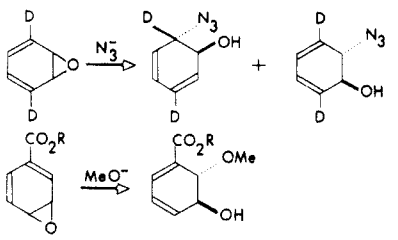
Scheme II



S_N2 reaction, with sulfur nucleophiles, opposite stereocontrol results from this frontier interaction as shown in Figure 14. Monointeractive nucleophiles exhibit syn stereochemistry in the S_N2' pathway while diinteractive nucleophiles provide more anti product and increase the S_N2/S_N2' product ratio in a competitive reaction. This competitive behavior of nucleophiles is clearly demonstrated in the careful studies by Stork et al. on the intermolecular reaction of piperidine and propanethiolate with the 2,6-dichlorobenzoates of *cis*- and *trans*-6-isopropyl-2-cyclohexen-1-ol which provide a product distribution[15a,b,17] (Chart II). Indeed, the sterically unencumbered intramolecular S_N2' cyclization involving a thiolate nucleophile proceeds exclusively with anti stereochemistry while an intermolecular version with a secondary amine proceeded exclusively with syn stereochemistry.[15a,b]

Extension of the analysis of the hyperconjugative interaction of a C–X fragment with a conjugated trans and cis diene (Figure 15) in a rigid system dictates that a monointeractive nucleophile is directed by orbital distortion in an anti–syn–anti alternating pattern relative to the C–X fragment. Although diinteractive reagents experience a dicotomy with respect to the two frontier interactions at the 3-position, the distortions of the diene HOMO and LUMO are both anti with respect to the C–X fragment at the 5-position. The elegant work of Berchtold[18] and Kishi[19] with an arene oxide as a cis dienoid model beautifully demonstrates the correctness of this application of the principle of orbital distortion (see Scheme II).

Finally, analysis of a C-X fragment hyperconjugated with an alkyne unit (Figure 9) gives a stereochemical result quite different from the alkene counterpart as a result of the appearance of a lowest lying antisymmetric vacant σ^* orbital. Mixing in second order of this σ^* orbital with a π^* orbital directs an incoming monointeractive reagent at the 3-position along a path anti to the bond of the C–X fragment. Diinteractive reagents take advantage of an additional interaction with the substrate highest occupied π molecular orbital, and as in the case of an alkenyl substrate, these orbitals are distorted in the anti direction. Since both substrate frontier π orbitals are distorted in concert, a high degree of stereospecificity should be demonstrated in the S_N2' reaction for alkyne substrates with both reagent types. This conclusion has been exemplified in the reaction of a number of 1-alkyn-3-ol derivatives with diinteractive and monointeractive reagents to give allenes derived from an anti reaction path.[20]

Acknowledgment. We express our sincere thanks to many colleagues whose encouragement and criticism provided the stimulus for this adventure. E.M.B. especially appreciates the support of the National Science Foundation which allowed unencumbered summers for the above work.

(18) G. A. Berchtold, R. M. DeMarinis, C. N. Filer, and S. M. Waraszkiewicy, *J. Am. Chem. Soc.*, **96**, 1193 (1974); G. A. Berchtold, D. M. Jerina, et al., *ibid.*, **96**, 6929 (1974).
(19) Y. Kishi, *J. Am. Chem. Soc.*, **98**, 6723 (1976).
(20) (a) W. H. Pirkle and C. D. Bolder, *J. Org. Chem.*, **43**, 1950 (1978); (b) E. J. Corey and W. T. Borden, *Tetrahedron Lett.*, 313, (1969); (c) P. Crabbie, J. L. Luche, E. Barreiro, and J. M. Dollat, *ibid.*, 4615 (1974); (d) O. J. Muscio, Jr., Y. M. Jun, and J. B. Philip, Jr., *ibid.*, 2379 (1978); (e) G. Tadema, et al., *ibid.*, 3935, (1979); (f) M. Midland, *J. Organomet. Chem.*, **156**, C5 (1978); (g) Y. Yamamoto, *ibid.*, **156**, C9 (1978).

GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures[1]

Raymond E. Carhart, Dennis H. Smith,* Neil A. B. Gray, James G. Nourse, and Carl Djerassi

Department of Chemistry, Stanford University, Stanford, California 94305

Received September 23, 1980

An interactive computer program called GENOA, for isomer generation based on overlapping and alternative substructures, is described. This program produces an exhaustive and irredundant set of structural isomers based on these substructures and thus has direct application as a computer aid to molecular structure elucidation. The key algorithm in GENOA, constructive substructure search, solves the problem of piecing together substructures which may overlap to any extent. This algorithm provides efficient, prospective use of the often ambiguous and redundant structural information collected on an unknown structure by using a variety of complementary spectroscopic and chemical techniques. Advantages of this approach to structure elucidation are discussed, including simplicity of use and direct interface to programs for automated analysis of spectroscopic data. Examples of use of GENOA in actual structure elucidation problems are presented. Novel aspects of the algorithm for structure generation are described.

A. Introduction

In recent years several computer programs have been written to perform the structure generation task in computer-assisted structure elucidation.[2] These programs have as their common goal the construction of computer representations of all isomeric molecular structures which obey a set of constraints on desired and undesired features

(1) Part 37 of the series "Applications of Artificial Intelligence for Chemical Inference". For Part 36, see N. A. B. Gray, A. Buchs, D. H. Smith, and C. Djerassi, *Helv. Chim. Acta*, in press.

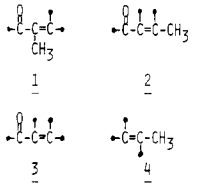
(2) "Computer-Assisted Structure Elucidation", D. H. Smith, Ed., ACS Symposium Series 54, American Chemical Society, Washington, D.C., 1977.

0022-3263/81/1946-1708$01.25/0

of an unknown structure. The resulting isomers then represent the set of structural candidates for the unknown. The constraints are usually in the form of substructural fragments of an unknown structure. The presence (or absence) of these substructures is inferred, manually or automatically, from spectroscopic and chemical data collected on the unknown. These programs, called structure or isomer generators include DENDRAL and CONGEN, developed by our own group,[3-5] the CASE system developed by Munk and co-workers,[6,7] the CHEMICS system of Sasaki and co-workers,[8,9] and the MASS module of the STREC system developed by Gribov and co-workers.[10] Other, less comprehensive structure generation methods have also been presented.[11,12]

These methods have important limitations. In the real world, a collection of spectroscopic and chemical information collected on an unknown structure tends to be highly *redundant* and often *ambiguous*. None of the computer programs mentioned above have good mechanisms for handling such information. Consider redundancy and ambiguity.

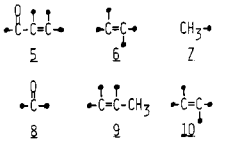
Redundancy. By *redundant* we mean *overlapping*, in the sense that structural inferences (substructures) derived from different techniques tend to overlap one another; although some overlaps are obvious, the extent of overlap is often unknown. Consider as a trivial, but illustrative example the presence in an unknown structure of an α,β-unsaturated ketone (e.g., inferred from UV) and a vinyl methyl group (e.g., inferred from ^{1}H NMR). Without additional information, and as long as the molecular formula allows, the double bond in the two substructures may or may not be the same. Thus, an investigator is forced to consider as alternative substructures either **1** or **2**, both of which assume the double bond is overlapping between the substructures, and the pair **3** and **4**, which assumes the double bonds are distinct. However, current programs for structure generation require *nonoverlapping* substructural units as the basis for beginning the constructive procedures. In other words, no substructure to be used as building block for the complete molecule should share any atoms with any other substructure also used as a building block.



Ambiguity. By *ambiguous* we mean that two or more substructures may be plausible alternatives for a given item of structural information. Although manual approaches to structure elucidation keep these alternatives in mind, no program has direct mechanisms for considering such alternatives. Rather, the programs require rigorous specification of a set of substructures. If there are alternatives they must be treated as separate problems.

These program limitations are severe for at least two reasons. First, the inability to use overlapping or alternative substructures directly in a structure-generating program can lead to terrible inefficiencies, resulting in fragmentation of one problem into several and consumption of inordinate amounts of computer time. Second, the programs, whose primary task is to emulate the structure-building capabilities of a chemist, have no or, at most, nonintuitive mechanisms for handling redundant and ambiguous structural information. This further complicates the task of the structural chemist who is supposed to use a program as an aid to structure elucidation.

Without going into details (which are in any case available in the literature) it is useful to discuss briefly the various ways in which current programs deal with potentially overlapping or alternative substructures. CONGEN and CASE depend on manual inference of substructural units (of arbitrary complexity). Painstaking manual analysis of the problem is required to ensure that the set of substructures ("superatoms"[5]) supplied to the programs do not overlap. This generally results in substructures of size smaller than the original inferences, which increases the computation time and the number of resulting isomers. Tests for substructures which potentially overlap other substructures (i.e., "GOODLIST"[5]) are deferred to the end of structure generation, further increasing computation time. For example, referring back to the example introduced above and assuming that there is in fact a second double bond, CONGEN[5] would use as starting points **5–7** (with a GOODLIST constraint that **7** must be bonded to **5** or **6**) or **8–10** (with a GOODLIST constraint that **8** must be bonded to **9** or **10**) in order to ensure completeness. A



different set of starting points would have to be used if there was uncertainty about the presence of the second double bond. The CASE[7] program does provide the capability of characterizing the immediate neighbor of each atom of a superatom which bears unfilled bonding sites, thereby implicitly allowing for one-atom overlaps. In addition, some statements can be made about a more general environment, such as ring size.[7] However, this is not a general solution to the problem of overlapping substructures.

The CHEMICS[9] and MASS[10] programs avoid overlapping substructures in a different way. Both programs break down the substructural inferences (usually derived automatically from computer-based interpretation of spectral data) into smaller substructures which are members of a library of small building blocks within which overlap is not possible. Structures are built by assembling these smaller substructures into complete molecules under the constraints that the molecular formula not be exceeded and that every resulting isomer possesses the originally inferred number of substructures of each type. Although this procedure avoids overlapping substructures, it is very inefficient for large molecules because most of the information on the original bonding among atoms of the large substructures is discarded until late in the computations

(3) R. E. Carhart, T. H. Varkony, and D. H. Smith, ref 2, p 126.
(4) J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, *J. Am. Chem. Soc.*, **91**, 2973 (1969).
(5) R. E. Carhart, D. H. Smith, H. Brown, and C. Djerassi, *J. Am. Chem. Soc.*, **97**, 5755 (1975).
(6) M. E. Munk, C. S. Sodano, R. L. McLean , and T. H. Haskell, *J. Am. Chem. Soc.*, **89**, 4158 (1967).
(7) C. A. Shelley, H. B. Woodruff, C. R. Snelling, and M. E. Munk in "Computer-Assisted Structure Elucidation", D. H. Smith, Ed., ACS Symposium Series 54, American Chemical Society, Washington, D.C., 1977, p 92.
(8) T. Yamasaki, H. Abe, Y. Kudo, and S. Sasaki, ref 7, p 108.
(9) S. Sasaki, Y. Kudo, S. Ochiai, and H. Abe, *Microchim. Acta*, 726 (1971).
(10) L. A. Gribov, M. E. Elyashberg, and V. V. Serov, *Anal. Chim. Acta*, **95**, 75 (1977).
(11) N. A. B. Gray, *Anal. Chem.*, **47**, 2426 (1975).
(12) G. Beech, R. T. Jones, and K. Miller, *Anal. Chem.*, **46**, 714 (1974).

(most partial or complete isomers generated will be discarded in such post-testing).

Alternative substructures are generally treated in all programs by solving separate structure generation problems and somehow combining the results. This procedure is successful for small structures (<10 nonhydrogen atoms) with a small number of alternatives. However, for larger structures the number of combinations which must be considered rises dramatically if there are several sets of alternatives and the method of solving separate problems quickly breaks down.

The limitations of current structure generators together with difficulties faced by some chemists using CONGEN led us to seek an alternative method for structure generation. The method we have developed and describe in this paper more closely emulates the manual problem-solving procedures of chemists in arriving at candidate structures for an unknown, specifically taking into consideration all potential overlapping and alternative substructures. Our initial effort resulted in an INTERLISP[13] version of a program which performed some aspects of the necessary computations.[3] Based on our experience with that version, we have recently finished a much more complete and interactive program, in the exportable BCPL programming language,[14] which incorporates many new algorithmic developments and which is interfaced to the CONGEN program as described in subsequent sections. We call this program GENOA, for structure *GEN*eration with *O*verlapping *A*toms. GENOA constructs structures by taking into account all possible overlaps of substructures and all alternative substructures at the very beginning of the computational procedure. Substructures of arbitrary complexity can be used. The program is flexible enough to use collections of substructures *known* to be nonoverlapping, thereby retaining some of the desirable features of CONGEN.

GENOA provides several advantages over CONGEN and other structure generators. (1) GENOA avoids the requirement for detailed manual analysis of structural information to determine nonoverlapping structures. Structural information from a variety of sources can be used exactly as it is determined, without regard to potential overlaps. This saves an investigator's time and makes maximum use of the structural information early in a problem to help keep the problem size and computational time small. Thus, the program operates much more "intuitively" than CONGEN and related programs. The previous example concerning the presence of an unsaturated ketone and a vinyl methyl group is simply solved by making both assertions directly to GENOA. GENOA deals automatically with the problem of all ways in which the substructures may, or may not, overlap and arrives quickly at the possible solutions represented by 1, 2 and the pair 3, 4.

(2) Certain problems can only be stated in terms of potentially or guaranteed overlapping substructures. In such cases GENOA represents the only efficient (in some cases the only possible) way to solve the problem. One such problem, involving a cembranolide, was mentioned briefly in a description of the preliminary version of the program[3] and is described in more detail below.

(3) GENOA allows an investigator to specify alternative substructural information if data are ambiguous. In fact, scores can be associated with each alternative, resulting eventually in a rank ordering of structural candidates based on the relative plausibility of the substructures they contain.

(4) GENOA can be interfaced directly to programs designed for automatic interpretation of spectral data.[1,15,16] Any structural inferences, including alternative substructures, derived from such interpretations can be supplied directly to GENOA and used exactly in that form without the necessity for libraries of small structural fragments.[8,10] The interactive nature of GENOA allows such structural inferences to be supplemented with any number of additional constraints inferred manually from other data.

In subsequent sections we introduce several important concepts of the program, provide examples of use of GENOA in actual structural problems, and describe the key algorithms in some detail. (See paragraph at the end of the paper about supplementary material.)

Table I. Basic GENOA Commands and Their Functions

GENOA command	function
DEFINE	To define the chemical context of a structural problem, including: (a) the characteristics of the computer terminal for subsequent structure drawings, (b) the molecular formula, (c) the name and valence of atoms other than C, N, O, X, and H, or (d) the name and structural definition of a new substructure.
FIX	To alter a definition previously created by using DEFINE.
SHOW	To display on the computer terminal any of the following: (a) a tabular representation ("connection table") of selected cases or substructures, (b) a summary of the current status of the problem, (c) a history of the problem including all constraints used so far, or (d) the contents, by substructure name, of any substructure library.
SEARCH	To retrieve selected substructures from a library file.
CONSTRAINT	To incorporate the next substructural constraint as part of the procedure for constructing new cases.
ALTERNATIVE	To specify a set of alternative substructures as a single constraint.
DRAW	To print, on the computer terminal, drawings of selected cases or substructures.
FORGET	To forget, by deletion, the definition of selected atoms or substructures, or to delete specified cases from the list of cases.
SAVE	To store on a specified file all information about the current session.
RESTORE	To retrieve from a specified file all information about a previous session.
GENERATE	To generate final structures for the structural problem and to transfer control to CONGEN.
EXIT	To terminate investigation of the problem.

B. Method

B.1. Basic Concepts of the GENOA Program. GENOA is an interactive program designed to be used by the

(13) W. Teitelman, "INTERLISP Reference Manual", XEROX Corp., Palo Alto Research Center, Palo Alto, CA, 1974 (revised edition, 1975).
(14) M. Richards and C. Whitby-Strevens, "BCPL - The Language and its Compiler", Cambridge University Press, Cambridge, England, 1979.
(15) N. A. B. Gray, C. W. Crandell, J. G. Nourse, D. H. Smith, M. L. Dageforde, and C. Djerassi, *J. Org. Chem.*, **46**, 000 (1981).
(16) N. A. B. Gray, J. G. Nourse, C. W. Crandell, D. H. Smith, and C. Djerassi, *Org. Magn. Reson.*, in press.

Table II. Possible Ranges of Occurrence of Substructures Allowed by GENOA in a Typical Structure Elucidation Problem

range of occurrence	action by GENOA
NONE	None of the substructure will be allowed in resulting cases and, eventually, final structures.
AT LEAST X	*X* of the substructure will be included in each case resulting from the constraint. Subsequent additional occurrences resulting from accidental construction of the substructure are allowed.
AT MOST X	*No* substructures will be constructed; however, GENOA will always ensure that no more than *X* are present.
EXACTLY X	*X* of the substructure will be included in each case. *No* additional copies of the substructure will be constructed or allowed in subsequent steps. Every case will have exactly *X* of the substructure, no more, no less.
RANGE X TO Y	*X* of the substructure will be included in each case. Additional copies of the substructure up to and including a total of *Y* will be allowed in subsequent steps, but never more than *Y*.

chemist for analysis of his or her own structural problems. Wrapped around the structure generation algorithm itself is a set of routines which control interaction betwen the chemist and the program. Also provided is a selection of modules which perform utility functions, such as substructure definition and file manipulation. These are summarized in Table I, which lists the basic GENOA commands and a brief description of their functions.

For structural problems involving more than six or seven nonhydrogen atoms, constraints are *essential* in order to avoid generation of vast numbers[17] of unwanted structures. It is GENOA's method for utilizing substructural constraints that allows an efficient solution to structural problems of realistic size.

GENOA uses a single type of constraint, the substructure. (See CONSTRAINT command, Table I—the ALTERNATIVE command merely provides a simple mechanism for expressing ambiguous inferences, allowing specification of multiple, alternative substructures as a single constraint.) Any structural inference, however complex, can be used as a constraint as long as it can be expressed in terms of a substructure. Substructures are supplied to GENOA together with a desired range of occurrence. The allowed ranges of occurrence are summarized in Table II and cover all useful statements about how often a substructure may occur in a structure. Note that "none" is a possible range of occurrence (Table II). This means of course that the substructure *cannot* be present, i.e., this is a convenient way of expressing "BADLIST".[5]

Substructures to be used as constraints can be defined by using a variety of commands to associate atom and bond properties with portions of the substructure, for example, bond orders, hydrogen ranges, whether or not atoms are aromatic, and so forth. These properties are summarized in Table III.

A single substructural constraint can be composed of two or more unconnected components (i.e., "disconnected" substructures). Although GENOA will allow interconnection of the components, the program assumes that every atom in a disconnected substructure is unique, i.e., cannot overlap with any other atom in the substructure. This feature provides a convenient mechanism for specifying to GENOA, as a single substructural constraint, those substructures which are known to be nonoverlapping. Most structural problems have several such substructures which are easily recognized by the chemist from examination of spectral data. With this mechanism, many of the efficiencies of expressing substructural information to CONGEN are retained.

(17) D. H. Smith, *J. Chem. Inf. Comput. Sci.*, 15, 203 (1975).

Table III. Atom and Bond Properties Used by GENOA To Characterize Substructures

property	possible values
	Atom Properties
name	Any name, not restricted to atom names in the periodic table. Default is carbon. C, H, N, and O are predefined in GENOA. X is reserved to mean "any" atom name. Atom names can also be variables ("polynames") with specified values, e.g., F, Br, Cl, I.
hydrogen range	A range of hydrogens (specified from minimum to maximum, min = max for exact number) may be allocated to an atom. Default is ANY unless free valences are used, when the default is to saturate unspecified valences on other atoms with hydrogen atoms.
free valence	Free valences, bonds with an unspecified terminus, are used to specify bonding sites for interconnection of substructures and atoms. Default is "all"; if no free valences are specified, every unfilled valence is a potential bonding site.
aromatic type	Atoms can be designated aromatic, nonaromatic, or either. Default is either.
hybridization	Hybridization of atoms can be specified to be sp^3, sp^2, sp^{1a} (e.g., alkynes, nitriles), sp^{1b} (e.g., allenes, ketenes) or "any". Default is "any".
configuration	Atom and double bond stereocenters can be designated to possess chirality ("zero" or "one") or to be cis or trans, respectively.
color	Atoms can be designated to possess one or more "colors" in order to differentiate them from other atoms which are otherwise the same. This is useful for fine control over the degree of overlap of substructures.
tag	Atoms can be tagged to modify the determination of the number of occurrences of a substructure during constructive substructure search,
	Bond Properties
bond order	Bonds can be specified to have orders 1, 2, 3 ... for single, double, triple ... Default is single. A bond order of "any" allows any possible bond order.

B.2. Illustration of Method—Constructive Substructure Search. GENOA begins its computations with limited chemical knowledge, consisting only of the atoms C, N, O, H and their standard valences and the reserved atom name X, which represents "any" atom type. All additional chemical information about a structural problem must be supplied to the program as outlined in the next section. However, GENOA does possess considerable knowledge of how structural information can be manipulated in the computer, including substructure searching, detection and use of symmetry, and constructive procedures for piecing together the substructural constraints. These aspects of GENOA are discussed separately. (See paragraph at the end of the paper about supplementary material.)

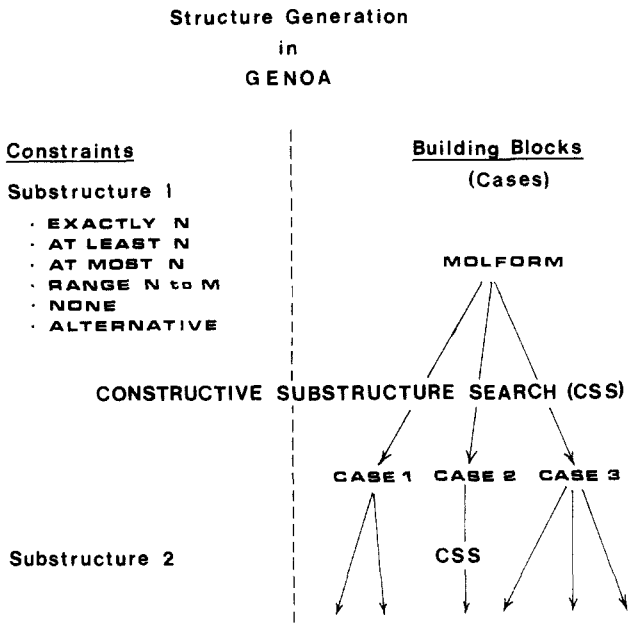


Figure 1. Basic steps used by GENOA to generate structural candidates for an unknown given a molecular formula and a set of substructural constraints (substructure 1, 2 ...). Given a substructure together with a desired range of occurrence of that substructure in final structures, the constructive substructure search algorithm constructs new partial structures (cases) from old cases obtained based on previous substructures.

The basic steps in using GENOA to generate structures for a structure elucidation problem are described in Figure 1.

The structure-generation procedure is stepwise, beginning with the atoms in the molecular formula as building blocks. The molecular formula can be regarded as "pot" of atoms from which GENOA will remove those atoms required to construct structures based on the specified constraints. As an example, consider the use of GENOA in determining candidate structures for the cembranolide derived from lemnalialoside[18] mentioned previously.[3] The molecular formula of the unknown is $C_{20}H_{34}O$.[18] This material is the reduced aglycon portion of lemnalialoside. The aglycon, $C_{20}H_{32}O$, possessed an aldehyde functionality which was reduced to a hydroxyl group[18] in the compound discussed below.

Constraints are supplied to GENOA as named substructures, one at a time. A given constraint may request several copies of the specified substructure (Table II) or may be an ALTERNATIVE constraint which requires the presence of at least one of the alternative substructures in each resulting structure. Specification of the first required substructure together with the desired range of occurrence (Figure 1) causes GENOA to obtain from the molecular formula ("MOLFORM") those atoms sufficient to construct the substructure, or multiple copies of the substructure if more than one was requested. A computer representation of the substructure(s) is actually constructed by GENOA, using the "constructive substructure search" algorithm. (See paragraph at the end of the paper about supplementary material.) For the cembranolide, assumption of the skeleton CEMBRANE (11) based on biogenetic considerations,[18] with potential bonding sites as specified in constraint C-1 (12), leads GENOA to construct the skeleton 12 as described in Figure 2. In this problem

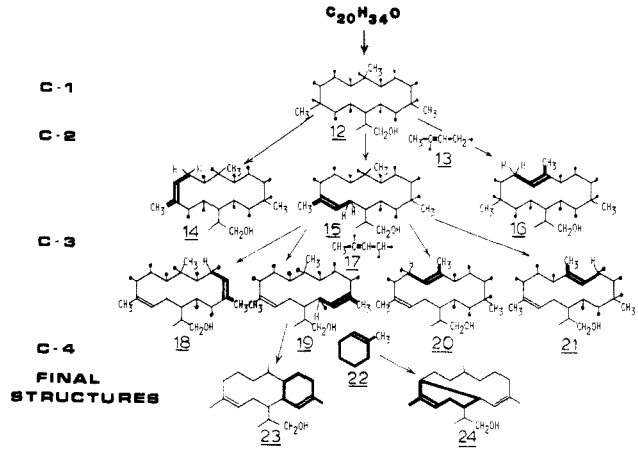
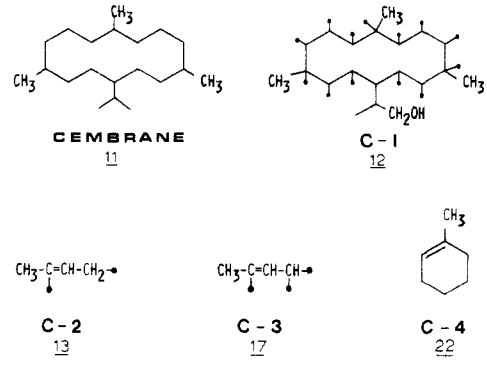


Figure 2. GENOA's stepwise expansion of the molecular formula $C_{20}H_{34}O$ to final structures given constraints C-1–C-4 (**12**, **13**, **17** and **22**). Complete results are shown only for **15** and **19**, leading to two of the seven final structures (see text for details).

seven hydrogens and three degrees of unsaturation (rings plus double bonds) remain to be allocated to **12**.



In general, there may be more than one way to satisfy the given constraint; each way represents a separate "case" (Figure 1) in which the requested number of copies of the desired substructure has been constructed and saved for the next step. A case is usually a partial structure, consisting of several components including disconnected substructures and perhaps several individual atoms not yet attached to specific substructures. A case may be a complete structure if sufficient constraints have been supplied to specify complete structures.

Each constraint subsequent to the first is handled as indicated schematically in Figure 1. Each case from the previous step (an "old" case) is examined by GENOA and the new substructure(s) is constructed in all possible unique ways, using whatever combination of atoms and partial structures in the old case is required to fulfill the new constraint. During these computations all possible overlaps of the new substructure with elements of the old case are considered explicitly by constructive substructure search (CSS). In general, for each old case there are several ways to fulfill the new constraint, each of which represents a "new" case to be passed to the next step (Figure 1, Figure 2). For the cembranolide, constraint C-2 (**13**), obtained from ^1H NMR,[19] can be incorporated in three unique ways into **12** as indicated in Figure 2 in bold face, yielding three new cases **14**–**16**. As described in the section on algorithms (see paragraph at the end of the paper about supplementary material), CSS works by noting that the methyl group of **13** can be matched to several methyl groups in **12**.

(18) C. Charles, License of Chemical Science Thesis, Faculty of Sciences, Free University of Brussels, 1974.

(19) C. Cheer, unpublished results.

However, 13 requires that the methyl be attached to a double bond. CSS notes that such double bonds do not exist in 12 but that the constraint can be fulfilled by *constructing* the required double bond. The terminal CH_2 group is then constructed by allocating one of the remaining hydrogen atoms to the appropriate carbon on the newly constructed double bond (see Figure 2). During these constructive procedures the symmetry of both the constraint and the case is taken into account. Thus, GENOA recognizes the symmetry of 12 so that *only* 14–16, and no symmetry duplicates, are constructed.

Construction proceeds in this fashion until all available constraints are incorporated. For the cembranolide, specification of constraint C-3 (17), obtained from ^1H NMR data,[19] yields several new cases, including the four cases 18–21 obtained from old case 15 as shown in Figure 2. Similarly, as may be verified by inspection, three new cases are obtained for each of 14 and 16. Finally, specification of constraint C-4 (22), obtained from chemical degradation experiments,[19] yields a set of seven new cases, two of which, 23 and 24, obtained from old case 19, are shown in Figure 2. Cases 20 and 21 yield one new case each, while no new cases can be obtained for 18 because there is no way to construct a six-membered ring containing a double bond. (Remember that the terminal >CH– group of 17 when built into 15 forces the corresponding >CH– group in 18–21 to participate in closing of the six-membered ring required by 22. Otherwise the characteristics of this problem would force it to be a $-CH_2-$ group which is illegal.) These seven cases are in fact the complete set of final structures which are structural candidates for this example, because the only problem remaining after specification of C-4 is allocation of the only remaining atoms in the "pot", six hydrogen atoms, to six remaining free valences. GENOA detects this trivial allocation problem and performs it automatically.

During CSS any old case in which a required number of a new substructure cannot be built is discarded because it can never lead to a legal final structure. This is in fact the situation with case 18 (Figure 2). This case is discarded from further consideration because it cannot incorporate the next constraint, 22. If the constraint represents a forbidden substructure (range of occurrence "none", Table II), then each old case is simply tested for the presence of the substructure and discarded if an instance is found. In addition, at each step, after incorporation of a constraint, GENOA checks to ensure that *all previous constraints are still met*. This involves checking each case to ensure that the specified *maximum* number of occurrences of a constraint has not been exceeded by accidental construction of additional copies. This check includes testing constraints whose range of occurrence was "none".

Consider a further illustrative example which schematically illustrates how GENOA considers potential overlapping substructures in a problem where many different overlaps are possible. The problem is presented in Figure 3. Assume that previous constraints have led to several cases, one of which is the "old" case, consisting of a cyclopropane ring (25) and a branched oxirane (26) together with three remaining carbon and nine remaining hydrogen atoms as shown in Figure 3. Assume further that the bonding sites on the three- and six-membered rings may be attached to any atom, including hydrogen. The new constraint 27 (assume at least one is required) consists of a chain of three carbon atoms, methine, methylene, methine. GENOA detects the symmetry of 27 (and 25) and employs it in subsequent construction, as discussed in the section on algorithms. (See paragraph at the end of the paper about supplementary material.)

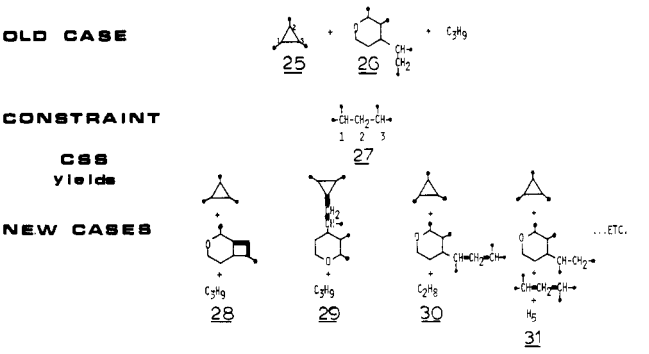


Figure 3. Illustration of the algorithm for constructive substructure search. Four (28–31) of the 17 new cases are shown to demonstrate various ways of overlapping constraint 27 with elements of the old case.

Constructive substructure search yields 17 new cases from the old case, four of which, 28–31 are summarized in Figure 3. The substructure 27, as it was incorporated, is indicated in bold face in 28–31. The first new case, 28, is obtained by forming an edge-fused 6,4 ring system by cyclizing the side chain on the six-membered ring in 26 in the old case. The second, 29, satisfies the constraint by joining the two partial structures 25 and 26 as indicated. The third new case, 30, satisfies the constraint by extending the side chain on 26, using one of the remaining carbon atoms. The fourth new case, 31, results from construction of the substructural constraint using all three previously unassigned carbon atoms. Note that new cases 28 and 29 are constructed by assuming *complete* overlap of the new substructure with atoms in partial structures previously constructed. Case 30 is constructed based on *partial* overlap, while case 31 is constructed from remaining atoms with *no* overlap with previous substructures.

The number of cases can potentially grow very rapidly in a real structure elucidation problem. However, in most problems it does not because of two factors. (1) Constraints are applied to a problem one at a time. The set of new cases from a single constraint is immediately available for examination at a computer terminal. These cases can be checked for the presence of undesired partial structures. Unwanted cases can be removed (by implementing a new constraint(s) which forbids the presence of undesired structural features, or by using the FORGET command, Table I) at that point to prevent them from being carried any further in the procedure. This is a typical example of the interactive nature of GENOA and its predecessor CONGEN, whereby the structural chemist works closely with the program to guide it to a set of structural candidates which fulfill all desired and undesired structural features.

(2) GENOA, during specification of constraints, never expands existing partial structures and remaining atoms into complete structures. *The only structural detail present during this part of the procedure is that which results from the input substructural constraints and all their possible overlaps.* Although each case might represent hundreds or thousands of complete structures if all the pieces were assembled, such assembly is not carried out until explicitly requested by the chemist using the GENERATE command (Table I).

B.3. Use of GENOA. The sequence of commands one would normally use for analyzing a structural problem is presented in this section to convey some flavor of the actual synergism between an investigator and the program.

However, commands for any problem may be issued in any consistent order. GENOA responds with helpful error messages if an inappropriate command is issued.

Step I. Definition of Structural Problem. All "chemistry" relevant to the problem must be given to the program prior to computations which require, for example, specific named substructural constraints. This chemical description includes the molecular formula, names and valences of any new atom types, and names and definitions of inferred substructures. There are two mechanisms for providing such information, the GENOA commands SEARCH and DEFINE (Table I). The SEARCH command is used to retrieve predefined substructures from libraries to avoid time-consuming redefinition of common fragments. The DEFINE command is used to specify the molecular formula and to define new atoms and substructures particular to the new structural problem. Definition of substructures is via an extension of the teletype-oriented structure editor described previously.[20] With this editor, substructures of arbitrary complexity can be defined quickly and visualized at the computer terminal.[21]

Step II. The CONSTRAINT and ALTERNATIVE Commands. The CONSTRAINT command is issued once for each substructure (specified by a name associated with the substructure), including the desired range of occurrence (Table II). The ALTERNATIVE command is used if there is ambiguity, e.g., more than one possible substructure inferred from a spectral signature. The constructive substructure search algorithm is applied for the substructure (or set of alternatives). Repeated use of these commands will normally be interspersed with DRAW and SHOW, below, in order to examine selected cases resulting from application of the previous constraint(s).

Step III. The SAVE Command. Results can be saved on a named disk file at any point, either for protection against computer failure or to save partial results for later analysis.

Step IV. The DRAW and SHOW Commands. The DRAW and SHOW commands are used, normally after each new constraint, to visualize structural features of new cases and to monitor the progress and history of the computations.

Step V. The EXIT and RESTORE Commands. When currently available constraints have been applied, the EXIT command can be used to leave GENOA, including saving the current status of the computations in a computer file. At any later time when new structural information is available, the RESTORE command is used to restore the file, thus returning GENOA to the exact point at which the problem was left on exit.

When no more constraints are available and when complete structures are required in order to carry out further tests to differentiate among them, it is possible to construct complete or final structures from all cases. An important result of the procedure for constructive substructure search is that *within each case all partial structures are guaranteed to be nonoverlapping.* This is important for future developments of GENOA and CONGEN because the existing mechanism for structure generation based on nonoverlapping substructures in CONGEN can be used to construct final structures for each case. However, the necessary interface between the two programs is not written. As a temporary alternative, final structures from all of the cases are constructed with a simple structure generator within GENOA (GENERATE command, Table I). This structure generator is relatively unsophisticated compared to its analogue in CONGEN. However, it accomplishes the desired task of generating all structures while ensuring that no duplicates result in the final set of structures. Although it is possible to generate structures at any point in GENOA, including from the original molecular formula, such a procedure would be foolhardy until a reasonable number of constraints is specified to restrict the problem.

The last step in execution of the GENERATE command is to pass control of the problem over to CONGEN. This interface to CONGEN gives the chemist access to modules of CONGEN which are not included as part of GENOA. Thus, he or she can continue with examination of the structures, application of additional constraints with the PRUNE command[5] and application of additional post-processing functions to the structures, including exploration of potential stereoisomers[22,23] under stereochemical constraints,[24] or prediction of spectral properties and rank ordering of candidate structures.[15,16,25]

Table IV. Substructural Constraints, Their Range of Occurrence, the Source of the Constraint and the Resulting Number of Cases Obtained by GENOA for the Structure of Uvidin A[26]

substructure name	range of occurrence	source of constraint	no. of cases
OH	exactly 1	IR, acetylation, ^{1}H NMR	1
CO	at least 1	IR	1
T-METHYL	exactly 4	^{1}H NMR	7
TBU	none	^{1}H NMR	5
ACETYL	none	^{1}H NMR	3
ME-C-O	exactly 1	^{1}H NMR	4
ISOL	exactly 1	^{1}H NMR	27
ISOLCH	exactly 2	^{1}H NMR	1532
CSP2	exactly 1	^{13}C NMR	1532
CH2	exactly 4	^{13}C NMR	100
BASICATOMS	at least 1	^{13}C NMR	100
TRIEPO	exactly 1	^{13}C NMR	33
CYCLOPROPYL	none	^{1}H, ^{13}C NMR	33

C. Results

In this section we describe the use of GENOA to analyze a structure elucidation problem taken from the literature. This particular problem is in no respect special; it is merely one of many used to test the program. This example is meant to illustrate several points mentioned earlier with regard to special features of the program. Notable in this regard are the following. (1) Substructural information is supplied to GENOA *exactly* in the order in which it was presented in the paper, with no need for manual analysis to determine potential overlapping substructures. This illustrates the utility of GENOA in *prospective* analyses of structures in that information can be supplied to the program as it is collected in the laboratory, with the program used as an aid in determining the effects of each additional piece of information on the potential structural variety and in guiding additional experiments. (2) In places where assumptions were made regarding major features of the structure we illustrate how GENOA can be used as an exploratory tool to test other assumptions. (3) The problem yields several structural candidates which can

(20) T. H. Varkony, R. E. Carhart, D. H. Smith, and C. Djerassi, *J. Chem. Inf. Comput. Sci.*, **18**, 168 (1978).
(21) R. E. Carhart, *J. Chem. Inf. Comput. Sci.*, **16**, 82 (1976).
(22) J. G. Nourse, *J. Am. Chem. Soc.*, **101**, 1210 (1979).
(23) J. G. Nourse, R. E. Carhart, D. H. Smith, and C. Djerassi, *J. Am. Chem. Soc.*, **101**, 1216 (1979).
(24) J. G. Nourse, D. H. Smith, R. E. Carhart, and C. Djerassi, *J. Am. Chem. Soc.*, **102**, 6289 (1980).
(25) N. A. B. Gray, R. E. Carhart, A. Lavanchy, D. H. Smith, T. Varkony, B. G. Buchanan, W. C. White, and L. Creary, *Anal. Chem.*, **52**, 1095 (1980).

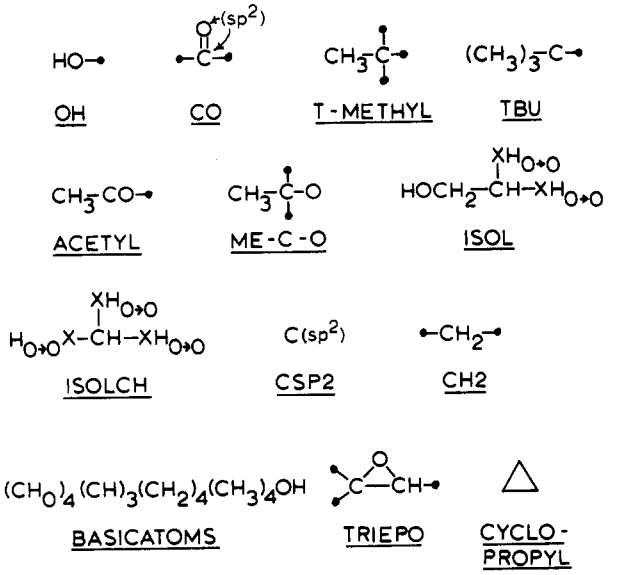
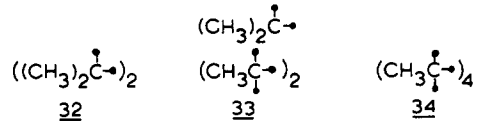


Figure 4. Substructural definitions of the constraints used in structure elucidation of uvidin A.[26] All bonds with an unspecified terminus (free valences) are to nonhydrogen atoms. Atoms named "X" represent *any* nonhydrogen atom. The subscript "0→0" on hydrogen atoms (e.g., in substructure ISOL) represents a hydrogen range of exactly 0, or *no* hydrogen atoms.

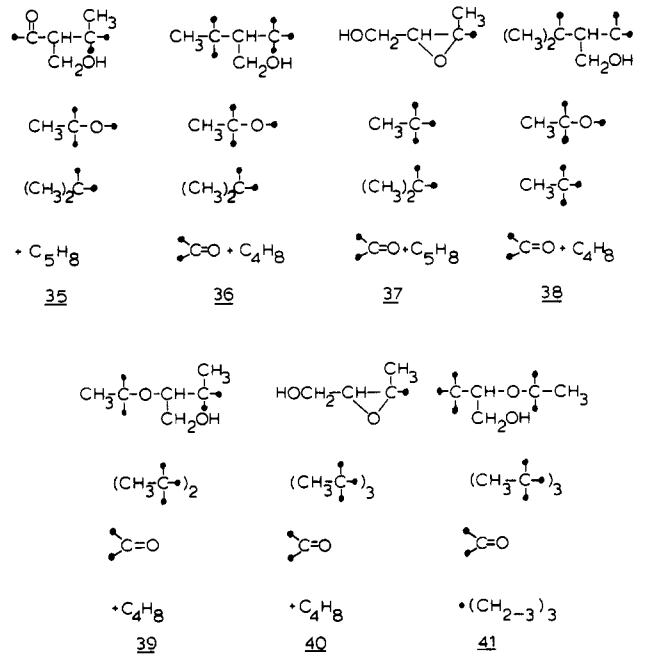
subsequently be tested by other modules of the program, specifically, prediction of spectral properties and ranking of candidates based on mass[25] spectral data.

The problem concerns the structure of uvidin A, published recently by De Bernardi et al.[26] Uvidin A has the molecular formula $C_{15}H_{24}O_3$. We present in Table IV the results obtained by GENOA for each successive substructural constraint. In Figure 4 we show the actual definitions of substructures whose names correspond to those given in Table IV. Table IV also indicates the source of data for each inference. There are obviously many alternative ways of expressing the structural inferences, some of which would be more efficient computationally. For reasons stated previously, we chose precisely the order given by De Bernardi et al.

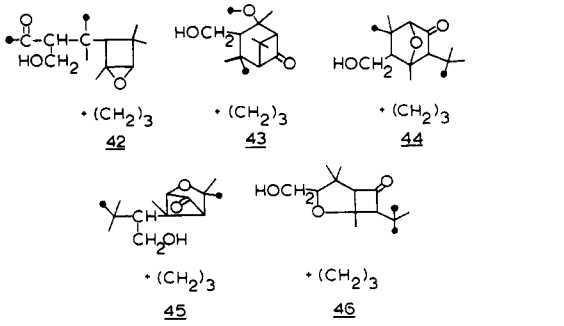
The first two constraints, OH and CO, record the presence of exactly one hydroxyl and at least one carbonyl functionality, leading to only one case for each constraint (Table IV and Figure 4). Note that at this point in the problem the investigator can only specify *at least one* carbonyl, because the degree of unsaturation and the number of oxygens may allow two. The constraint T-METHYL species exactly four tertiary methyl groups. Investigation of the seven resulting cases revealed that two possessed *tert*-butyl groups and two possessed acetyl groups; these four cases were discarded by specifying *none* of the substructures TBU and ACETYL. The three resulting cases distribute the four tertiary methyl groups in the three possible ways **32–34**.



Specifying that exactly one of the tertiary methyls must be geminal to an oxygen atom, ME-C-O, yields four new cases. Old case 1 (**32**) fails to yield new cases because it must yield two such methyls. Two cases are obtained from each of **33** and **34**, resulting from connection to either the OH group or the remaining oxygen in each case. Specifying that the OH group must be attached as shown in the substructure ISOL (Figure 4) leads to 27 new cases, examples of which are shown as **35–41** to indicate the variety of structural environments possible in the absence of additional information.



The ^1H NMR data revealed the presence of exactly two isolated methine groups, substructure ISOLCH, Figure 4. This constraint produces a large number of structural possibilities (Table IV) primarily because nothing has yet been stated about the possible environments of the methines. Consideration of the ^{13}C NMR spectrum, however, rapidly cuts down the number of possible cases as illustrated by the next sequence of constraints in Table IV. Specification of a single sp^2 hybridized carbon has no immediate effect on the number of cases (CSP2 constraint, Table IV) because there has as yet been no opportunity to construct any additional carbons of this hybridization. However, this constraint serves to prevent future construction of additional sp^2 carbons as discussed in the methods section. The ^{13}C spectrum also revealed exactly four CH_2 groups, and this constraint alone dramatically reduces the number of cases from 1532 to 100. The constraint BASICATOMS is a simple expression of the degree of each of the carbon atoms from the ^{13}C spectrum. This does not reduce the number of cases; previous constraints have, explicitly or implicitly, specified the degree of each carbon atom. Examples of cases (out of the total of 100) at this stage in the problem **42–46**. The structures are now beginning to take definite shape.



(26) M. De Bernardi, G. Mellerio, G. Vidari, and P. Vita-Finzi, *J. Chem. Soc., Perkin Trans. 1*, 221 (1980).

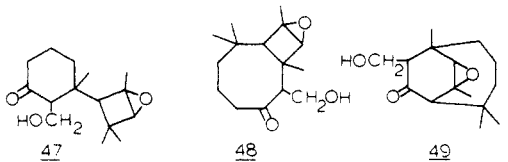
Table V. Various Assumptions Concerning the Structure of Uvidin A

structural hypothesis	resulting no. of cases	no. of final structures
no additional assumptions	33	33
decalin ring system	6	6
three isoprene units linked head-to-tail	4	4

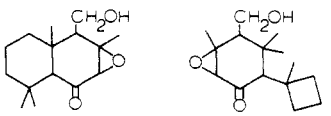
Spectral data revealed the presence of a trisubstituted epoxide functionality, TRIEPO, Figure 4. Specification of this constraint yields 33 new cases (Table IV); 67 of the previous 100 could not incorporate this constraint. For example, **42** already meets the constraint, but no legal case can be constructed from **43–46**. The final constraint, CYCLOPROPYL, was specified to express the fact that there was no evidence for a cyclopropyl ring in the structure. This has no immediate effect on the number of cases (Table IV) but will prevent the construction of such rings in subsequent steps.

These constraints represent the sum total of the "hard" facts gleaned from the literature about the structure. At this point in the structural problem, De Bernardi et al. made the assumption that the presence of two fused six-membered rings would explain the level of unsaturation and accommodate the functional groups.[26] Using GENOA we can test not only this hypothesis but several others as well. We investigated the hypotheses about the structure which are summarized in Table V.

If no other assumptions are made, then the total number of structural possibilities at this point in the problem is obtained by issuing the GENERATE command (Table I) to GENOA. For this problem, each of the 33 cases (Table IV) yields only a single, unique structure, resulting in 33 final structures (Table V). Examples illustrating the structural variety are **47–51**; several edge-fused, disjoint, and bridged ring systems are possible. Structure **50** is the assigned structure of uvidin A.[26]

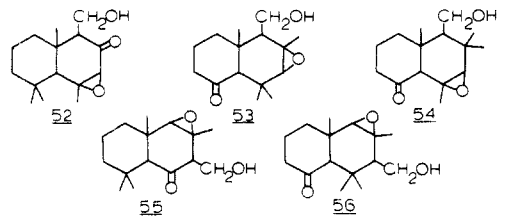


47 48 49



50 51

De Bernardi et al. state that several structures are possible if the existence of a decalin ring system is assumed.[26] The exact number can be constructed by GENOA by first defining a substructure comprised of two edge-fused six-membered rings and then using that substructure as a constraint applied to the 33 cases obtained previously (Table IV). There are only six cases which can be constructed, and each case leads to a single, unique structure, as summarized in Table V. The six structures are **50** and **52–56**.



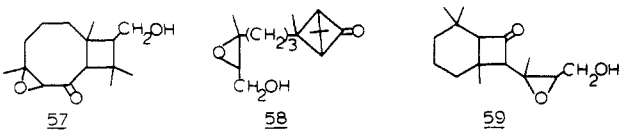
52 53 54 55 56

Table VI. Results of Comparison of Predicted and Observed Mass Spectral Data for the Structural Candidates for Uvidin A under Various Structural Hypotheses (Structure 50 is the Assigned Structure for Uvidin A)

structural hypothesis	no. of structures	rank order[a] of structure 50 based on mass spectrum: partial	complete
no additional assumptions	33	7 (6)	5 (2)
decalin ring system	6	1 (3)	1 (2)
three isoprene units linked head-to-tail	4	2	1

[a] Numbers in parentheses are the numbers of structures with the same score as **50**.

Another assumption which might be made is that the unknown possesses an unrearranged isoprenoid skeleton comprised of three isoprene units linked head-to-tail. GENOA can test this assumption by using a constraint a substructure expressing this linkage. Given the 33 cases of Table IV, there are only four structures which can be assembled based on this assumption, **50** and **57–59**.



57 58 59

Several additional tests of the final structures obtained under various assumptions can be made by using the post-generation "structure-checking" program.[27] One method of evaluating structural candidates is to compare them against a library of previously observed structures of skeletons. For uvidin A we can test the 33 structural candidates against a library of known bicyclic sesquiterpene skeletons.[28,29] Only one structure, **50**, contains a skeleton in this library (the drimane skeleton).

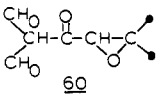
One of the methods used by De Bernardi et al.[26] to locate the position of substituents on the skeleton was mass spectrometry. The prominent peaks at m/z 123 and 109, the former representing cleavage of the C-9,10 and C-5,6 bonds (see **50**) with concomitant hydrogen transfer, the latter representing the same cleavage followed by loss of a methyl group,[30] indicate the absence of substitution in ring A. Given the complete set of structural candidates and the low-resolution mass spectrum of the unknown,[26] we can apply this test automatically. We use a method of prediction of mass spectra for each candidate, followed by rank-ordering of the candidates based on how well the predicted spectrum agrees with the observed spectrum.[25] In addition, we can apply a more rigorous test by using all peaks in the observed spectrum and looking for the candidates which serve best to explain the *complete* spectrum.

We present in Table VI the results obtained for these tests by indicating the ranking of the assigned structure (**50**) in each of the sets of candidate structures from Table V. Considering row one of Table VI, the assigned structure is ranked seventh among a group of 14 structures which

(27) C. Djerassi, D. H. Smith, and T. H. Varkony, *Naturwissenschaften*, **66**, 9 (1979).
(28) T. K. Devon and A. I. Scott, "Handbook of Naturally Occurring Compounds, Vol. II, Terpenes", Academic Press, New York, 1972.
(29) D. H. Smith and R. E. Carhart, *Tetrahedron*, **32**, 2513 (1976).
(30) S. C. Sharma, J. S. Tandon, H. Uprety, Y. N. Shukla, and M. M. Dhar, *Phytochemistry*, **14**, 1059 (1975).

should be taken as candidates based on the distribution of the scores,[25] considering only the partial spectrum m/z 123 and 109. The ranking improves slightly by considering the complete spectrum. Under the assumptions of a decalin skeleton, three structures, **50**, **52**, and **55** are ranked equally well; all are unsubstituted in ring A. However, only **50** and **55** provide the most plausible rationalization of the complete spectrum (Table VI). Under the assumption of the presence of a head-to-tail isoprene skeleton, **50** is ranked behind **58** considering the partial spectrum because the latter structure provides a "simpler" explanation of the spectrum (only single-bond cleavages are required to yield the selected ions). However, when the complete spectrum is considered, **50** again provides the most plausible explanation.

A subsequent chemical experiment, borohydride reduction of the carbonyl group revealed the environment of that functionality as expressed in substructure **60**.[26]



60

Applying this constraint to the final structures summarized in Table V leads to three possibilities for the set of 33, **50**, **51**, and **57** only two of which, **50** and **51**, provide good explanations of the mass spectrum. The set of six (Table V) is reduced to two structures, **50** and **57**, the former already being top-ranked on the basis of its mass spectrum (Table VI).

Taken together, the above results are strongly supportive of **50** as the correct structure for uvidin A. De Bernardi et al.[26] assigned **50** as the correct structure after additional chemical transformations to a previously characterized structure.

Analysis of this problem with GENOA, including exploration of all alternative hypotheses and spectrum prediction and ranking, required approximately 1 h of central processor time (see Experimental Section).

D. Conclusions

The description of GENOA has been necessarily brief. It is simply not possible to describe adequately a complex program in the chemical literature. Nor is it possible to select examples of applications which illustrate the wide variety of approaches which can be taken to solve a particular problem or which illustrate the generality of the program to many organic structural problems. The following paragraphs summarize some of the features of the program which were not discussed previously in an attempt to at least touch upon these points.

GENOA has built into it limited chemical knowledge. However, the program does possess general knowledge of chemical valence, aromaticity, and hybridization and how these chemical concepts can be utilized to avoid construction of nonsense chemical structures. The program is not limited to particular atom types, nor even to atoms in the periodic table. It would be possible to add a variety of general constraints on molecular structures to avoid particular systems which most organic chemists would consider unreasonable. However, our experience with the CONGEN program and with GENOA has shown that there are such a wide variety of potential applications that built-in chemical knowledge would prevent effective use of the program by significant segments of the chemical community. Therefore, we depend on a friendly, tolerant interface to the program so that an investigator using GENOA can make use of a variety of schemes for defining general chemical constraints of use in his or her own applications and, subsequently, invoke such constraints automatically in a problem.

A singular advantage of GENOA is its utility as a "dry" laboratory to explore alternative structural hypotheses, such as was done for structure of uvidin A. At any point in a problem an investigator can specify different assumptions about the structural characteristics of an unknown, either through explicit use of the ALTERNATIVE command or by subdividing the problem along different pathways to solution. In this way the net effects of every assumption can be determined by examining the structural variety which results. Experiments can then be planned to differentiate among the hypotheses with a clear idea of the effects of each result on the overall problem. In this way both valuable time and sample can be conserved. This alone seems sufficient justification for utilizing GENOA (and CONGEN) as routine tools in all but the most trivial structure elucidation problems.

All computer-based approaches to generation of structures attempt to solve the same problem. All seek to establish the interconnection of atoms in a molecular formula in order to obtain complete chemical structures which obey given constraints. *In terms of the actual computations performed, however,* GENOA *represents a completely novel approach to the problem.* Given a connection table representation of the molecular formula, GENOA constructs structures not by arbitrary interconnection of atoms, each of which must be tested against the constraints, but by building in the constraints in all ways as the computations progress. *In many problems this approach dramtically reduces the necessary computation time.*

Our brief experience with GENOA has convinced us that the advantages for such an approach (see the introduction) are real and important to the utility of computer-assisted approaches to structure elucidation. The ability to interface GENOA directly to programs for spectral data interpretation in a straightforward way has already been demonstrated for both mass[1] and ^{13}C NMR[15] data. Equally important is the completely natural way of expressing substructural constraints to the program. The fact that they may be specified as they are obtained experimentally, and that GENOA will always ensure that potential overlaps have been considered, means that the program can be a real aid during the course of solving an unknown structure.

The current program has several limitations, some of which were mentioned previously. One limitation is, of course, accessibility. A computer program is useful to chemists only if they can access it. In the Experimental Section we mention various alternatives for access. One can foresee that programs such as GENOA will be available on the next generation, microprocessor-based laboratory systems, but that has not yet happened because of the newness of the microprocessor technology. Other significant limitations involve chemical concepts such as tautomerism and aromaticity. The former is not treated by GENOA; both tautomers of a resonating pair will be produced. The latter is treated only partially; meta- and para-bridged, implausible aromatic systems are constructed unless they are explicitly excluded. Another limitation arises *because* we free the scientist from detailed consideration of a structural problem. It is possible to state certain problems in such a way that the number of intermediate cases numbers in the thousands or tens of thousands. To cite a relatively simple example, if GENOA were given the basic cholestane skeleton as a constraint, followed by information from the ^{13}C spectrum on the

number of methyl, methylene, methine, and quaternary carbons, large numbers of cases result. This comes about because GENOA is forced to construct all cases in what is a large "labeling" problem; there are many ways of allocating the carbons in various ways about the skeleton in the absence of other constraints! It is difficult to anticipate such situations and, in general, difficult even to predict the size of problems in terms of eventual numbers of cases or final structures. These limitations represent areas where further program development is needed for GENOA to become a program of true utility to a broad community of investigators.

The completely thorough and unbiased exploration of structural possibilities for an unknown carried out by GENOA (and CONGEN) suggests another useful application of the programs. There are several chemical journals which devote significant space to reports of structure elucidation of new natural products. We suggest that our computer programs could be made a useful adjunct to preparing papers for such publications. For those unknowns where structural assignment is based on X-ray determination, unambiguous synthesis, or relation to previously characterized structures, the programs are obviously not necessary. For the remaining problems, however, it would be quite simple to determine, based on reported spectroscopic and chemical data and the structural inferences derived therefrom, whether or not a proposed structure was in fact the only structure allowed by the data. If other structural possibilities were found which could be eliminated by additional experiments, reported structural assignments could be made on much firmer ground.

E. Experimental Section

GENOA is implemented in the ALGOL-like BCPL programming language[14] on a Digital Equipment Corporation KI-10 computer at the SUMEX-AIM facility at Stanford University. The program is available to an outside community of collaborators via a nationwide computer network (TYMNET), to the limits of available resources. Export of the program to other DEC PDP-10 or PDP-20 systems is possible. Information on the possibility of export to the other types of computers or on additional algorithmic details can be obtained from us.

Acknowledgment. We thank the National Institutes of Health for their generous financial support (RR-00612). This work was supported in part by a grant from the United Kingdom Science Research Council (B/RF/4955, to N.A.B.G.). Computer resources were provided by the SUMEX facility at Stanford University under separate NIH support (RR-0785). Special thanks to Claire Cheer for experimental work used to derive the constraints for the cembranolide derived from lemnalialoside and to Ben Tursch and Claude Charles for a sample of this compound.

Supplementary Material Available: This material (23 pages, including six tables, two figures, structures, and references) describes in some detail the key algorithms in GENOA which are necessary for constructive substructure search. Ordering information is given on any current masthead page.

Notes

DAST-Induced Epimerization of a 2-(Acetoxymethyl)myoinositol

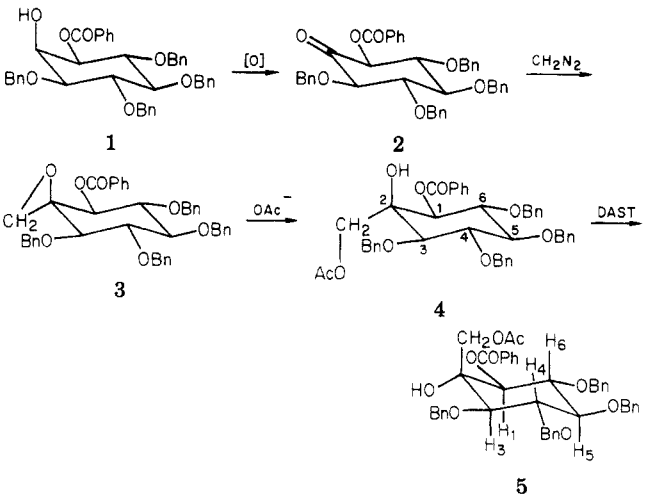
Shu Shu Yang* and Thomas R. Beattie

Merck Sharp & Dohme Research Laboratories, Division of Merck & Co., Inc., Rahway, New Jersey 07065

Received August 26, 1980

Diethylaminosulfur trifluoride (DAST)[1] has recently been introduced as an excellent reagent for converting ROH to RF. More recently application has been extended to steroids[2-4] and carbohydrates.[5,6] In contrast to the considerable interest in synthesis and chemistry of fluorinated carbohydrates, no fluoro analogues of cyclitols have been reported. While pursuing synthetic studies aimed at preparation of ring-fluorinated inositols,[7] we observed an unusual epimerization. When DL-2-*C*-(acetoxymethyl)-1-*O*-benzoyl-3,4,5,6-tetra-*O*-benzylmyoinositol (**4**) (prepared from DL-1-*O*-benzoyl-3,4,5,6-tetra-*O*-benzylmyoinositol (**1**)[8,9] in three steps as shown in Scheme I) was treated with DAST in methylene chloride at 0 °C, a new

Scheme I[10]



Bn = CH_2Ph

(1) W. J. Middleton, *J. Org. Chem.*, **40**, 574 (1975).
(2) S. S. Yang, C. P. Dorn, and H. Jones, *Tetrahedron Lett.*, 2315 (1977).
(3) B. L. Onisko, H. K. Schnoes, and H. F. DeLuca, *Tetrahedron Lett.*, 1107 (1977).
(4) J. L. Napoli, M. A. Fivizzani, A. H. Hamstra, H. K. Schnoes, H. F. DeLuca, and P. H. Stern, *Steroids*, **32**, 453 (1978).
(5) M. Sharma and W. Korytnyk, *Tetrahedron Lett.*, 573 (1977).
(6) T. J. Tewson and M. J. Welch, *J. Org. Chem.*, **43**, 1090 (1978).
(7) Synthesis of 1-fluoro-1-deoxyscylloinositol will be reported elsewhere.
(8) D. E. Kiely, G. D. Abruscato, and V. Baburao, *Carbohydr. Res.*, **34**, 307 (1974).
(9) V. P. Schevchenko, T. Yu. Lazurkina, Yu. G. Molotkovsky, and L. D. Bergelson, *Bioorg. Khim.*, **3**, 252 (1977).
(10) For convenience, D-myoinositol configurations were used to denote DL-myoinositol throughout Schemes I–III.

0022-3263/81/1946-1718$01.25/0