The Role of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Enhanced Structure Elucidation[1]

Bradley D. Christie[†] **and Morton E. Munk***

Contribution from the Department of Chemistry, Arizona State University, Tempe, Arizona 85287-1604. *Received October 15, 1990*

Abstract: Two-dimensional NMR spectroscopy is a powerful structural probe, but often in structure elucidation some of the most significant information derived therefrom—long-range carbon-hydrogen signal correlations—is more commonly used to distinguish between proposed alternatives later in the process than to construct compatible structures very early in the process. It is the uncertainty about the number of intervening bonds between correlated atoms that introduces substantial overall ambiguity in a set of such correlations and complicates their interpretation. A strategy linking a versatile, computer-based procedure for the interpretation of 2-D NMR-derived signal correlations (INFER2D) to a recently developed, structure-reduction-based structure generator (COCOA) is described which has demonstrated considerable promise in *prospectively* utilizing such ambiguous data in directly constructing compatible molecules, even in the presence of molecular symmetry. The software, although at a early stage of development, already possesses some problem-solving capabilities which are illustrated with several examples. Current program scope and limitations are described.

Introduction

Two-dimensional NMR spectroscopy is a powerful structural probe in the elucidation of the structure of complex organic compounds because of its ability to reveal carbon-carbon connectivity.[2] Since the method may not uncover the complete carbon skeleton of a compound—perhaps because of experimental difficulties (e.g., in 2-D INADEQUATE) or lack of sufficient coupled nuclei (e.g., in COSY)—it complements rather than replaces other spectroscopic techniques. Thus, 2-D NMR is ideally suited as *one* spectroscopic source of structural information in a computer-based structure elucidation system. This paper describes the impact of the addition of 2-D NMR interpretation capabilities to such a system, SESAMI (*s*ystematic *e*lucidation of *s*tructure *a*pplying *m*achine *i*ntelligence), which is currently under development.

SESAMI is successor to CASE,[3] the final version of which included a 2-D NMR interpreter (INTERPRET2D) linked to a version of the structure generator ASSEMBLE.[4] INTERPRET2D was designed to produce units of connected carbon atoms, i.e., discrete substructures, from ^1H-^1H COSY (three-bond H-H correlations)/^1H-^{13}C COSY (one-bond C-H correlations) experiments and from the 2-D INADEQUATE (one-bond C-C correlations) experiment. The observed correlations are entered as pairs of NMR chemical shifts, element type, and number of bonds between correlated atoms. In the interpretation process, input data are first reduced to carbon-carbon signal "connectivity," i.e., one or more units of "connected" carbon signals. The 2-D INADEQUATE experiment gives this information directly, but from ^1H-^1H COSY/^1H-^{13}C COSY data it is derived by identifying carbon signals corresponding to carbon atoms that bear coupled vicinal hydrogens. In the absence of molecular symmetry in the unknown, the interpretation is complete because carbon-carbon signal connectivity is then the same as carbon-carbon atom connectivity. In the presence of molecular symmetry, algorithms based on group theory perceive symmetries consistent of the 1-D ^{13}C NMR data, and for *each case*, convert units of connected carbon signals to one or more discrete substructures. Since more than one compatible symmetry is usual, multiple interpretations are generally produced.

ASSEMBLE2D[4] accepts the discrete substructures produced by INTERPRET2D, and any other structural information entered by the user, and generates all compatible molecular structures. If there is more than one interpretation of the 2-D NMR data, each must be treated as a separate structure generation problem. It is this latter requirement that precludes the utilization of information-rich, long-range ^1H-^{13}C COSY correlations even in the absence of molecular symmetry.

Experiments that correlate nonadjacent hydrogen and carbon atoms often produce many separate correlations, each of which is ambiguous in the sense that it is consistent with at least two, but possibly three different interpretations. In the usual case the number of bonds between the two nuclei is two or three, i.e., *H*·C·*C* or *H*·C·A·*C*, but it can be four, i.e., *H*·C·A·A·*C*, where "A" is any nonhydrogen atom and "·" is any bond. If discrete substructures are required as input to the structure generator, a serious problem arises because the observed set of many ambiguous correlations gives rise to many different sets of unambiguous correlations. Each such set corresponds to a different discrete substructure (or set of discrete substructures), i.e., a different interpretation. In the monochaetin problem described later, there are 22 ambiguous long-range carbon-hydrogen correlations (two or three intervening bonds). This would give rise to 2^{22} (over 4 million) different sets of 22 unambiguous correlations. Clearly, the generation of all of these discrete interpretations of the 2-D NMR data, and the treatment of each as a separate structure generation problem, does not comprise a feasible approach to problem solving. For similar reasons, chemists involved in conventional structure elucidation often use long-range carbon-hydrogen correlations retrospectively to distinguish between plausible alternative structures, rather than prospectively to construct compatible structures.

Only one other direct application of 2-D NMR spectroscopy to computer-based structure elucidation has been reported. CHEMICS has been recently expanded by Funatsu, Susuta, and Sasaki[5] to include the interpretation of data from 2-D NMR spectroscopy. However, only the results of the 2-D INADEQUATE experiment are utilized, and there is no indication that compounds with molecular symmetry can be treated. In the earlier work of Lindley et al.,[6] the structural fragments derived from 2-D NMR experiments were included as input to GENOA, but it was the user who deduced them from the observed data.

Brief Overview of SESAMI

Conventional structure elucidation (computer-unassisted) can be viewed as a process consisting of two separate and distinct stages. The first is time-intensive and multi-step in nature. It involves collecting and interpreting the results of chemical and

* To whom correspondence should be addressed.

[†] Current address: Molecular Design Ltd., 2132 Farallon Dr., San Leandro, CA 94577.

[1] Presented in part at the 1989 International Chemical Congress of Pacific Basin Societies, Honolulu, Dec 12-17, 1989; Paper no. 1, Information Transfer.

[2] Kessler, H.; Gehrke, M.; Griesinger, C. *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 490.

[3] Munk, M.; Shelley, C.; Woodruff, H.; Trulson, M. *Fresenius' Z. Anal. Chem.* **1982**, *313*, 473.

[4] Christie, B.; Munk, M. *Anal. Chim. Acta* **1987**, *200*, 347.

[5] Funatsu, K.; Susuta, Y.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 6.

[6] Lindley, M.; Schoolery, J.; Smith, D.; Djerassi, C. *Org. Magn. Reson.* **1983**, *21*, 405.

0002-7863/91/1513-3750$02.50/0

spectroscopic experiments and is most likely followed by additional experiments and interpretation. The "interpretation" is expressed as structural fragments believed to be part of the unknown. The set of fragments at any point in this process represents a *partial* structure of the unknown. Stage one is complete when the chemist can express a partial structure as *all* of the molecular structures compatible with it. This is generally possible only when the number of *all* structures is sufficiently small to be manageable.

In stage two, the emphasis shifts to a distinction between the alternative structural assignments. In contrast to stage one, stage two is consummated relatively quickly because experienced chemists are proficient at readily identifying the correct structure of an unknown from among a limited number of alternatives. The process usually involves a review of the spectroscopic data in the light of each structure, but may require additional experiments; if the latter is so, the set of structures provides invaluable guidance in their design.

In terms of productivity, the bottleneck is the first stage. If the chemist's initial involvement in the structure elucidation could begin at the *second* stage, that is, with the "shortlist" of plausible alternative molecular structures, much time would be saved and substantially increased productivity achieved. Thus, the goal of the SESAMI project is the creation of computer software capable of directly reducing the collective spectroscopic properties of an unknown to some manageable number of plausible molecular structures compatible with these data. The chemist is still a required player in the structure elucidation process, but is spared its most time-intensive component. The nature of SESAMI follows from the goal: an interactive program for computer-*enhanced* structure elucidation. The major target of the SESAMI system is the compound of complex structure—complex in its skeletal intricacy and functionalization—such as is commonly encountered in compounds of natural origin. Initial software development will concentrate on compounds of up to 50 nonhydrogen atoms. Spectroscopic data will be the exclusive source of structural information since it is believed that the collective spectroscopic properties alone can, if the experiments are thoughtfully selected, be sufficiently information-rich to narrow the compatible molecular structures to a manageable number.

SESAMI seamlessly and efficiently links the tasks of spectrum interpretation and structure generation. It is still under development, not a finished product. However, it already possesses substantial power in solving structure elucidation problems utilizing modern 2-D NMR experiments (see examples described below). Its present organization is shown in Figure 1.

INTERPRET is a two-track spectrum interpretation procedure. On one track, PRUNE, the molecular formula and collective spectroscopic properties are reduced to a shortlist of uniformly sized, precisely defined structural fragments predicted to be present in the unknown. These fragments serve as the structural building units for the structure generator COCOA. They are referred to as ACFs:[7] one-concentrically-layered, atom-centered fragments (e.g., $=\mathrm{CH}—\mathit{CH_2}—\mathrm{O}—$) built of element groups[7] that define an element, its attached hydrogens, if any, and each of the partial bonds by which one element group joins to another (e.g., $-\mathrm{CH_2}-$, $=\mathrm{O}$). PRUNE produces the set of possible structural building units in the unknown (the ACF shortlist) by deleting ACFs from an *exhaustive* list of ACFs. Currently the exhaustive list contains all possible ACFs that can be derived from elements most commonly encountered in natural products: carbon, hydrogen, oxygen, trivalent nitrogen, divalent sulfur, and each of the monovalent halogens. Important structural features excluded by these elemental limitations can be conveniently added as "super" element groups. The nitro group was included in this way. The chemically stable, "exhaustive" ACF list excludes those ACFs that would clearly confer chemical instability on compounds containing them (e.g., $-\mathrm{CH_2}-\mathit{C}(\mathrm{OH})_3$) and currently includes about 5100 ACFs.

PRUNE is modular in nature and currently consists of routines that remove ACFs from the exhaustive list that are not compatible with the molecular formula of the unknown, its ^{1}H and ^{13}C NMR data, and the results of 2-D NMR experiments. (An infrared-based pruning routine is under development.) PRUNE is interactive; the ACF shortlist can be conveniently examined and further pruned in accord with the chemist's insights.

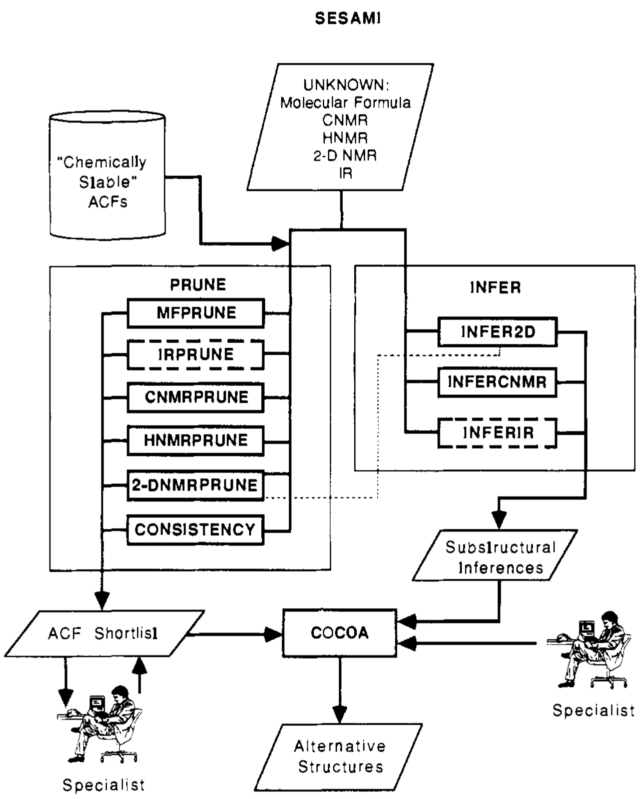


Figure 1. Information flow in SESAMI.

The ACF is too small a structural unit to permit a distinction to be made between each and every one based solely on spectroscopic properties. Thus, in practice, the ACF shortlist will usually contain more invalid ACFs (those not present in the unknown) than valid ones. This presents no major problem to COCOA as most structures containing invalid ACFs are eliminated prospectively during structure generation. Those that are not, lead to plausible, but incorrect alternative structural assignments.

The approach used in producing the set of structural building units (ACF shortlist) recognizes the enormous diversity of structure found in compounds of natural origin and the need for SESAMI to reveal the entire range of structures compatible with the spectroscopic data, *without exception*. If the list of structures presented to the chemist is to exclude no plausible alternative, the ACF shortlist must exclude *no* fragment compatible with the spectroscopic data. It is because of this that the initial list of ACFs on which PRUNE acts must be exhaustive, and that if PRUNE is to err, it is better to retain an invalid ACF than to delete a valid one.

INFER, the second track of INTERPRET, produces the substructural inferences that serve as *constraints* on the structure generation process, thereby limiting the number of plausible alternative structures produced from the structural building units. It likewise consists of separate routines, the output of each of which is one or more substructures predicted to be present or absent in the unknown. No restrictions are placed on the number or the size of the substructures predicted, the degree of ambiguity in expressing them, or the extent to which substructures derived from the same or different routines may overlap. Alternative interpretations of the data may also be produced. Only INFER2D and INFERCNMR, an interpretive ^{13}C NMR library search system,[8,9] are fully operational. (Work on an infrared interpreter,

[7] Munk, M.; Lind, R.; Clay, M. *Anal. Chim. Acta* **1986**, *184*, 1.

[8] Shelley, C.; Munk, M. *Anal. Chem.* **1982**, *54*, 516.

[9] Velu, V.; Munk, M. Manuscript in preparation.

INFERIR, is in progress and other substructural inference makers are in the planning stages.) INFER, like PRUNE, is interactive. The output of INFER may be viewed, edited, and supplemented by any structural information known to the chemist from whatever the source.

The dual output of INTERPRET—structural building units (ACFs) and constraints—is handed directly to the structure generator COCOA, the output of which is displayed in the conventional structural language of the chemist.

SESAMI is both more versatile and powerful than its predecessor, CASE. In large measure this is because of the intrinsic limitations of the *structure assembly* procedure on which CASE's structure generator ASSEMBLE,[10] and other reported structure generators,[11-15] are based. However, SESAMI's structure generator, COCOA, is based on an entirely different concept: structure reduction.[16] Five major features account for COCOA's improved performance over structure assembly procedures: (1) it perceives and uses symmetry information *prospectively*;[17] (2) it prospectively uses potentially overlapping substructural information (structural building units or constraints) directly without preprocessing of any kind; (3) it prospectively uses the *required substructure constraint* (a constraint that requires the presence of a substructure); (4) it prospectively uses alternative substructural inferences (such as those derived from long-range ^{1}H–^{13}C COSY); and (5) it provides for efficient interaction between applied constraints.

SESAMI Input

An instructive, user-friendly routine (INPUT) guides the entry of information to SESAMI. The molecular formula and 1-D ^{1}H and ^{13}C NMR spectroscopic data are *required* for program execution. The ^{13}C NMR data must include the chemical shift and hydrogen atom multiplicity of each signal. SESAMI assumes no fortuitous overlap of signals; fewer carbon signals than atoms are used as a measure of molecular symmetry in the unknown.[18] Chemical shifts, integrals, and hydrogen exchangeability (addition of D_2O) are *required* ^{1}H NMR data, the latter information being used to count heteroatom-attached hydrogens. Unresolved hydrogen signals may be entered as multiplets with an average chemical shift, but the appropriate integral. Hydrogen signal patterns and coupling constants, where discernible by the user, may be entered, but currently these data are used only to a limited extent by the program. Provision for entry of data from other spectroscopic sources (IR, MS, UV) is built into INPUT, but these data are not yet used by INTERPRET.

The results of all types of 2-D NMR experiments are conveniently entered as pairs of correlated signals and the number of intervening bonds between the atoms they correspond to, either an exact number or a range. (At this time SESAMI only recognizes topology, not topography; therefore, only through-bond correlations can be utilized.) For example, a 2-D INADEQUATE correlation would be entered as C169.20 C42.90 1,1. This input statement is interpreted by the program as requiring carbon atoms assigned to these chemical shifts to be separated by exactly one bond. The range for ^{1}H–^{1}H COSY correlations is 2,3; geminal coupling is through two bonds; vicinal, through three bonds. (In many cases the user will be able to distinguish between geminal and vicinal coupling and should assign correlations accordingly, e.g., H5.09 H1.66 3,3.) For the ^{1}H–^{13}C COSY correlation the number of intervening bonds is set to one (1,1); for the long-range ^{1}H–^{13}C COSY, it is usually two or three (e.g., C28.4 H1.73 2,3), but four intervening bonds are also possible; the user must decide on the range to be set. As each connectivity is entered, the program adds it to a table which is handed to the interpretation routine, INTERPRET, for use by INFER2D.

INFER2D/2-DNMRPRUNE

The 2-D NMR data are used by both tracks of INTERPRET. INFER2D produces substructural inferences that serve both as powerful constraints on the structure generation process and as input to 2-DNMRPRUNE, enhancing PRUNE's ability to discriminate between valid and invalid structural building units (ACFs).

In developing INFER2D, the successor to INTERPRET2D, a major requirement was the *efficient* interpretation of *all* through-bond 2-D NMR correlations in the presence or absence of molecular symmetry. The key to realizing this was the recognition that molecular symmetry and alternative interpretations of 2-D NMR data are best treated in the structure generation step, not within INTERPRET. Thus, INFER2D only produces carbon–carbon *signal connectivity*. No consideration of molecular symmetry is required of INFER2D. This is possible because COCOA, as indicated, uses symmetry information prospectively in structure generation, and efficiently and prospectively processes alternative substructural inferences in a single structure generation sequence.

COCOA does not treat hydrogen atoms explicitly; therefore, connectivity correlations that involve hydrogen are first translated by the program to carbon signal connectivity information. Two passes are made through the table of 2-D NMR connectivities produced by INPUT. The first collects all one-bond carbon–hydrogen signal connectivities in a table. The second uses this information to convert observed ^{1}H–^{1}H COSY and long-range ^{1}H–^{13}C COSY correlations to a table of paired ^{13}C signal connectivities. From this latter table, constraints are created as input to COCOA.

Unambiguous connectivity between ^{13}C NMR signals is obtained from 2-D INADEQUATE and ^{1}H–^{1}H COSY/^{1}H–^{13}C COSY. If the unknown possesses no molecular symmetry (determined by comparing the number of ^{13}C NMR signals to the number of carbon atoms in the molecular formula), carbon signal connectivity is the same as carbon atom connectivity and INFER2D attempts to build multicarbon atom substructures from the table of ^{13}C NMR signal connectivities by looking for atom overlaps. Larger substructures as constraints (required substructure constraints) are used more efficiently by COCOA than smaller ones. For example, in the solution of the monochaetin problem described later, the six ^{1}H–^{1}H COSY correlations (Table II) initially yield four different pairs of connected carbon atoms—C52.13•C43.66, C46.70•C14.39, C46.70•C26.27 and C11.45•C26.27—which are then reduced to two discrete carbon atom substructures: C52.13•C43.66 and C14.39•C46.70•C26.27•C11.45, where "•" indicates *any* bond type. Thus, INFER2D provides no information on bond type or attached heteroatoms. If any symmetry is present, then more than one "interpretation" is possible (each expressed as a fragment or set of fragments) from a set of carbon signal connectivities.[4] In this case, the building of substructures is *not* attempted and the information is passed to COCOA as a set of constraints that reveal only the deduced connectivity of each carbon signal, as in the tetrahydrobinor-S problem (Figure 6). As an example, the first statement in Figure 6 requires each carbon atom of chemical shift δ 49.8 ppm (in an unknown, but symmetrical molecule, a given chemical shift may represent more than one carbon atom) to be joined to carbon atoms of chemical shifts δ 39.3, 37.8, and 37.0 ppm. The symmetry test in COCOA ensures that the molecular structures generated using these constraints will possess symmetry compatible with the ^{13}C NMR spectrum.[16]

[10] Shelley, C.; Hays, T.; Munk, M.; Roman, R. *Anal. Chim. Acta* **1978**, *103*, 121.

[11] Carhart, R.; Smith, D.; Gray, N.; Djerassi, C. *J. Org. Chem.* **1981**, *46*, 1708.

[12] Abe, H.; Okuyama, I.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 220.

[13] Bremser, W.; Fachinger, W. *Magn. Reson. Chem.* **1985**, *23*, 1056.

[14] Carabedian, M.; Dagane, I.; Dubois, J. *Anal. Chem.* **1988**, *60*, 2186.

[15] Lipkus, A.; Munk, M. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 9.

[16] Christie, B.; Munk, M. *J. Chem. Inf. Comput. Sci.* **1988**, *20*, 27.

[17] If information is used *prospectively* in structure generation, invalid molecular structures, i.e., those incompatible with the information, are rejected *before* they are generated. This is made possible by using the *constraints* derived from INFER and those intrinsic to the structure generation procedure. Use of such information prospectively rather than retrospectively is much more efficient since it is not necessary to use computer time to generate a complete molecule that in the end will be rejected as invalid.

[18] If two or more signals are known to fortuitously overlap, each can be given a slightly different chemical shift, e.g., δ 33.45 and 33.46 ppm. The program then treats those signals as separate and distinct.

If the 2-D NMR experiment reveals a range of possible connectivities, as with long-range ^{1}H–^{13}C COSY[2] and HMBC (heteronuclear multiple bond multiple quantum coherence)[19] spectroscopies, this ambiguity is conveyed to COCOA by statements of alternative constraints that are generated by INFER2D. For example, together with the ^{1}H–^{13}C COSY datum C14.41 H1.73 1,1, the long-range carbon–hydrogen correlation C28.4 H1.73 2,3 generates the constraint: C28.4·C14.4|C28.4·A·C14.4 (where "·" is any bond type, "|" is "or", and A represents any nonhydrogen atom); i.e., a carbon atom whose chemical shift is δ 28.4 ppm is connected by any bond type *either* directly to a carbon atom whose chemical shift is δ 14.4 *or* to some atom A (carbon or a heteroatom) which in turn is connected to a carbon atom with that chemical shift. Because of the high ambiguity in a set of such alternative constraints, no attempt to build discrete substructures is made, even in the absence of molecular symmetry.

The output of INFER2D also serves as the knowledge base for a routine, 2-DNMRPRUNE (part of PRUNE, Figure 2), which comes into play late in the generation of the ACF shortlist. In the operation of PRUNE, MFPRUNE initially removes ACFs from the exhaustive set of chemically stable ACFs that are incompatible with the molecular formula of the unknown. CNMRPRUNE and HNMR follow, and each requires a data base containing allowed ^{13}C NMR chemical shift ranges and signal multiplicities for the central carbon atom of each carbon-centered ACF, and allowed ^{1}H NMR chemical shift ranges for all ACFs with hydrogen-bearing central carbon atoms, respectively. (Some information on the structural implications of ^{1}H NMR signal multiplicities is also stored by HNMRPRUNE and currently used only in a limited way.) These routines compare the stored data for each ACF surviving MFPRUNE with the observed ^{1}H and ^{13}C NMR spectra and delete those that are incompatible.

Surviving ACFs are organized into groups based on the observed ^{13}C NMR spectrum. For each chemical shift, there is a list of ACFs, each of whose assigned central carbon chemical shift range and signal multiplicity match those of that observed signal. Following carbon-centered ACFs are separate lists of compatible heteroatom-centered ACFs for each heteroatom in the unknown.

The new routine, 2-DNMRPRUNE, examines each surviving carbon-centered ACF for compatibility with the inferences made by INFER2D. For example, suppose INFER2D predicts a connection between a specific methine carbon (δ 30.70) and a methyl carbon (δ 16.60). Any surviving ACF in the group of ACFs assigned to chemical shift δ 30.70 that does not have a neighboring methyl group will be removed. In a sense, this pruning is redundant, since the same information is also presented to COCOA by INFER2D as a substructure constraint. However, a reduced ACF shortlist can lead to greater efficiency during structure generation because a major step in structure generation by COCOA involves the selection of ACFs.[16] Since structure generation is a nonpolynomial (i.e., exponential) problem and PRUNE uses only linear algorithms, pruning as many ACFs as possible during interpretation reduces the amount of work necessary during structure generation.

During the pruning process, a routine, CONSISTENCY, examines the list of surviving ACFs for conflicting structure information. The routine attempts to restore any observed internal inconsistency by deleting the ACF(s) causing it. For example, if all methyl-centered ACFs for a given unknown have methylene or methine carbon as first-layer neighbors, then all quaternary carbon-centered ACFs bearing methyl groups as first-layer neighbors would be deleted. This routine cycles until no further deletions are possible. The surviving ACFs are output as the ACF shortlist which may be examined by the user in an abbreviated, but informative way and further pruned with a convenient editor, if desired.

Problem Solving

A few examples of problem solving will illustrate the current status of the software. The first problem, monochaetin, is a fungal metabolite of molecular formula $C_{18}H_{20}O_5$. The structure assignment was disclosed in 1986[20] and depended in part on 2-D NMR experiments, particularly those revealing long-range carbon–hydrogen relationships. Table I is a tabulation of the 1-D ^{1}H and ^{13}C NMR data reported in the paper and required as input. The observed 2-D NMR data are shown in Table II. One-bond carbon–hydrogen correlations are entered first, followed by three-bond hydrogen–hydrogen correlations and long-range carbon–hydrogen correlations. Each of the 27 ambiguous, long-range carbon–hydrogen correlations is compatible with two different interpretations.

Table I. 1-D ^{1}H and ^{13}C NMR Data for Monochaetin[19]

^{13}C NMR		^{1}H NMR			
shift (ppm)	mult	shift (ppm)	integral	mult	exch[a]
205.94	S	6.79	1		
191.77	S	6.02	1		
169.10	S	5.29	1		
158.52	S	4.05	1		
145.52	S	3.76	1		
143.30	D	3.19	1		
116.22	S	2.13	3	S	
107.04	D	1.81	1		
105.73	D	1.48	1		
82.55	S	1.32	3	S	
52.13	D	1.11	3	D	
46.70	D	0.97	3	T	
43.66	D				
26.27	T				
19.49	Q				
18.92	Q				
14.39	Q				
11.45	Q				

[a] An "E" is entered if a signal disappears upon addition of D_2O.

```
Substructure constraint:  C 52.13 · C43.66
Substructure constraint:  C 14.39 · C46.7 · C 26.27 · C11.45


Substructure constraint:  C 19.49 · C107.04 | C 19.49 · A · C107.04
Substructure constraint:  C 14.39 · C 26.27 | C 14.39 · A · C 26.27
Substructure constraint:  C 18.92 · C 43.66 | C 18.92 · A · C 43.66
Substructure constraint:  C 18.92 · C 82.55 | C 18.92 · A · C 82.55
Substructure constraint:  C 43.66 · C 82.55 | C 43.66 · A · C 82.55
Substructure constraint:  C 82.55 · C105.73 | C 82.55 · A · C105.73
Substructure constraint:  C105.73 · C107.04 | C105.73 · A · C107.04
Substructure constraint:  C 43.66 · C116.22 | C 43.66 · A · C116.22
Substructure constraint:  C105.73 · C116.22 | C105.73 · A · C116.22
Substructure constraint:  C107.04 · C116.22 | C107.04 · A · C116.22
Substructure constraint:  C116.22 · C143.30 | C116.22 · A · C143.30
Substructure constraint:  C 43.66 · C143.30 | C 43.66 · A · C143.30
Substructure constraint:  C107.04 · C145.52 | C107.04 · A · C145.52
Substructure constraint:  C143.30 · C145.52 | C143.30 · A · C145.52
Substructure constraint:  C 19.49 · C158.52 | C 19.49 · A · C158.52
Substructure constraint:  C107.04 · C158.52 | C107.04 · A · C158.52
Substructure constraint:  C143.30 · C158.52 | C143.30 · A · C158.52
Substructure constraint:  C 52.13 · C169.10 | C 52.13 · A · C169.10
Substructure constraint:  C 18.92 · C191.77 | C 18.92 · A · C191.77
Substructure constraint:  C 14.39 · C205.94 | C 14.39 · A · C205.94
Substructure constraint:  C 43.66 · C205.94 | C 43.66 · A · C205.94
Substructure Constraint:  C 52.13 · C205.94 | C 52.13 · A · C205.94

Substructure Constraint:  END
      *** End of substructure constraints from INTERPRET ***
```

Figure 2. Monochaetin: INFER2D-generated substructure constraints.

INFER2D recognizes the absence of molecular symmetry in monochaetin and builds the discrete two and four carbon atom substructures shown first in Figure 2 from the carbon signal connectivity derived from ^{1}H–^{1}H and ^{1}H–^{13}C COSY correlations. The atom adjacencies revealed by these two discrete fragments make the information in four of the 27 long-range carbon–hydrogen correlations—C43.66 H4.05 2,3; C52.13 H3.76 2,3; C46.70

[19] Bax, A.; Summers, M. *J. Am. Chem. Soc.* **1986**, *108*, 2093.

Table II. 2-D NMR Correlations for Monochaetin[19]

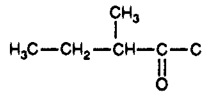
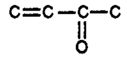
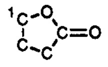
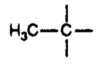
signal 1[a]	signal 2[a]	min[b]	max[c]
C143.30	H6.79	1	1
C107.04	H6.02	1	1
C105.73	H5.29	1	1
C43.66	H3.76	1	1
C52.13	H4.05	1	1
C46.70	H3.19	1	1
C26.27	H1.81	1	1
C26.27	H1.48	1	1
C11.45	H0.97	1	1
C19.49	H2.13	1	1
C18.92	H1.32	1	1
C14.39	H1.11	1	1
H4.05	H3.76	3	3
H3.19	H1.11	3	3
H3.19	H1.81	3	3
H3.19	H1.48	3	3
H0.97	H1.48	3	3
H0.97	H1.81	3	3
C18.92	H3.76	2	3
C19.49	H6.02	2	3
C26.27	H1.11	2	3
C43.66	H1.32	2	3
C43.66	H4.05	2	3
C46.70	H1.11	2	3
C52.13	H3.76	2	3
C82.55	H1.32	2	3
C82.55	H3.76	2	3
C82.55	H5.29	2	3
C105.73	H6.02	2	3
C107.04	H2.13	2	3
C107.04	H5.29	2	3
C116.22	H3.76	2	3
C116.22	H5.29	2	3
C116.22	H6.02	2	3
C116.22	H6.79	2	3
C143.30	H3.76	2	3
C145.52	H6.02	2	3
C145.52	H6.79	2	3
C158.52	H2.13	2	3
C158.52	H6.02	2	3
C158.52	H6.79	2	3
C169.10	H4.05	2	3
C191.77	H1.32	2	3
C205.94	H1.11	2	3
C205.94	H3.76	2	3
C205.94	H4.05	2	3

[a] Element, chemical shift (ppm). [b] Minimum number of intervening bonds. [c] Maximum number of intervening bonds.

H1.11 2,3; and C26.27 H1.11 2,3—redundant, and these are deleted by INFER2D. The information in two of the remaining 23 correlations is redundant leaving a total of 22 alternative constraints (corresponding to about 4.2 million different sets of 22 unambiguous constraints) generated by INFER2D (Figure 2). Each of the 22 either-or statements serves as a separate constraint on structure generation.

The interactive nature of PRUNE was utilized in developing the final ACF shortlist handed to COCOA. It is important for the user of SESAMI to recognize that broad ^1H and ^{13}C NMR chemical shift ranges are assigned to ACFs in the knowledge bases of PRUNE since, if PRUNE is to make an error, it is better to retain an invalid ACF than to delete a valid one. SESAMI may have to work harder with a larger ACF shortlist and may possibly (though not necessarily) produce a greater number of plausible alternative structures in the end, but the risk of eliminating the correct structure will be at a minimum. Increasing the risk should be a user decision. With the ACF shortlist editor, the user can make assumptions at a level of risk comparable to that he or she would make in a conventional structure elucidation and further prune the ACF shortlist. In addition, the IR-, MS-, and UV-derived structural information that is currently beyond the scope of INTERPRET may be brought to bear by the user on ACF shortlist generation.

Table III. Monchaetin: User-Entered Substructure Constraints

	substructure	entry code
1.		CH3CH2CH(CH3)C(=O)C
2.	$H_3C—C=C$	CH3C=C
3.		C=CC(=O)C
4.		1:C—C—C—C(=O)—O—1
5.		CH3—G1 G1 = C
6.	$H_3C—C=$	CH3—G2 G2 = C=

In the monochaetin problem, PRUNE assigns tricoordinate carbon-centered ACFs other than carbonyl carbon-centered ACFs for the carbon signals at δ 205 and 191 ppm, and tetracoordinate as well as tricoordinate carbon-centered ACFs for signals at δ 145 and 143 ppm. (The surviving ACFs predicted by PRUNE for signals at δ 169 and 158 ppm are all tricoordinate carbon-centered.) Using the ACF shortlist editor, the central carbon atoms of the ACF sets corresponding to these two groups of signals were restricted to carbonyl carbon and tricoordinate carbon, respectively. Similarly, the ACF set for the signal at δ 82 ppm was limited to tetracoordinate carbon-centered ACFs. Additionally, dicoordinate carbon-centered ACFs were removed in the editing process, largely on the basis of infrared evidence.

The structure proof of monochaetin as reported[20] did make use of substructural information (Table III, substructures) deduced earlier on the basis of chemical behavior and spectroscopic sources other than NMR. In elucidating the structure with SESAMI, these substructures were *not* added as user-entered constraints in order to get a sense of the information content of the collective 2-D NMR data alone and the ability of INFER2D to extract it. In its solution, SESAMI did indeed reduce the input (molecular formula, Tables I and II) to a manageable number of alternative structures, six in all (Figure 3), the first of which (structure **1**) is identical with that reported in 1986.

Of the six structures, only structure **1** contains three carbon–carbon double bonds; the remaining five have only two. This latter class of structures arises because PRUNE, with its broad chemical shift ranges, considers both tricoordinate and tetracoordinate carbon-centered ACFs as valid assignments for carbon signals at δ 116, 107, and 105 ppm. (A less conservative user, allowing only tricoordinate carbon-centered ACFs for these signals—in addition to those at δ 145 and 143 ppm—would have had a SESAMI output of only structure **1**.)

The substructural information in Table III can be invoked in problem solving by taking advantage of the interactive nature of INFER. User-defined constraints are keyboard-entered using a simple linear code (Table III, Entry Code) that closely mimics conventional structural language (the absence of bonds in an atom sequence implies any possible bond). Cycles are entered in linear format by labeling an atom (1 followed by a semicolon) and forming a bond between that atom and another distant atom by referencing the label (in the entry code for the five-membered lactone, the cycle is formed by a bond between oxygen and the atom assigned label 1). In preparing user-defined substructures as input, no consideration need be given by the chemist to the potential overlap between substructures, e.g., whether the carbon–carbon double bonds of fragments 2 and 3 are one and the same or separately present in the unknown. (INFER allows the user to specify nonoverlap between fragments if this is known, but in solving real-world problems, that information is often not easily inferred.) COCOA accepts the list of user-entered constraints and examines each in generating all compatible structures.

[20] Steyn, P.; Vleggaar, R. *J. Chem. Soc., Perkin Trans. 1* **1986**, 1975.

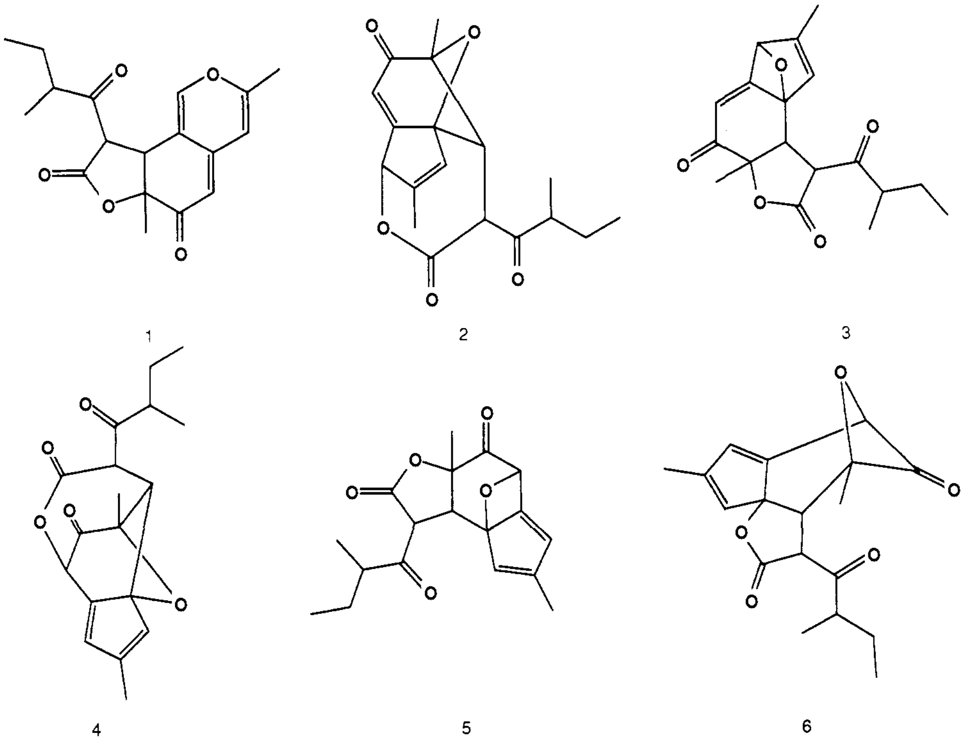


Figure 3. SESAMI output for the monochaetin problem.

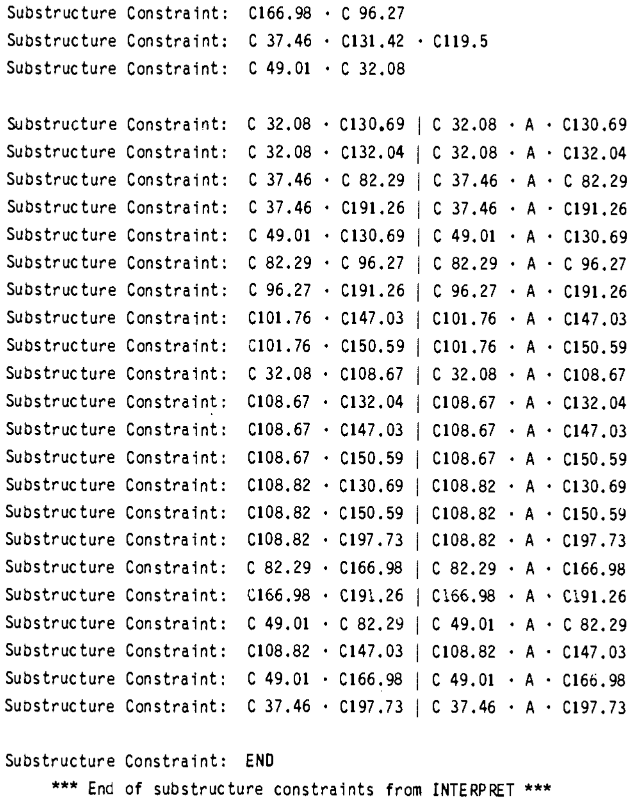
Table IV. 1-D ^1H and ^{13}C NMR Data for the Wasserman Compound

^{13}C NMR		^1H NMR			
shift (ppm)	mult	shift (ppm)	integral	mult	exch[a]
197.73	S	7.70	1		
191.26	S	6.80	1	S	
166.98	D	6.60	1	S	
150.59	S	5.95	2	S	
147.03	S	5.55	1		
132.04	S	5.20	1		
131.42	D	5.05	1		
130.69	S	4.95	1		
119.50	T	4.05	1		
108.82	D	3.70	1		
108.67	D	3.40	1		
101.76	T	3.15	1		
96.27	D	2.85	1		
82.29	S	2.75	1		
49.01	T				
37.46	T				
32.08	T				

[a] An "E" is entered if a signal disappears upon addition of D_2O.

Since the information is used prospectively, it is used efficiently.

Running the monochaetin problem with the user-defined constraints reduces the SESAMI output from six structures to two, structures 1 and 3 (Figure 3). However, even without these constraints, SESAMI achieves the goal of a "manageable" output that would significantly facilitate the assignment of structure. If nothing were known about monochaetin other than the collective spectroscopic data used by SESAMI, it is unlikely the chemist would attempt to construct all compatible molecular structures because of the highly ambiguous nature of the major source of skeletal information, the long-range carbon–hydrogen correlations. In contrast, SESAMI, because of its capacity to utilize alternative constraints prospectively and without preprocessing, is ideally suited to such a task.



```
Substructure Constraint:  C166.98 · C 96.27
Substructure Constraint:  C 37.46 · C131.42 · C119.5
Substructure Constraint:  C 49.01 · C 32.08

Substructure Constraint:  C 32.08 · C130.69 | C 32.08 · A · C130.69
Substructure Constraint:  C 32.08 · C132.04 | C 32.08 · A · C132.04
Substructure Constraint:  C 37.46 · C 82.29 | C 37.46 · A · C 82.29
Substructure Constraint:  C 37.46 · C191.26 | C 37.46 · A · C191.26
Substructure Constraint:  C 49.01 · C130.69 | C 49.01 · A · C130.69
Substructure Constraint:  C 82.29 · C 96.27 | C 82.29 · A · C 96.27
Substructure Constraint:  C 96.27 · C191.26 | C 96.27 · A · C191.26
Substructure Constraint:  C101.76 · C147.03 | C101.76 · A · C147.03
Substructure Constraint:  C101.76 · C150.59 | C101.76 · A · C150.59
Substructure Constraint:  C 32.08 · C108.67 | C 32.08 · A · C108.67
Substructure Constraint:  C108.67 · C132.04 | C108.67 · A · C132.04
Substructure Constraint:  C108.67 · C147.03 | C108.67 · A · C147.03
Substructure Constraint:  C108.67 · C150.59 | C108.67 · A · C150.59
Substructure Constraint:  C108.82 · C130.69 | C108.82 · A · C130.69
Substructure Constraint:  C108.82 · C150.59 | C108.82 · A · C150.59
Substructure Constraint:  C108.82 · C197.73 | C108.82 · A · C197.73
Substructure Constraint:  C 82.29 · C166.98 | C 82.29 · A · C166.98
Substructure Constraint:  C166.98 · C191.26 | C166.98 · A · C191.26
Substructure Constraint:  C 49.01 · C 82.29 | C 49.01 · A · C 82.29
Substructure Constraint:  C108.82 · C147.03 | C108.82 · A · C147.03
Substructure Constraint:  C 49.01 · C166.98 | C 49.01 · A · C166.98
Substructure Constraint:  C 37.46 · C197.73 | C 37.46 · A · C197.73

Substructure Constraint:  END
      *** End of substructure constraints from INTERPRET ***
```

Figure 4. Wasserman compound: INFER2D-generated substructure constraints.

A sample of a compound of synthetic origin kindly provided by Professor Harry Wasserman of Yale University is the basis for a second illustration of the potential of INFER2D in structure elucidation. The 1-D and 2-D NMR data shown in Tables IV and V for this compound of molecular formula $C_{17}H_{15}NO_4$ are the input to SESAMI. For this problem, the long-range carbon-hydrogen coupling data were derived from two experiments, ^1H–^{13}C COSY[2] and HMBC.[19] The output of INFER2D (Figure 4) reveals three discrete carbon atom substructures and 22 pairs of alternative carbon–carbon connectivities. No user-defined constraints were entered in the Wasserman problem, but the ACF shortlist was edited to require only carbonyl carbon-centered ACFs for the carbon signals at δ 197 and 191 ppm, and to exclude all dicoordinate carbon, decisions based in part on infrared evidence. SESAMI generated a *single* structure (Figure 5), identical with

Table V. 2-D NMR Correlations for the Wasserman Compound

signal 1[a]	signal 2[a]	min[b]	max[c]
C166.98	H7.70	1	1
C108.82	H6.80	1	1
C108.67	H6.60	1	1
C101.76	H5.95	1	1
C131.42	H5.55	1	1
C119.50	H5.20	1	1
C119.50	H5.05	1	1
C96.27	H4.95	1	1
C49.01	H4.05	1	1
C49.01	H3.70	1	1
C37.46	H3.40	1	1
C32.08	H3.15	1	1
C37.46	H2.85	3	3
C32.08	H2.75	3	3
H7.70	H1.81	3	3
H5.55	H1.48	3	3
H5.55	H1.48	3	3
H5.55	H1.81	3	3
H5.55	H3.76	2	3
H4.05	H6.02	2	3
H3.70	H1.11	2	3
H3.70	H1.32	2	3
C130.69	H4.05	2	3
C130.69	H1.11	2	3
C132.04	H3.76	2	3
C82.29	H1.32	2	3
C82.29	H3.76	2	3
C119.50	H5.29	2	3
C119.50	H6.02	2	3
C131.42	H2.13	2	3
C191.26	H5.29	2	3
C130.69	H3.76	2	3
C166.98	H5.29	2	3
C191.26	H6.02	2	3
C147.03	H6.79	2	3
C150.59	H3.76	2	3
C132.04	H6.02	2	3
C147.03	H6.79	2	3
C150.59	H2.13	2	3
C130.69	H6.02	2	3
C150.59	H6.79	2	3
C197.73	H4.05	2	3
C191.26	H1.32	2	3
C32.08	H1.11	2	3
C37.46	H3.76	2	3
C82.29	H4.05	2	3
C82.29	H7.70	2	3
C197.73	H2.85	2	3
C166.98	H3.70	2	3
C166.98	H4.05	2	3
C147.03	H6.80	2	3
C108.67	H2.75	2	3
C108.67	H3.15	2	3
C96.27	H7.70	2	3
C82.29	H3.70	2	3
C49.01	H7.70	2	3
C32.08	H3.70	2	3
C32.08	H4.05	2	3

[a] Element, chemical shift (ppm). [b] Minimum number of intervening bonds. [c] Maximum number of intervening bonds.

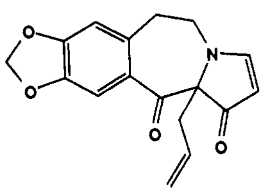


Figure 5. SESAMI output for the Wasserman compound.

that described by Wasserman. In this case a single structure results even though both tricoordinate (correct) and tetracoordinate (incorrect) carbon-centered ACFs were assigned by PRUNE to carbon signals δ 147, 132, and 130 ppm. Thus, the presence of invalid ACFs on the ACF shortlist does not necessarily lead to the generation of invalid structures. The Wasserman compound is another example of a structure problem not likely to be readily solved by the chemist solely on the basis of the NMR data used by SESAMI.

Table VI. 1-D ^{1}H and ^{13}C NMR Data for Tetrahydrobinor-S[20]

^{13}C NMR		^{1}H NMR			
shift (ppm)	mult	shift (ppm)	integral	mult	exch[a]
49.80	D	1.00[b]	20		
39.30	D				
37.80					
37.00	D				
32.40	T				
32.20	T				
24.10	T				

[a] An "E" is entered if a signal disappears upon addition of D_2O. [b] A broad unresolved multiplet.

Table VII. 2-D NMR Correlations for Tetrahydrobinor-S[20]

signal 1[a]	signal 2[a]	min[b]	max[b]
C49.80	C39.30	1	1
C49.80	C37.80	1	1
C49.80	C37.00	1	1
C39.30	C37.00	1	1
C39.30	C32.40	1	1
C37.80	C32.40	1	1
C37.80	C32.20	1	1
C37.00	C24.10	1	1
C32.20	C24.10	1	1

[a] Element, chemical shift (ppm). [b] Minimum number of intervening bonds.

```
Substructure constraint:  C49.8(C39.3) (C37.8) (C37)
Substructure constraint:  C39.3(C49.8) (C37) (C32.4)
Substructure constraint:  C37.8(C49.8) (C32.4) (C32.2)
Substructure constraint:  C37(C49.8) (C39.3) (C24.1)
Substructure constraint:  C32.4(C39.3) (C37.8)
Substructure constraint:  C32.2(C37.8) (C24.1)
Substructure constraint:  C24.1(C37) (C32.2)
Substructure constraint: END
      *** End of substructure constraints from INTERPRET ***
```

Figure 6. Tetrahydrobinor-S: INFER2D-generated substructure constraints.

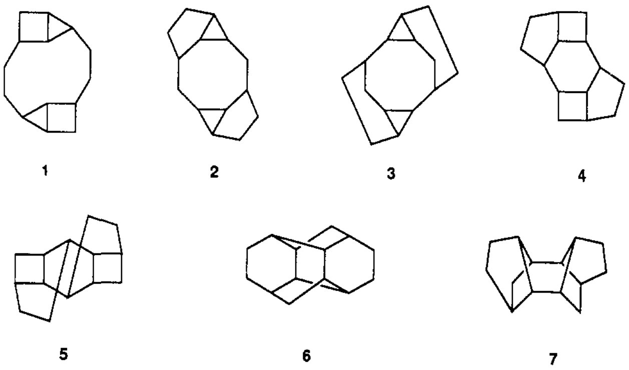


Figure 7. SESAMI output for the tetrahydrobinor-S problem.

The hydrocarbon tetrahydrobinor-S[21] illustrates SESAMI's ability to treat problems involving molecular symmetry. SESAMI input consisted of the molecular formula, $C_{14}H_{20}$, the 1-D ^{1}H and ^{13}C NMR data (Table VI), and the results of one-bond carbon–carbon correlations derived from the 2-D INADEQUATE experiment (Table VII). *Complete* carbon–carbon *signal* connectivity was determined by INFER2D (Figure 6), but because of the symmetry of the compound, more than one arrangement

[21] Krishnamurthy, V.; Shih, J.; Olah, G. *J. Org. Chem.* **1985**, *50*, 3005.

of carbon atoms in the skeleton is possible. Therefore, INFER2D attempts no further substructural analysis and forwards the signal connectivity directly to COCOA. In running this problem, there was no editing of the ACF shortlist and no user-defined substructures were entered. SESAMI produced seven structures (Figure 7), all of which conform to the observed carbon–carbon signal connectivity. Structure **1** was assigned by Olah.[21] (The same seven structures were produced by INTERPRET2D/ASSEMBLE2D using the same input information.[4])

Conclusions

It has been demonstrated that the addition of 2-D NMR data interpreting capability (INFER2D/2-DNMRPRUNE) to SESAMI, a computer-based system of structure elucidation, substantially enhances its power to solve real-world structure problems. A significant advantage of INFER2D is that it allows the structural implications of *all* through-bond signal correlations, including the all-important long-range carbon–hydrogen signal correlations, to be utilized. Inherent in the commonly encountered, large set of long-range correlations is substantial ambiguity, often more than the chemist can contend with if the information is to be used to directly construct compatible structures. SESAMI's ability to do so is made possible by a new structure generating procedure (COCOA) that prospectively utilizes alternative interpretations of spectroscopic data. The same procedure permits SESAMI to readily accomodate unknowns possessing molecular symmetry. The power of SESAMI is further enhanced by its interactive nature. Information known to the user may be conveniently communicated to the program.

Although SESAMI already possesses some problem-solving capability, considerable enhancements to it are envisaged and made possible by an underlying framework for computer-enhanced structure elucidation of considerable promise.

Experimental Section

NMR spectra for the Wasserman compound were recorded on a Bruker Model AM 400 spectrometer (400 MHz for ^{1}H and 75 MHz for ^{13}C). Multiplicities of ^{13}C NMR signals were determined from APT data. Standard pulse sequences were used for the ^{1}H–^{1}H COSY, ^{1}H–^{13}C COSY, long-range ^{1}H–^{13}C COSY, and HMBC experiments.

INFER2D is written in FORTRAN. Earlier program development was on a Prime 450 minicomputer dedicated to this project. Currently program development is continuing on a VAXstation 3500 running under VMS 5.1. COCOA is a Pascal program.[16]

Acknowledgment. The financial support of this work by the National Institutes of Health (GM 37963), The Upjohn Company, and Sterling Drug, Inc., is gratefully acknowledged. Stimulating and informative discussions with Dr. Martin Badertscher contributed to the development of this manuscript. We are grateful to Professor Harry H. Wasserman of Yale University for a generous sample of the "Wasserman compound" described herein and to Dr. Ronald Nieman and Ms. Sharon Ma of the NMR Laboratory at Arizona State University for determining its spectroscopic properties.

Structure Elucidation of a Novel Antibiotic of the Vancomycin Group. The Influence of Ion–Dipole Interactions on Peptide Backbone Conformation

Nicholas J. Skelton,[†] **Dudley H. Williams,***[,†] **Michael J. Rance,**[‡] **and John C. Ruddock**[‡]

Contribution from the University Chemical Laboratory, Lensfield Road, Cambridge, CB2 1EW England, and Central Research, Pfizer, Sandwich, Kent, CT13 9NJ England.

Received October 30, 1990

Abstract: UK-69542, a novel glycopeptide belonging to the vancomycin group of antibiotics, has been isolated from *Saccharothrix aerocolonigenes* fermentation broth and has had its structure determined by a combination of fast atom bombardment mass spectrometry (FABMS) and two-dimensional NMR. The aglycon of the antibiotic is identical with that of aridicin but contains novel groups attached to the peptide core. Proton NMR studies revealed that this novel antibiotic exists in two conformational forms in DMSO solution. The use of NOESY experiments implicated a cis to trans amide bond isomerization as the cause of the conformational difference. This marked conformational change observed for UK-69542, but not for aridicin, is deduced to arise from a charge–dipole interaction involving an aryl sulfate ester; this functional group is not present in aridicin. The observation of this change highlights the ability of electrostatic interactions to stabilize polypeptide secondary structure.

Introduction

The vancomycin group of antibiotics have received considerable attention in recent years owing to the lack of bacterial strains developing resistance to their antibiotic activity.[1] For this reason, vancomycin—the original member of the group to be isolated and brought into clinical use—has become the main line of defense against methicillin resistant *Staphyloccocus aureus* infections.[2] Antibiotics belonging to this group of antibiotics have been isolated from *Actinomycete* cultures obtained from soil samples from diverse geographical locations. The pharmaceutical industry has been actively involved in isolating these novel antibiotics in the hope of finding compounds either with increased efficacy, with a wider spectrum of activity, or with fewer side effects than those currently in use. The determination of structure for the novel antibiotics of the group[3] and the preparation of semisynthetic derivatives[4] have helped to determine which structural features are needed for the antibiotics to bind to peptide analogues of bacterial cell wall components.[5] More importantly, NMR ex-

[1] Foldes, M.; Munro, R.; Sorrell, T. C.; Shanker, S.; Tooley, M. J. *J. Antimicrob. Chemother.* **1983**, *11*, 21.

[2] (a) Klaystersky, J.; et al. *J. Antimicrob. Chemother.* **1983**, *11*, 361. (b) Richardson, J. F.; Marples, R. R. *J. Med. Microbiol.* **1982**, *15*, 475.

[3] Barna, J. C. J.; Williams, D. H. *Annu. Rev. Microbiol.* **1984**, *38*, 339–357.

[4] (a) Herrin, T. R.; Thomas, A. M.; Perun, T. J.; Mao, J. C.; Fesik, S. W. *J. Med. Chem.* **1985**, *128*, 1371–1375. (b) Barna, J. C. J.; Williams, D. H.; Williamson, M. P. *J. Chem. Soc., Chem. Commun.* **1985**, 245–256.