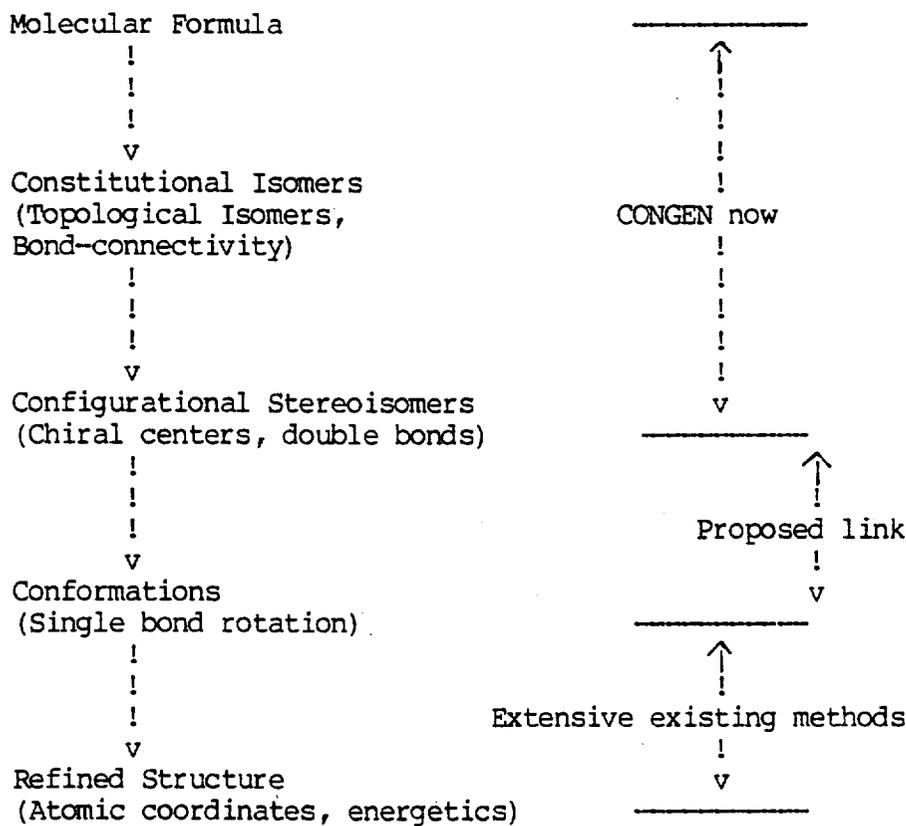


Figure 9. Proposed Link of CONGEN to existing coordinate-based methods

C.4.a Total Conformation Generation.

In its current state of development, CONGEN will generate, if desired, all possible configurational stereoisomers from an input molecular formula (Figure 9). This generation is complete and irredundant, that is, all possibilities consistent with chemical valence are included without duplication. We feel that complete and irredundant generation is one of CONGEN's strongest selling points as it assures no possibilities are overlooked in a structure elucidation problem. We propose to provide the same assurance for conformation generation. However, since there are infinitely many conformations possible for most structures, as torsional angles vary continuously, a complete and irredundant generation is possible only if a finite number of discrete values are allowed for the torsional angles. To accomplish the desired total conformation generation it will be necessary to develop:

- i) a generation algorithm,
- ii) a means of representing conformations,
- iii) an assurance of irredundance.

C.4.a.i Generation Algorithm.

We propose to generate conformations for a CONGEN stereoisomer by labelling

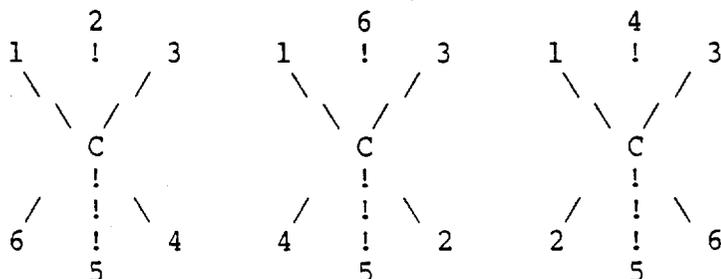
the bonds (edges) of the graph representing the stereoisomer with the possible bond torsional states. The algorithm will be developed so that any number of possible torsional states can be specified. The possible torsional states might be the familiar trans, gauche (+), and gauche (-) states which correspond to the minima of the torsional potential for single carbon-carbon bonds (e.g., the central bond of n-butane). Alternatively, the states might be the six (three staggered, three eclipsed) possible for a single carbon-carbon bond. The former choice might be made in cases where only local bond potentials are considered important, while the latter might be made when global potentials are important. The point here is that the algorithm would have to handle either case or any other choice of possible torsional states [82].

Part of the algorithm would involve recognizing the various kinds of rotatable bonds (e.g. sp³-sp³, sp³-sp², sp²-sp², etc.) and treating each accordingly. The problem of deciding which are rotatable bonds would be handled in a manner analogous to that used for determining stereocenters in the algorithm for configurational stereoisomer generation [25]. This method would start out by assuming every bond is rotatable and then proceed to reject those known to be fixed by intrinsic or input constraints (section C.4.b, below). Generation would proceed on the remaining bonds with further testing done as the conformations are generated. This method is similar to that used successfully for other structural generation problems in CONGEN.

We propose to do the generation of conformations as a labelling since our extensive experience with such algorithms suggests that these are very fast and efficient algorithms for problems of this kind. In particular, we believe that this method would be faster than an algorithm based entirely on atomic coordinates. The result of such a labelling would be a CONGEN structure represented as a graph with labels on some of the edges (bonds). The result of a generation of this kind would be a list of structures with bonds labelled in all possible ways and including structures not consistent with constraints (such as those imposed by intrinsic structural features such as ring closure). Labelling algorithms assure completeness since the labels are distributed in all possible ways. In our experience they are also constrainable, a particularly important consideration for conformations.

C.4.a.ii Canonical Representation.

In order to deal with conformations in the computer, some canonical (unique) representation for each conformation is required. We will likely construct such a canonical representation by making use of the atom numbering of the input stereoisomer. This is in analogy to the canonical representation used for the configuration of the stereocenters in the stereoisomer [24]. For example, consider the problem of uniquely describing the three staggered rotamers of a structure numbered as:



A method which would give unique representations for each of these three rotamers

would be to choose the lowest numbered substituent on each of the two atoms involved in the bond and express the bond state in term of them. In the example, this would be 1,2-gauche(+), 1,2-trans, and 1,2-gauche(-) respectively. The convention of designating clockwise rotation with a positive sign was used. If any or all of these were equivalent by symmetry, an ordering of the torsional states would be required, and the canonical representative would be the lowest. For example, if the ordering based on torsional angle were used: gauche(+) < trans < gauche(-), and all three were equivalent in the above case (because of equivalent substituents) the canonical representative would be 1,2-gauche(+). This method of canonical representation assumes that the numbering of the atoms was determined before the edge labels are added. This sequence corresponds to that which will be used in CONGEN since the atom numbering is done when the constitution (bond to bond connectivity) is determined.

C.4.a.iii Irredundance.

Assuring irredundance (no duplicate structures generated) in the labelling algorithm being proposed for conformation generation requires the proper use of the symmetries involved in the problem. Since CONGEN can produce structures of arbitrary symmetry, the conformation generator must be able to deal with any possible symmetrical structure. It is for this reason that symmetry must be dealt with so generally, not because of any claim of preponderance of symmetrical structures in biomedical problems. To our knowledge, this general problem of considering arbitrary symmetry in conformation generation has not yet been solved. The problems for which conformation generation has been treated either have no symmetry at all or symmetry of just one type (such as that in cyclic hydrocarbons). This problem can be solved by using a method similar to that used to irredundantly generate the configuration stereoisomers [24]. For the conformation generation problem, the symmetry group of the input CONGEN stereoisomer is represented by its effect on the possible torsional states for each rotatable bond. The equivalence classes of this group on the set of all possible conformations are the symmetry distinct conformations. A lowest representative of each class will be the canonical representative. This method will also readily give an enumeration formula for conformations which will be of use to the program since it will provide an independent check of the conformation generator and will allow faster computation of the number of conformations.

The method for irredundant generation of conformations will be illustrated on tetrachloromethylmethane $(\text{ClCH}_2)_4\text{C}$. The objective is to find the distinct rotamers of this structure with only perfectly staggered rotamers allowed. To do this it is necessary to represent the symmetry group of the structure by its effect on the possible torsional states of the rotatable bonds. The symmetry group of this structure is its Configuration Symmetry Group [24] and is isomorphic to the tetrahedral point group. This group is computed by the current stereoisomer generator in CONGEN. It is this group which must be represented on the possible torsional states of the bonds. This group has five "kinds" of symmetry operations: the identity, eight C3 rotation axes, three C2 rotation axes, six reflection planes, and six S4 alternating axes. Each of these five will be considered in turn to ascertain their effect on the possible torsional states. The identity operation has no effect on the possible torsional states. The C3 operations, however, do have an effect. Each C3 axis is colinear with one of the C-C bond axes. Rotation about this axis has the effect of interchanging the three possible torsional positions about that axis. It has no effect on the torsional positions about the other three axes. Each C2 axis interchanges pairs of bonds and the torsional states around each. With this much information it is possible to compute the number of possible rotamers using the method of Kerber [83]. The identity operation can be expressed

by the permutation (1)(2)(3)(4) of atoms. The C3 operations can be expressed as permutations like (123)(4). There are eight of these. The C2 operations can be expressed as permutations (12)(34) and there are three of these. The number of rotamers can be computed by substituting 3 for each orbit in the permutation and multiplying by the number of permutations, summing over all types of permutations and dividing by the order of the group. This is done as stated for the identity and the C2 axes. However, because the C3 axes interconvert the possible rotamers about one axis, this term contributes zero to the total. This conclusion can be reached without rationalization by simply using the formula derived by Kerber [83]. This yields $(1/12)*((3)*(3)*(3)*(3)+3*(3)*(3)) = (1/12)*(81+27)=9$ which is the correct number of rotamers for this structure [84]. The number of enantiomeric pairs can be computed by following through the procedure for the other symmetry operations. This procedure has computed the number of equivalence classes for this permutation group. A generation algorithm can be developed by simply taking each possible rotamer in turn and letting the symmetry group represented in this manner act on it. This process will actually construct the equivalence classes and the lowest member of each class is the canonical representative. While this is not the most efficient algorithm, it serves to illustrate the method.

C.4.b Constrained Generation of Conformations.

As in the previous efforts to develop generators of constitutional isomers and configurational stereoisomers, once a total generator is devised, it is necessary to modify it so that constrained generation can be done. We propose to develop a constrained generator of conformations in this manner. Constraints relevant to conformation generation include 1) intrinsic structural constraints such as that imposed by ring closure, 2) constraints input by the user based on partial structural information, and 3) constraints imposed by dynamics since most chemical structures exist in several conformations in rapid equilibrium.

C.4.b.i Intrinsic constraints.

The intrinsic constraints on conformation generation imposed by ring closure in polycyclic structures are probably the most difficult problem facing a conformation generation program of the type proposed here. Not all possible values of torsional angles for bonds are consistent with ring closure in these cyclic systems. For a polycyclic structure, the conformations possible for each ring are constrained by the conformations of adjacent rings, in addition to the constraint of ring closure. The relative constraints of overlapping rings can be determined by establishing the overlapping bonds and the configurations of any stereocenters involved. For example, in cis-decalin the common bond to the two six-membered rings has the same signed torsional angle (with respect to the ring carbons) in both rings. In trans-decalin, the common bond has opposite signed torsional angles in the two rings. For a given polycyclic structure these relationships can be computed by finding all rings in the structure and establishing the relative configurations of the stereocenters involved. The ring-finding can be done using a currently available function in CONGEN. Hence, the general problem reduces to that of determining the possible conformations for a single ring with constraints on the values for the torsional angles of some bonds. The ring-finding would have to be done only once for all the stereoisomers of given constitution. This is an efficiency because the list of structures coming out of CONGEN has all stereoisomers of the same constitution together. Another efficiency would be to consider only nonenveloping rings (those which do not "contain" smaller rings).

The process of generating conformations for a single ring with constraints

will be the "unit operation" of the conformation generator and will have to be optimized for efficiency besides being complete and irredundant. Six possible methods can be suggested based in part on the work of others.

1) The most direct method would be to systematically vary each bond in turn and check for end-to-end distance and overlapping atoms trigonometrically using coordinates. This would be a depth first generation with pruning when atoms overlap or the end-to-end distance becomes too large to allow closure.

2) A variation on this approach successfully used by Barry [64] would be to vary "fold axes" or dihedral angles which have as their axis the line connecting two nonbonded atoms. This method is more efficient because ring closure is more easily maintained but is not always exhaustive for one set of fold axes. The same distance checking would be done here.

3) A different method would be to use internal coordinates which measure the "pucker" of the ring and directly take account of the ring symmetry. These coordinates are usually based on displacement from a reference plane of the ring but could probably be expressed in terms of torsional angles. This method would probably have to be modified to assure exhaustion for large rings. This method also requires coordinates, but in a somewhat simpler fashion [58,60,61].

4) Another method makes use of ideas developed by workers in conformational analysis [85]. Families of conformations can be differentiated by the sequence of the signs of the bond torsional angles. For example, the chair conformation of cyclohexane has six torsional angle sign changes as one goes around the ring. The boat-twist family has four torsional angle sign changes. Similar patterns are evident for families of conformations of larger rings. This observation suggests that conformations might be generated by first discerning such families and then generating within each family. The families might be generated by first labelling the atoms with locations of torsional angle sign changes (only an even number of these are possible for any size ring) and then labelling the bonds with the values of the torsional angles. The first labelling reduces the number of possibilities for the second labelling. This is a reductionist approach which resembles the vertex graph method used so successfully in the first version of the CONGEN cyclic structure generator.

5) A related labelling method would make use of the polygon classification method for conformations developed by Dale [56]. Conformations are considered as polygons made of straight chain edges of varying lengths. This has the effect of reducing a ring of n atoms to a polygon of $m < n$ sides. Generation with a method like this would involve generating the possible polygons and labelling the edges in all possible consistent ways with numbers of bonds. This is also a reductionist approach similar to the one above.

6) A different method would be to make use of a "catalog" of the possible conformations for each ring size with a backup generator for cases involving rings larger than those in the catalog. Each ring in the structure would then be labelled with each of the possible conformations consistent with any constraints. This method effectively trades faster runtimes for a larger storage requirement.

These six possibilities have been ordered by their probable speed. The choice of one of these, a combination of them, or another method entirely will depend on speed, programmability, and constrainability along with an assurance of completeness and irredundance.

Particular attention will be paid to common conformational features such as six-membered rings. This will probably be done most efficiently by making use of a "catalog" of six-membered ring conformations (item 6 above). The possible six-membered ring conformations will be generated by reference to the catalog and any current constraints. This method will simplify the interpretation of conformational observations in this common case and will easily permit use of notions most familiar to chemists (chair, twist, axial, equatorial, etc.).

Other common intrinsic constraints involve rigid substructures such as multiple bonds and three-membered rings. These features are easily recognized with currently available parts of CONGEN which will simply be used in the conformation generator as well.

C.4.b.ii Input Constraints.

These are the constraints which will be input by the user when solving a particular structural problem. Common constraints of this type will arise from interpretation of NMR coupling data (vicinal and longrange), steric effects on C13 NMR (e.g. axial vs. equatorial substituent shifts) and interpretation of CD or ORD spectra (sector rules, Brewster's rules, etc.). Such constraints will be expressed as desired or undesired substructures which include designations for absolute configurations and bond torsional angles. They will be dealt with in a manner similar to that in CONGEN now, that is, by graph matching and pruning. Particular interest will be directed toward expressing constraints dealing with CD spectra because of ongoing efforts in Prof. Djerassi's research group on this topic and because of recent developments in new methods such as Magnetic Circular Dichroism (MCD) [86], Vacuum ultraviolet Circular Dichroism (VUVCD) [87] and Vibrational Circular Dichroism (VCD) [88]. The latter (VCD) is particularly interesting because this method will likely provide direct evidence about individual chiral environments of many chromophores. Some or all of these new methods will be particularly useful to structure determination of biomedically relevant molecules because almost all such structures are chiral. Another type of input constraint arises from observations about symmetrically equivalent atoms. The symmetry of conformations will always be less than or equal to the symmetry of the configurational stereoisomer from which they are generated. Thus, to take proper account of observations about the symmetry of a unknown structure, consideration of the possible conformations is crucial.

C.4.b.iii Dynamics.

A different sort of constraint arises in the conformation generation problem since most flexible molecules are in rapid equilibrium among several different conformations. Thus a structure determination of the type discussed here which includes conformations may lead to several final structures rather than just one. This sort of information could be expressed as a constraint which requires at least n conformations or requires only one be present. Alternatively, there may be an observation which requires that there be interconversion between two known partial substructures (chair-chair interconversion in cyclohexane for example). To make use of this information it will be necessary to predict flexibility in conformations. For certain situations such a prediction is fairly simple. For example, most acyclic substructures are flexible if not too heavily substituted. In many structure determination problems there may be no interest in flexibility of acyclic substructures, hence these could be recognized and ignored in further conformation generation leading to a savings in time and storage. Other kinds of flexibility which are fairly easily recognized involve ring pseudorotation, large unconstrained

rings or parts of rings, chair-chair interconversion, etc. However, predictions of flexibility in this way can only go so far since this property depends strongly on energetics. A typical structure elucidation problem might end with the observation of several final structures and the question of whether or not they can be interconverted and their relative populations. Resolution of this question would require an energy calculation of some type. The conformation generator will allow computation of all the possibilities, an important piece of information in problems of this kind.

C.4.c Interface and Applications.

CONGEN Interface.

The conformation generator program will eventually become part of the CONGEN program. Conformation generation will take place after configuration stereoisomer generation. Information will flow to the conformation generator which describes the constitution and configuration of the stereoisomer for which conformation generation is to be done. The conformation generator will construct a list of possible conformations subject to input constraints and will return information about continued candidacy of the input structure in the ongoing structure elucidation problem. Thus it may be possible to eliminate a stereoisomer or even a constitutional isomer from further consideration based on results of the conformation generation. The user interface will resemble that currently in CONGEN and in fact will use many of the program sections already written for this purpose.

Other Applications.

Besides improving CONGEN as a tool for structure elucidation, this proposed addition of a conformation generator will lead to potential new biomedical applications. The CONGEN program with a conformation generator will output structures complete with internal (torsional angle) coordinates. Since many existing programs require as input a structure with coordinates, the opportunity exists to interface CONGEN with these programs. Examples of such programs are empirical force-field energy calculations (molecular mechanics), quantum mechanical energy calculations (probably semi-empirical), graphics programs, and finer grid conformation generation (smaller torsional angle increments with van der Waals checking). These programs and methods frequently see biomedical applications; an example is in the determination of structure-activity relationships. By interfacing CONGEN to these existing programs, the opportunity for new biomedical applications will be expanded. Since these programs have been extensively developed by others, it will not be necessary for us to spend the time to develop them. Instead, we propose to use collaborative efforts to take advantage of these very interesting new applications. (See Section F, Collaborative Arrangements, section C.5.d Graphics Interface).

Another application would be to the field of conformational analysis. The conformation generator would be useful to such efforts by exhaustively and irredundantly generating the possible conformations which must be considered in such a study. We propose to explore this possibility collaboratively as well (Section F).

The conformation generator program will also be written to exist by itself (besides being a part of CONGEN) by providing a mechanism for inputting structures

from other sources. Since we expect this program to be fast and efficient and able to deal with large lists of structures, it is conceivable that the conformation generator could be used on lists of structures from other data bases to "upgrade" a list of structures which do not yet contain any conformational information. This could lead to applications in pharmacology or toxicology and other areas which make use of large chemical structure data bases.

C.5 Resource Sharing.

C.5.a Access to Programs via SUMEX and Local Dedicated Computer and Resource Management.

In Section A.2, Background, we discussed the relationship of our project to the SUMEX resource. In that discussion we outlined the conflicts between program development and extensive "production" use of resulting applications programs by collaborators both within and outside of the SUMEX community. We propose to resolve these conflicts by providing separate machines for development and production uses, each available to local and network users.

We propose that program development take place, as it has in the past, on the SUMEX PDP-10 system. This system provides the requisite facilities for languages, editors, message handling and so forth, to support effectively such development. In addition, SUMEX will provide the gateway for access to production use the proposed dedicated computer (see Figure 10). In the past, few collaborators have participated directly in program development. Their contributions have been, for the most part, indirect in that many suggestions resulting from trial use of CONGEN were incorporated by our group into programs, resulting in improved performance and greater chemical "intelligence". Now that more chemical and biochemical research groups are becoming sophisticated in their use of computers, we expect that, during this proposal period, our collaborators may desire more direct involvement, Cowburn at Rockefeller for example. We will encourage such collaboration and will carry out such work on SUMEX.

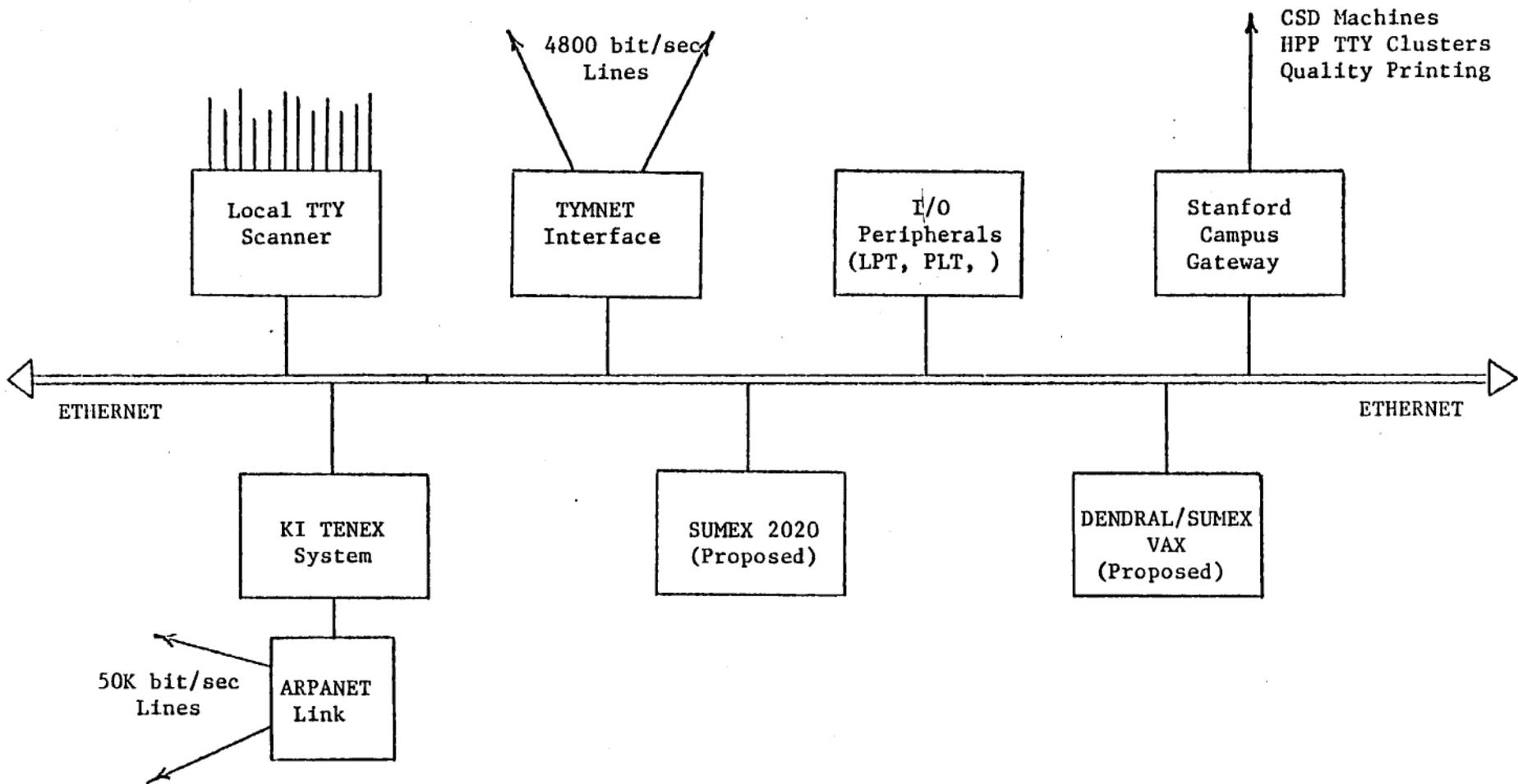


Figure 10. Access to DENDRAL programs on SUMEX

Djerassi, Carl

The production machine will provide more routine access to programs whose development and testing has progressed to the point where outside use of the programs by a wide community can be encouraged. Currently, CONGEN and auxiliary programs or functions, (STEREO for unconstrained generation of configurational isomers, SURVEY for examining structures, MSANALYZE for mass spectral prediction and ranking and REACT for simulation of chemical reaction sequences) are in this category. This use of our programs may involve local users, i.e. persons in our own group or local collaborators, or remote users accessing programs via TYMNET and SUMEX. We propose that access for applications, or production, use be handled by a computer linked to, but separate from the SUMEX system, specifically a Digital Equipment Corporation VAX-11/780. The schematic diagram of Figure 10 conveys our proposed design to allow users to access either machine, taking advantage of all existing network communications facilities on SUMEX. Whether computations will be carried out on SUMEX or the VAX will depend on the users and the program to be run. Only SUMEX users involved in DENDRAL will be enabled to run on the VAX and vice-versa. By exercising such administrative controls we will assure allocation of the DENDRAL machine for biochemical work while maintaining flexibility for network mail routing and approved file transfers. The VAX will have its own disc system, for system programs and page-swapping in the time sharing system of the VAX. The network configuration does not make the DENDRAL VAX dependent on SUMEX being up for access, yet still takes advantage of existing SUMEX communications hardware and other selected peripheral equipment (e.g., printer, plotter, TTY ports) without duplicating them.

We have discussed previously (see Budget Remarks) the reasons for our preliminary selection of the VAX computer. Briefly, in our opinion, it represents an optimum trade-off of several factors including interactive computing, large address space, compatibility with the biochemical and computer science communities with which we interact, long-term trends within the computing industry and cost-effectiveness. Several of the persons attending our workshops last year (see Annual Report, Appendix I) have, or soon will have, a VAX system. The National Resource for Computation in Chemistry will soon obtain a VAX. Taken together, these factors point towards the VAX as at least a good representative of the future shape of scientific computing appropriate to our applications.

The implementation of the VAX system as a part of the SUMEX resource, to be used as discussed above, will involve primarily the efforts of the two scientific programmers identified in the budget, under the guidance of the scientific staff and the Advisory Committee (see below). Their responsibilities, outlined in the Budget Remarks, are discussed here in more detail. We identify these programmers in the following discussion as befits their responsibilities, "systems programmer" and "applications programmer." Although we will be able to call upon the advice and expertise of the existing SUMEX staff in the implementation and continued maintenance and support of the VAX, in our opinion a full-time systems programmer and a full-time applications programmer (concerned primarily with the VAX) represent the minimum staffing required to support our VAX system. Initial implementation of the VAX will include the following:

i) select either DEC's, or Bell Lab's UNIX, operating system (the responsibility of the entire staff).

ii) together with the SUMEX staff, make the necessary operating system modifications to allow access as outlined in Figure 10 (systems programmer).

iii) obtain, (or write) a BCPL compiler for the VAX in order to run the current CONGEN, GENOA and related programs written in BCPL (applications programmer).

While (i-iii) are being carried out, SUMEX will be used for applications under community controls in effect at the present time (GUEST access, strict control over SUMEX resource allocation (computer cycles) for local users). The applications programmer will coordinate this access and provide required documentation.

Once the BCPL compiler has been brought up, our next step will be to put the current production versions of CONGEN and GENOA on the VAX. At this stage we will have full availability of these programs and the system to the outside community, thereby relieving SUMEX of the bulk of the current production use (only programs in INTERLISP and experimental versions of new programs open for testing will be on SUMEX at this stage) (applications programmer).

At this time we will schedule a workshop, aggressively inform outside users of the system's availability and encourage them to thoroughly test the current versions of the programs and the VAX system. Based on our past experience with workshops and outside users we expect, because of a responsive system and many newly enthusiastic users (workshop participants), to have an active community of outside collaborators using our programs on the VAX. We expect this situation to occur in approximately late 1980 to early 1981 depending on the delivery time of the new system.

Following this initial period we will enter a phase of stepwise development of new programs on SUMEX, experimental testing of new applications programs by selected outside collaborators on the VAX and, finally, community-wide announcement of production versions of these programs on the VAX. (The scientific staff will be responsible for most of the new program development. The applications programmer, with the aid of the scientific staff and the systems programmer, will have important responsibilities in the resource sharing of experimental and production versions.)

Future DENDRAL program developments on SUMEX will be in languages for which compilers exist, or will soon exist, on VAX e.g. PASCAL, FORTRAN, MAINSAIL and BCPL. Obtaining these languages, and other system support packages such as text editors, will be the responsibility of the systems programmer. It should be noted that these efforts plus maintenance of the operating system and ancillary programs is a full-time, continuing job for a system of the complexity of the VAX.

There are proposals for an INTERLISP on VAX, which is probably a year or more away from implementation. Therefore, DENDRAL programs, such as REACT, MAXSUB and so forth, which are in INTERLISP and which we do not propose to convert to an exportable, algorithmic language, will remain accessible solely through SUMEX until an INTERLISP for VAX is available.

We plan, with the aid of the programmers requested in the budget, to implement on the VAX a number of chemistry-related programs which will directly support our new research and the work of our collaborators. We are particularly interested in molecular modelling programs such as MMI, CAMSEQ and PCILO, as an example. We plan to modify or adapt OMNIGRAPH to simplify our program developments for graphical input and output of structural information. We will work closely with others possessing VAX systems in order to share software and avoid duplication of effort as far as possible (e.g., VAX systems are being obtained by the National Resource for Computation in Chemistry with whom we are in contact, and Dr. David Pensak at DuPont who has been a collaborator with us in the past).

In a sense, we will be acting as brokers of some chemistry-related programs for the new generation of computers such as the VAX, for which little applications software exists currently. Our responsibilities as brokers would extend only to providing network access by qualified collaborators using our programs on the VAX,

and exporting applications programs to other VAX or VAX-compatible sites working on biomedically-relevant problems. We do not plan to compete with the quantum chemistry program exchange (QCPE), which distributes primarily programs that perform numerical computations. Our programs are primarily involved with symbolic computations. Also, we specifically do not intend to embark on language development for the VAX (e.g., INTERLISP), or to implement software which is not directly related to our goals or the goals of our collaborators. Although we may well utilize such programs and languages, such efforts are best done by other groups. We will monitor these efforts closely in order to take advantage of new software. Further into the grant period (two to three years) we expect that VAX and similar machines will be widely utilized within the biomedical community and that, as a result, we will be able to exploit the software available from the work of others.

Resource Management

To promote an orderly implementation of the dedicated computer system and to ensure that the respective goals of SUMEX-AIM and DENDRAL are being met, we propose to establish an Advisory Committee. This Committee will consist of the following persons:

Professor Edward Feigenbaum (chairman) - Principal Investigator - SUMEX

Professor Carl Djerassi - Principal Investigator - DENDRAL

Thomas Rindfleisch - Resource Manager - SUMEX

Dennis H. Smith - Co-Investigator - DENDRAL

Bruce G. Buchanan - Adjunct Professor, Computer Science - DENDRAL, MYCIN

The persons making up the committee are those with primary responsibilities for the conduct of both the SUMEX and DENDRAL efforts. They represent complementary knowledge and expertise, with Feigenbaum, Rindfleisch and Buchanan providing guidance on the computer science aspects of our research and on specific questions of the hardware and software interface with SUMEX and the dedicated machine, while Djerassi and Smith bring to the committee a strong emphasis on specific chemical and biochemical research problems of our group and those of our collaborators. Persons named above have worked extremely closely in the past (Prof. Feigenbaum, for example, was at one time PI on the DENDRAL grant); this committee will formalize an already existing close working relationship and focus attention on the specific area of resource sharing of our results.

This committee will advise on: a) planning the purchase, installation and development of the dedicated computer system; b) planning the hardware and software interfaces providing access to either the SUMEX or DENDRAL machines; c) assigning priorities for development of applications software on the VAX; d) promoting resource sharing activities such as lectures and workshops; and e) ensuring equitable access to the computer resources at Stanford by our collaborators. This committee will meet at two week intervals during the initial phases of the grant, and once per month thereafter.

C.5.b Exportable Programs.

To meet one of the goals of our present research, we have recently completed and begun distribution of an exportable version of CONGEN (see attached Annual Report). As we develop GENOA and the complete SASSES system, we will ensure that they retain at least the same degree of exportability. The proposed version for the VAX computer will certainly improve the likelihood that outside scientists will be able to run the program in their own laboratories.

An intriguing philosophical question is why produce exportable versions for different machines when the program can be run at one site, accessible nationwide by computer networks. There are some practical reasons for this. Industrial firms require a high degree of secrecy for key structural problems and feel strongly that, to retain control and guarantee secrecy, they need a version operating on their own computers. Some academic laboratories have free access to mini/midi computer systems and cannot afford communication and commercial network/computer costs. When these practical reasons are exhausted, there remains a significant group of persons who simply want a version on their own machine under their own control

There is, however, another practical reason and that is, despite our efforts over the past few months, we have been unable to provide even experimental network access to CONGEN except the restricted access at SUMEX. There are two likely sites to which CONGEN can in principle be exported, both of which are easily accessible by computer network. The first is the NIH/EPA Chemical Information System (CIS), which operates a PDP-10 system on which CONGEN in its current form can be run without modification. For a year or more we have been negotiating, with persons responsible for CIS, for funds to support the conversion required to integrate CONGEN into CIS. Such funding has, we understand, received favorable Division of Research Resources/NIH review. However, we have recently learned that such funding will not occur now, and the future seems to be in question. It is unlikely that this situation will be resolved before the current proposal is submitted.

Another possible site is the Control Data (CDC) 6600 system, accessible at the Lawrence Radiation Laboratory (LBL) in Berkeley, through the National Resource for Computation in Chemistry (NRCC). We have arranged for some trials of the interactive BCPL system existing there to explore how CONGEN might operate in that environment. We do not expect these experiments to be fruitful because CONGEN relies heavily on intermediate files (to and from which structural information is communicated many hundreds of times during a problem); access to the file system is one of the primary bottlenecks of the LBL CDC system, access times being up to many seconds during heavy use of the system. However, the imminent purchase of a VAX system by the NRCC would change the situation dramatically. This factor, plus the uncertainties of the CIS lend considerable emphasis to our desires to implement readily-accessible versions of our programs on our own VAX system, with distribution to a wider community of persons handled eventually by the NRCC or the CIS.

C.5.c Workshops.

Our successful experience with workshops on the use of the new CONGEN program prompts us to propose a continuing series of such workshops designed to serve purposes similar to the previous effort. We want to introduce new collaborators to applications programs as they are developed, and to promote new applications and wider use of the programs. We want both new and current collaborators to evaluate new programs so that bugs, and clumsy or unclear aspects of the interaction can be resolved before development of the final version for application. Most importantly, these workshops will provide a mechanism for keeping our group up-to-date on requirements of the community for computer assistance in their efforts and provide a far more diverse set of structural problems than we would encounter in work here at Stanford. We have budgeted funds for one workshop per year, to be organized by our personnel and held here at Stanford University (or possibly elsewhere, the network providing access to the system).