Applications of a HOUDINI-Based Structure Elucidation System

K.-P. Schulz, A. Korytko, and M. E. Munk*

Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona 85287

Received March 27, 2003

SESAMI, a comprehensive program for the elucidation of the structure of complex compounds of carbon, incorporates a structure reduction-based structure genearator (COCOA). Observed limitations with this program in the solution of higher molecular weight unknowns prompted the development of a structure generator (HOUDINI) which embodies a new concept, convergent structure generation. A comparison of the performance of COCOA-based and HOUDINI-based SESAMI using a set of complex, naturally occurring compounds as a test set of unknowns revealed faster execution times and more efficient processing of ambiguous structural information for the latter.

INTRODUCTION

The elucidation of the structure of compounds of carbon has been and continues to be of widespread importance in the chemical, biological, and medical sciences. Compounds of natural origin in particular are of great interest because of their potential as powerful chemotherapeutic agents. Although several computer-based structure elucidation systems have been described,[1] their application to the characterization of such compounds can pose special difficulties because quite often they not only have high molecular weights but also complex skeletons and extensive functionality as well. The scope and limitations of one of these systems, SESAMI, in solving real-world structure problems have been extensively studied.[2−6]

SESAMI shares the goal of other computer-based structure elucidation systems; the assignment of the molecular structure of an unknown directly from its collective spectral properties. In an ongoing project, SESAMI is being developed to achieve that goal in a two-step process. In the first step, a spectrum interpretation program attempts to generate a pool of structural inferences from the spectral data that is sufficiently rich in information content to dramatically limit the number of plausible molecular assignments, preferably to one. In the second step, the collective structural inferences are handed directly to a structure generator that exhaustively constructs all molecular structures compatible with that input. In the event that more than one molecular structure is produced, viewing the structures provides invaluable guidance for the chemist in arriving at the final assignment. Chemists experienced in their craft are generally sufficiently insightful to readily and reliably identify the correct structure among a small number of alternatives based either on evidence already at hand or that derived from a minimum number of selectively designed spectroscopic experiments. Such a computer-based system can significantly augment the productivity of the chemist and at the same provide some assurance that no equally compatible structure has been overlooked.

In the event that more than one plausible structure is generated, a third step—spectrum prediction and comparison—can at times be helpful in ranking the generated structures in the order of decreasing probability of being correct. However, a rather sophisticated spectrum prediction program may be required because with a small number of alternative candidates, they are generally structurally similar.

Extensive problem-solving studies using a version of SESAMI incorporating a structure reduction-based structure generator (COCOA)[2] revealed factors which impact the performance of that system. One factor is the number of atoms in the molecular formula of the unknown. Although caution is required in interpreting the significance of comparative execution times for different structure problems, it does appear that as the number of atoms increases by a factor of *x*, the computational requirements of structure generation increase by a factor greater than *x*. Of course, the nature and contents of the input influence execution time. Explicit information provided to the structure generator—e.g., a user-entered, explicitly defined, required substructure—usually decreases execution time.

Ambiguity of the input to the structure generator is a second factor. Although ambiguous structural inferences do serve to constrain structure generation, they do not do so as effectively as explicitly defined inferences. A commonly encountered source of ambiguity is the 2D NMR-derived, long-range carbon−hydrogen correlation (HMBC experiment), of which there can be very many reported for a complex natural product. Typically, the HMBC correlation is limited to either two or three intervening bonds between the chemical shift-labeled hydrogen and carbon atoms; however, four bond (and even five bond) correlations are known. The allowed intervening bond range for one or more HMBC correlations can be increased accordingly; however, this further increases ambiguity and can and often does increase execution time and the number of compatible structures generated.

The information-rich ACF Shortlist,[1] produced by the spectrum interpretation module INTERPRET, is another source of input ambiguity. The ACF Shortlist consists of a family of alternative fragments for *each non-hydrogen atom* in the unknown, only one of which is correct. Each fragment (an ACF) is explicitly defined and consists of a valence-

* Corresponding author phone: (602)965-4430; fax: (602)965-2747.

10.1021/ci034058j CCC: $25.00
Published on Web 08/15/2003

Table 1. Summary of Input to SESAMI

		1D NMR[a]		2D NMR[g]		
compound	mol formula	^{1}H NMR[b]	^{13}C NMR	HMQC	HMBC	COSY
1. DDP	$C_{23}H_{37}NO_7$	31	23	26	78	26
2. stenocarpine	$C_{21}H_{31}O_3$	27[c]	21	23	87	22
3. 1-demethylwinkleridine	$C_{22}H_{35}NO_6$	31[d]	22	26	58	26
4. guttiferone A	$C_{38}H_{50}O_6$	31[c]	38	28	88	0
5. lacinan-8-ol	$C_{15}H_{26}O$	18[e]	15	17	71	22
6. α-botryoxanthin	$C_{74}H_{112}O_2$	72	74	72	227	4
7. syringolide	$C_{15}H_{24}O_6$	17[f]	15	15	21[h]	6

[a] Number of signals. [b] All H atoms accounted for. [c] Three exchangeable hydrogens. [d] Five exchangeable hydrogens. [e] One exchangeable hydrogen. [f] Two exchangeable hydrogens. [g] Number of reported atom−atom correlations. [h] Five of HMBC correlations assigned a range of two to four intervening bonds between hydrogen and carbon.

satisfied, central atom and one concentric layer of nearest neighbors at least one of which is not valence satisfied. Although only one ACF in each family is present in the correct structure, other ACFs in that same family may lead to other plausible structures which are compatible with the collective spectral data. The number of alternative ACFs in a given family—which can reach 100 and more in some cases—is dependent on the richness of the spectral data entered and any initial constraints designated by the user. Experience has revealed that as the size of the ACF Shortlist in a given problem decreases (i.e., as the number of ACFs in one or more families decreases), execution time generally decreases.

A third source of ambiguity is uncertainty in the hybridization assigned in the spectrum interpretation step to one or more of the non-hydrogen atoms in the unknown. The structure generator COCOA is especially sensitive to this factor since it currently requires explicit designation of hybridization of all atoms in its representation of the initial problem state. In cases involving one or more atoms of uncertain hybridization, more than one initial problem state is needed to represent the entire search space, each of which is created automatically by COCOA and treated as a separate structure generation problem.[1] As a result, computational time is generally increased, in some cases substantially.

HOUDINI−BASED SESAMI

Observed limitations in the efficiency of the COCOA-based SESAMI system with increasing number of atoms and complexity of the unknown, and increasing ambiguity in the structural inferences, signaled the need for a structure generator whose performance is less adversely impacted by these factors. HOUDINI, a structure generator based on an entirely new concept—convergent structure generation—was developed to enhance the efficiency of SESAMI and thereby expand its useful range and scope.[1]

Several features of the new structure generator are noteworthy. First, HOUDINI, in contrast to COCOA, requires only a single representation of the initial problem state, even in the presence of hybridization uncertainties. Second, HOUDINI accepts the set of structural inferences from INTERPRET and initially preprocesses them to produce a *single, integrated representation* of the collective information. Not only can this single representation be utilized more efficiently in structure generation than the original set of individual structural inferences (which COCOA uses), but it is also richer in information than the original set. Third, relative to COCOA, information processing is more intelligently *managed* so as to eliminate large families of invalid structures earlier in the structure generation process. Fourth, HOUDINI processes sets of alternative inferences in a concerted rather than a sequential treatment. The latter is especially important since so much of the input to the structure generator is in the form of alternative inferences.

The study described in this paper was designed to evaluate the potential of the concept of convergent structure generation without the investment required for full-scale development. Toward this end, an initial working model of convergent structure generation was created and linked to INTERPRET for purposes of testing. The performance of this prototype of the next generation of SESAMI (SESAMI H) was compared to that of the earlier COCOA-based SESAMI (SESAMI C).

Table 2. 1D ^{1}H NMR Spectrum of DDP[a]

shift	integral	shift	integral	shift	integral
4.48	1	2.90	1	2.00	1
4.00	1	2.87	1	1.87	1
3.84	1	2.72	1	1.86	1
3.69	1	2.60	1	1.58	1
3.64	1	2.36	1	1.53	1
3.41	3	2.21	1	1.52	1
3.40	1	2.11	1	1.49	1
3.39	3	2.07	1	1.11	3
3.06	1	2.01	1		

[a] Five signals corresponding to rapidly exchangeable hydrogens 4−5 ppm are not included.

COMPARATIVE TESTING

All of the structure problems selected for this comparative study involve natural-occurring compounds. With one exception, the structures of the compounds were originally assigned by other investigators using conventional means largely on the basis of extensive spectral data. The exception is Syringolide, a naturally occurring fungiside whose characterization included the application of SESAMI C.

The structure of 18-demethyl-14-deacetylpubescenine ("DDP") was reported in 1997.[7] The paper included extensive 1D NMR data and 2D shift correlated NMR data (Tables 1−4). The following description of the computer-based structure elucidation of DDP is illustrative of problem-solving via SESAMI. The spectral input to SESAMI was taken from the publication and consists of three components. The first two—the molecular formula ($C_{23}H_{37}NO_7$) and 1D NMR data—are required input. The latter includes both ^{1}H and ^{13}C

Table 3. 1D ^{13}C NMR Spectrum of DDP

shift	multiplicity	shift	multiplicity	shift	multiplicity
85.6	S	57.7	T	41.5	D
84.2	D	56.7	Q	39.9	S
81.0	S	53.3	Q	31.0	T
75.4	D	51.4	T	29.8	T
73.3	D	47.9	D	29.6	T
70.6	D	47.8	S	29.2	T
70.0	T	46.5	D	14.0	Q
64.5	D	44.3	D		

NMR spectral data (Tables 2 and 3, respectively). Five of the ^{1}H NMR signals (not listed in Table 2) correspond to rapidly exchangeable hydrogens. SESAMI uses this information to identify the presence of five heteroatom-bearing hydrogens.

The third component, 2D NMR correlation data, although optional, is necessary for the solution of large, complex structures such as DDP. The 2D NMR data summarized in Table 4 describe the results of the three commonly utilized experiments: HMQC, one-bond carbon−hydrogen correlations; COSY, three-bond hydrogen−hydrogen correlations; and HMBC, long-range carbon−hydrogen correlations. Although SESAMI is currently an NMR-based program, this is not an inherent limitation in its design. Programs interpreting data from other spectroscopic sources are planned for INTERPRET.

SESAMI is a flexible program that can adapt to the style of the user. In this laboratory, characterizing a new compound about which nothing structural is known (such as is assumed in the case of DDP problem) is often initiated by applying a standard set of constraints. The purpose is to rapidly reveal a "likely" solution or solutions and provide some insight with regard to plausible assignments. The standard set of applied constraints—referred to as the *standard edit set*—includes the following*:*

1. HMQC correlations set to one intervening bond.
2. COSY correlations set to three intervening bonds.
3. HMBC correlations set to two or three intervening bonds.
4. INADEQUATE correlations set to one intervening bond.
5. Sp hybridized atoms forbidden.
6. ^{13}C NMR signals greater than 115 ppm not assigned to sp^3 hybridized carbon atoms.
7. ^{13}C NMR signals less than 80 ppm not assigned to sp^2 hybridized carbon atoms.
8. Of the group of singlet and doublet ^{13}C NMR signals *above* 170 ppm, up to four signals with the highest chemical shifts assigned to carbonyl or thiocarbonyl carbon atoms if the appropriate number of oxygen and/or sulfur atoms are present in the unknown. If the spectrum reveals at least one rapidly exchangeable hydrogen whose ^{1}H NMR chemical shift is equal to or greater than 9 ppm (enolic hydrogen), the number of such carbon signals assigned to carbonyl or thiocarbonyl is reduced to one.
9. Compounds containing three-membered rings excluded.
10. Compounds containing a member of a predefined library of highly strained structural features excluded.

In succeeding runs of the problem, chemist insight may suggest modification or disabling of one or more of these initial constraints. The nature and extent of the changes are

Table 4. 2D NMR Data of DDP

^{13}C shift	HMQC[a]	HMBC[b]	^{1}H-^{1}H COSY[c]
85.6		4.48, 2.60, 2.07	
84.2	3.40	4.00, 3.39, 2.60, 2.21, 2.11, 1.58	2.60, 2.11
81.0		4.48, 4.00, 3.41, 2.72, 2.60, 2.11, 2.01, 1.87	
75.4	4.00	2.21, 2.00, 1.58	2.21, 2.01
73.3	3.64	1.87, 1.86	1.52, 149
70.6	4.48	2.72	2.07
70.0	3.84	2.87, 2.07, 1.86	
	3.69		
64.5	2.72	4.48, 3.06, 2.90, 2.36, 2.07, 1.87	
57.7	2.87	3.84, 3.69, 3.06, 2.90, 2.72, 2.07, 1.86, 1.53	
	2.36		
56.7	3.39		
53.3	3.41		
51.4	3.06	2.87, 2.72, 1.11	1.11
	2.90		1.11
47.9	2.01	2.60, 2.21	4.00, 1.87
47.8		2.72, 2.07, 1.87, 1.58	
46.5	2.07	3.84, 3.69, 2.72, 2.36	4.48
44.3	1.87	3.64, 2.72, 221, 2.07	2.01, 2.00, 1.58
41.5	2.21	2.60, 2.00, 1.58	4.00, 2.00
39.9		4.48, 3.84, 3.69, 2.87, 2.36, 2.07, 1.86	2.21, 1.87
31.0	2.00	2.01, 1.87	2.21, 1.87
	1.58		1.87
29.8	1.52	1.53	3.64, 1.86, 1.53
	1.49		3.64, 1.86, 1.53
29.6	1.86	3.84, 3.69, 2.87, 2.36	1.52, 1.49
	1.53		1.52, 1.49
29.2	2.60	2.21	3.40
	2.11		3.40
14.0	1.11	3.06, 2.90	3.06, 2.90

[a] Chemical shift of H atom(s) attached to the corresponding C atom. [b] Chemical shifts of H atoms that correlate to the corresponding C atom in the HMBC experiment. [c] Chemical shifts of H atoms that correlate to the corresponding H atom (HMQC column) in the COSY experiment.

Table 5. Comparative Testing

compound/constraints	INFER2D[e] explicit[e]	INFER2D[e] ambiguous[f]	execution time[k] SESAMI C	execution time[k] SESAMI H	no. of structures
1. DDP					
all 2D NMR data					
SE[a]	5 (10)	38[g]	3.5	0.9	1
SE + sp atoms[b]	5 (10)	38[g]	4.5	1.1	1
SE + 3 rings[c]	5 (10)	38[g]	3.8	1.3	1
AA[d]	5 (10)	38[g]	31.4	1.5	1
COSY data deleted					
SE[a]	0	50		431.4	1
SE + sp atoms[b]	0	50		472.9	1
SE + 3 rings[c]	0	50		766.1	1
AA[d]	0	50		994.5	1
2. stenocarpine					
SE[a]	4 (7)	41[h]	27.5	0.6	1
SE + sp atoms[b]	4 (7)	41[h]	27.6	0.7	1
SE + 3 rings[c]	4 (7)	41[h]	413.0	0.9	1
AA[d]	4 (7)	41[h]	431.1	1.5	1
3. 1-demethylwinkleridine					
SE[a]	5 (10)	34[i]	28.0	1.1	2
SE + sp atoms[b]	5 (10)	34[i]	29.1	1.3	2
SE + 3 rings[c]	5 (10)	34[i]	88.6	1.7	2
AA[d]	5 (10)	34[i]	229.9	1.9	2
4. guttiferone A					
SE[a]	0	55	-	12.9	4
SE + sp atoms[b]	0	55	-	13.8	4
SE + 3 rings[c]	0	55	-	14.2	4
AA[d]	0	55	-	55.0	4
5. lacinan-8-ol					
SE[a]	1 (17)	0	<0.1	<0.1	1
AA[d]	1 (17)	0	<0.1	<0.1	1
SE,[a] 4 bond COSY allowed	0	40	14.0	0.2	1
AA,[d] 4 bond COSY allowed	0	40	35.2	0.4	1
6. α-botryoxanthin					
SE[a]	2 (3)	131[g]	-	3.9	1
SE + sp atoms[b]	2 (3)	131[g]	-	4.6	1
SE + 3 rings[c]	2 (3)	131[g]	-	5.7	1
AA[d]	2 (3)	131[g]	-	6.8	1
7. syringolide					
SE[a] + 4 user-entered substructures	2 (3)	16[j]	2.4	10.0	2
SE[a]	2 (3)	16[j]	679.4	42.8	13

[a] Standard edit constraints applied. [b] Standard edit, but sp atoms allowed. [c] Standard edit, but three-membered rings allowed. [d] All allowed. Constraints 5−9 disabled. [e] Number of explicitly defined fragments. The number in parentheses is the sum of carbon−carbon connections in all of the fragments. [f] Number of paired inferences: C−C or C−A−C, where A is any non-hydrogen atom. [g] Two of the HMBC correlations reduced to only C−A−C fragments. [h] Five of the paired inferences reduced to only C−A−C fragments. [i] Four of the paired inferences reduced to only C−A−C fragments. [j] Two of the HMBC correlations are of the form C−C or C−A−C or C−A−A−C. [k] In seconds.

guided in part by the level of assurance required by the chemist that no plausible structure has been overlooked.

In the constraints controlling parameters of the 2D NMR data (constraints 1−4), the number of intervening bonds allowed should be increased to four (and possibly five) for one or more COSY and/or HMBC correlations where concern about a possible extended range exists. Doing so, however, increases the ambiguity of the input, which can increase execution times and the number of generated structures. (Alternatively, in specific cases where doubt about the range of intervening bonds exists, those correlations can be deleted from the input.) Although constraint 4 is included, the 2D INADEQUATE experiment is not extensively employed due to experimental difficulties.

Although the chemical shift ranges assigned in constraints 6−8 are generally applicable, they can be either tightened or relaxed by the user. Such adjustments can affect the size of theACF Shortlist, which in turn can also alter execution times and the number of generated structures. Constraint 8 pertaining to the carbonyl group has been found to be useful based on experience in problem-solving. The library of specific structural features that lead to the exclusion of highly strained molecular structures (constraint 10) can be viewed and one or more can be disabled by the chemist. For maximum assurance that no plausible structure has been overlooked, constraints 5−9 should be disabled, intervening bond range extensions (or deletion) of one or more HMBC and COSY correlations should be considered, and the library of strained structural features should be carefully considered. It should be noted that SESAMI allows additional interaction with the chemist. User-defined constraints (e.g., explicitly or ambiguously defined required or forbidden substructures) can be entered and the output of INTERPRET (e.g., the ACF Shortlist) can be edited.

Eight variations of the DDP problem were studied with both SESAMI H and SESAMI C (Table 5, number 1). All were executed in automated mode, i.e., there was no intervention by the chemist at any stage, and no input other than the molecular formula and the spectral data described above was used. Each of the runs in the first set of four in

Table 5 incorporate *all* of the spectral information listed in Tables 2−4. The 1D NMR data input includes ^{1}H chemical shift and signal integral and ^{13}C chemical shift and signal multiplicity. The 2D NMR data are input as signal−signal couplings followed by the allowed intervening bond range, e.g., H4.48-C85.6 2 3 for the very first HMBC correlation listed in Table 4.

For input, the collective 2D NMR data in Table 4 are expressed as 130 signal−signal correlations: 26 HMQC correlations, 26 COSY correlations, and 78 HMBC correlations, the latter of which are ambiguous in that they do not distinguish between two and three intervening bonds between the designated hydrogen and carbon atoms. In the second set of four runs, *all* 26 COSY correlations are deleted from the input. The standard edit set of 10 constraints is applied in the first run of each of the sets of four. After the initial run, constraints are disabled in each succeeding run as follows: in the second run, sp hybridized atoms are allowed (only constraint 5 disabled); in the third run three-membered rings *and* sp hybridized atoms are allowed (only constraints 9 and 5 disabled); and in the fourth run, sp atoms and three-membered rings are allowed, and, in addition, constraints 6−8 are disabled (the so-called "all allowed" or "AA" run).

The output of INTERPRET, which is independent of the structure generator used, is handed directly to the structure generator in automatic mode. It consists of two components, the ACF Shortlist and the set of inferences—fragments of chemical shift-labeled carbon atoms—derived from the 2D NMR data. The ACF Shortlist represents 31 ACF families (one for each non-hydrogen atom of DDP), each of which consists of one or more ACFs whose central atom corresponds to one of the non-hydrogen atoms of the unknown. Within a given family, each ACF is a plausible description of the immediate structural environment of the non-hydrogen atom it represents. The ACF Shortlist is generated by a program of INTERPRET (PRUNE) which, for each ^{13}C NMR signal in the spectrum of the unknown, deletes those ACFs from an *exhaustive* list of carbon-centered ACFs whose assigned central carbon chemical shift range or signal multiplicity are inconsistent with that signal. Proton chemical shifts are also used in the "pruning" of ACFs from the exhaustive list. If the observed chemical shift of the hydrogen(s) attached to the central atom of an ACF lies outside of the assigned range, the ACF is deleted. The assigned chemical shift ranges of the central carbon atom and attached hydrogens, if any, of each ACF in the exhaustive list are derived from libraries of assigned NMR spectra. Heteroatom-centered ACF families are also produced by a pruning procedure beginning with an exhaustive list. To ensure exhaustivity of the generated molecular structures, in deleting ACFs from the exhaustive list, PRUNE operates on the principle that it is better to retain an invalid ACF on the Shortlist than to delete the correct ACF. The size and composition of the ACF Shortlist is dependent on the spectral data input to SESAMI and the constraints applied by the user. Note that within a given family of ACFs, ACFs with more than one hybridization of the central atom are possible

A partial list of alternative ACFs for some of the ACF families produced in the first of the eight DDP problem runs (Table 5) is shown in Figure 1. In the linear representation of an ACF, the first atom, with its attached hydrogens, if any, is the central, valence-satisfied atom. This atom is

```
1.  Chemical shift:  85.600

    1.    (C)     (-NH2)     (-OH)      (=C --)
    2.    (C)     (-NH2)     (-CH2 -)   (=CH -)
    25.   (C)     (-NH -)    (-O -)     (=C --)
    242.  (C)     (-N --)    (-C ---)   (-C ---)   (-C ---)

2.  Chemical shift:  84.200

    1.    (CH)    (-CH2 -)   (=C --)
    19.   (CH)    (-CH2 -)   (-N --)    (-C ---)

14. Chemical shift:  47.800

    1.    (C)     (-NH2)     (-CH2 -)   (-CH2 -)   (-CH2 -)
    18.   (C)     (-NO2)     (-CH2 -)   (-CH2 -)   (-CH2 -)
    95.   (C)     (-C =-)    (-C ---)   (-C ---)   (-C ---)

23. Chemical shift:  14.000

    1.    (CH3)   (-CH2 -)

Nitrogen centered ACFs:

    1.    (NH2)   (-CH2 -)
    2.    (NH2)   (-O -)
    28.   (NH)    (-C ---)   (-C ---)
    100.  (N)     (-C ---)   (-C ---)   (-C ---)

Oxygen centered ACFs:

    1.    (OH)    (-CH2 -)
    22.   (O)     (-CH2 -)   (-C=-)
    36.   (O)     (-C ---)   (-C ---)
```

Figure 1. Abbreviated ACF Shortfist from a "Standard Edit" run of the DDP problem.

followed by first-layer atoms, each with their attached hydrogens, if any. The bond to the left of each first-layer atom is the bond by which that atom joins to the central atom; bonds to the right indicate the nature of the bonding sites by which first-layer atoms join to other atoms. The last number in the left most column for each ACF family indicates the number of ACFs in the family. Each carbon-centered ACF family is listed separately and labeled with the ^{13}C NMR chemical shift of the central carbon atom. In the printed output of the ACF Shortlist, all compatible heteroatom-centered ACFs are listed only by element type.

The second component of INTERPRET's output, produced by INFER2D, is a set of chemical shift-labeled carbon fragments derived from the reported 2D NMR signal couplings. This output is dependent only on these data and the values set for constraints 1−3. In the four runs of the first set, standard edit values for constraints 1−3 were used, and all 130 signal couplings were processed by INFER2D giving rise to an output of 43 carbon fragments (Figure 2).

In the multistep interpretation process leading to that output, *explicitly defined* units of two connected, chemical shift-labeled carbon atoms are deduced from combined COSY/HMQC data by identifying carbon signals corresponding to carbon atoms bearing coupled vicinal hydrogen atoms. The combination of HMBC and HMQC data produces a set of *ambiguous* inferences, each of which is expressed as a pair of fragments (separated by a vertical line) requiring two, chemical shift-labeled carbon atoms to be either directly bonded to one another or separated by a single, undefined non-hydrogen atom, atom A (i.e., C−C or C−A−C, e.g., see line 6, Figure 2). The program next attempts to construct larger, explicitly defined carbon fragments from the COSY-derived, two-carbon fragments by first identifying labeled

```
C84.20 .C29.20
1:C75.40( .C47.90 .C44.30 .C31.00 .C41.50( .1))
C73.30 .C29.80 .C29.60
C70.60 .C46.50
C51.40 .C14.00
C85.60 C70.60 | C85.60 A C70.60
C85.60 C46.50 | C85.60 A C46.50
C85.60 C29.20 | C85.60 A C29.20
C84.20 C75.40 | C84.20 A C75.40
C84.20 C56.70 | C84.20 A C56.70
C84.20 C41.50 | C84.20 A C41.50
C84.20 A C31.00
C81.00 C75.40 | C81.00 A C75.40
C81.00 C70.60 | C81.00 A C70.60
C81.00 C64.50 | C81.00 A C64.50
C81.00 C53.30 | C81.00 A C53.30
C81.00 C47.90 | C81.00 A C47.90
C81.00 C44.30 | C81.00 A C44.30
C81.00 C29.20 | C81.00 A C29.20
C73.30 C44.30 | C73.30 A C44.30
C70.60 C64.50 | C70.60 A C64.50
C70.60 C39.90 | C70.60 A C39.90
C70.00 C57.70 | C70.00 A C57.70
C70.00 C46.50 | C70.00 A C46.50
C70.00 C39.90 | C70.00 A C39.90
C70.00 C29.60 | C70.00 A C29.60
C64.50 C57.70 | C64.50 A C57.70
C64.50 C51.40 | C64.50 A C51.40
C64.50 C47.80 | C64.50 A C47.80
C64.50 C46.50 | C64.50 A C46.50
C64.50 C44.30 | C64.50 A C44.30
C57.70 C51.40 | C57.70 A C51.40
C57.70 C46.50 | C57.70 A C46.50
C57.70 C39.90 | C57.70 A C39.90
C57.70 C29.60 | C57.70 A C29.60
C47.90 C29.20 | C47.90 A C29.20
C47.80 C46.50 | C47.80 A C46.50
C47.80 C44.30 | C47.80 A C44.30
C47.80 A C31.00
C46.50 C44.30 | C46.50 A C44.30
C46.50 C39.90 | C46.50 A C39.90
C41.50 C29.20 | C41.50 A C29.20
C39.90 C29.60 | C39.90 A C29.60
```

Figure 2. Output of INFER2D for a "Standard Edit" of the DDP problem.

carbon atoms present in more than one two-carbon fragment. Fragments are extended accordingly. Labeled carbon atoms common to an explicitly defined fragment and one or more of the paired HMBC inferences can reduce the ambiguity of the latter and in some cases allow further explicit connectivity extensions. Inconsistencies among the initially derived, paired HMBC inferences can also reduce one or more to a single fragment. Such steps, and redundancy checks, account for the first five explicitly defined carbon fragments (lines 1−5, Figure 2) representing a total of 10 carbon−carbon connectivities; the reduction of two paired HMBC inferences to a single fragment (lines 12 and 39); and a reduction in the number of final inferences. Note that only connectivity, not bond type is defined in the fragments.

All eight variations of the DDP problem gave rise to a *single* structure (Figure 3a)—the same as that assigned by the original investigators—with both SESAMI H and SESAMI C. (This provides evidence that these investigators did not overlook any plausible alternative.) The results recorded in Table 5 (number 1) for the eight runs permit a number of observations regarding the impact of changes in input and constraints on performance. In the first set of four runs utilizing all of the 2D NMR data, the performance of the two programs, although comparable for practical purposes, does reveal a trend. Relaxing the initially applied standard edit set of constraints increases execution times for both SESAMI C and SESAMI H. The impact of disabling one or two of the 10 constraints (runs 2 and 3, respectively) is minimal for both programs. Disabling five constraints (run 4, constraints 5−9 disabled) decreases the performance of SESAMI H negligibly (from 0.89 to 1.51 s) but produces a more significant adverse impact on SESAMI C (from 3.49. to 31.38 s).

Increased execution times with relaxation of constraints is to be expected. For example, allowing sp hybridized atoms can lead to an increase in the number and scope of ACFs in the ACF Shortlist. Relaxing the chemical shift ranges in constraints 6−8 can increase the number of allowed hybridizations of one or more carbon atoms and consequently add ACFs to the Shortlist. However, SESAMI C is likely to be more adversely impacted than SESAMI H by such increases for two reasons. First, increasing the number of allowed hybridizations of one or more atoms increases the number of initial problem states that COCOA must separately process. For example, in the DDP problem using all spectral data, the change from standard edit to five constraints disabled (5−9) increases the number of COCOA's initial problem states from 80 to 21441. In contrast, HOUDINI uses a single representation for the initial problem state regardless of the constraints applied. Second, COCOA processes the information from INTERPRET by a more stepwise, less efficient pathway than HOUDINI.[1]

The results of the second set of four runs, in which *all* COSY data were deleted from the input, support the above-described trends. However, in this case the difference in performance between the two programs is dramatic. Although execution times for SESAMI H increased from seconds to minutes without COSY data—at the extremes, from 0.89 s for the standard edit run with COSY data to almost 17 min without COSY data and five constraints disabled—none of the four SESAMI C runs completed even after several days of computational time. Again, an increase in execution time is to be expected with the deletion of the COSY data. The five, information-rich, explicitly defined substructures inferred from COSY data are absent from the output of INFER2D. Without interaction between the between the COSY and HMBC derived fragments, the number of paired HMBC-derived correlations increases (to 50). Additionally, the size, and therefore ambiguity, of the ACF Shortlist is increased since the COSY-derived, explicitly defined connections between carbon atoms is one of the means by which ACFs are pruned from the Shortlist. These comparative data suggest that COOCA is far more susceptible than HOUDINI to degradation in performance as the amount of information provided to the structure generator decreases. Also note that the relative change in execution times for SESAMI H with increasing relaxation of constraints is comparable for runs with and without COSY data.

The DDP problem is an interesting example since a single structure is produced in all completed runs, in particular, with and without COSY data. Thus, the information provided by the COSY data must duplicate that in the HMBC data (some of which are also redundant). However, the observed timings for both programs with and without COSY data clearly indicate that spectral data giving rise to explicitly defined fragments in the spectrum interpretation step significantly facilitates structure generation. Also noteworthy in this particular problem, the increase in the size of the ACF Shortlist that results with the deletion of the COSY data does

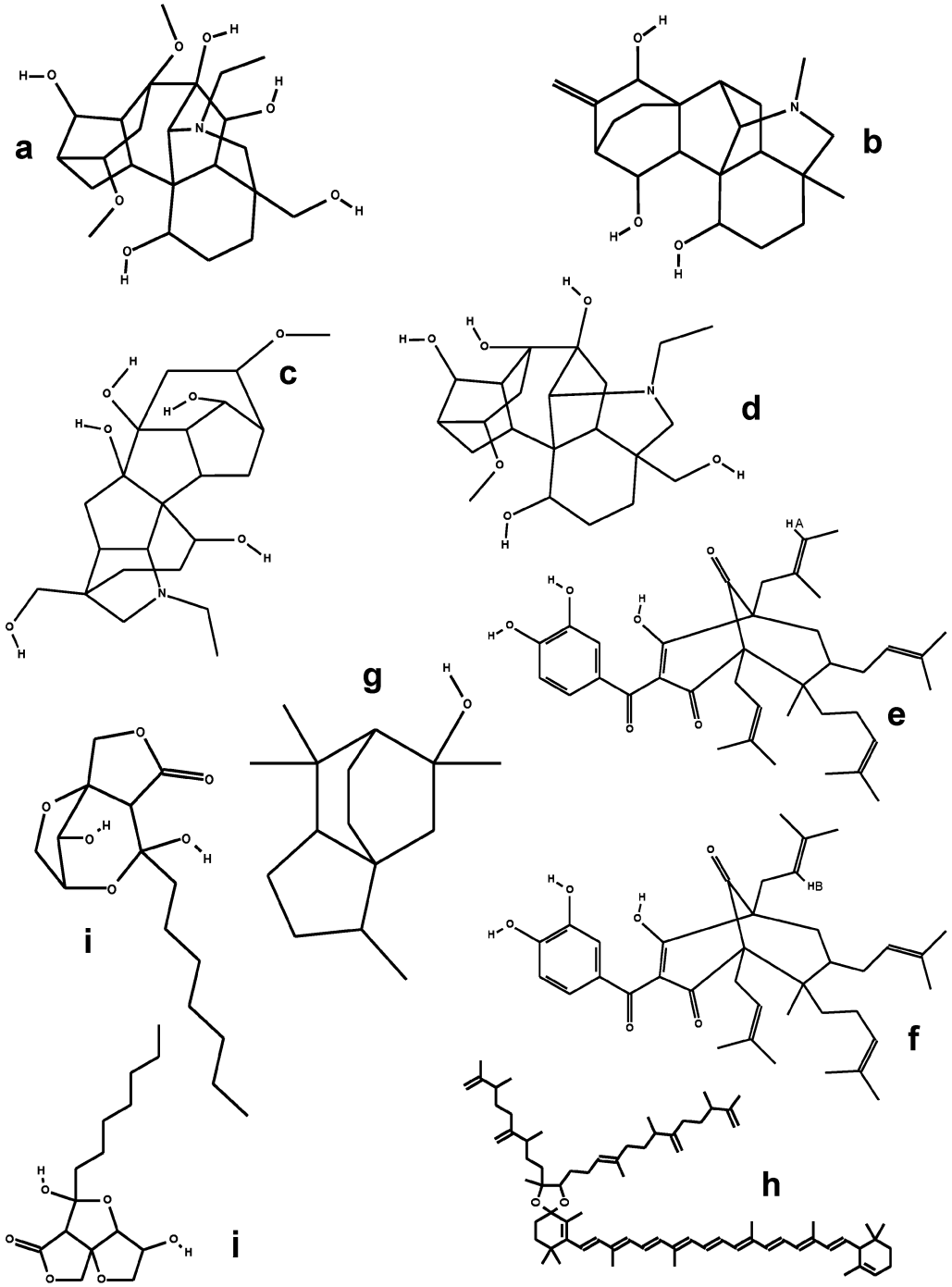

Figure 3. Structural output of SESAMI.

not lead to any additional structures. Such is not always the case.

An interesting and anomalous timing result is observed in the COCOA runs of the DDP problem. As discussed above, relaxation of constraints is expected to result in increased execution times. However, in run 3 relative to run 2, an additional constraint is relaxed—three-membered rings are allowed—but a *slight* decrease in execution time is observed instead (from 4.5 to 3.8 s). In the other cases of comparable runs (stenocarpine and 1-demethylwinkleridine, Table 5), expected behavior is observed. An examination of the "internal counters" collected by COCOA for each run suggested at least one factor that could account for the anomaly. In run 2, in which three-membered rings are forbidden, the test for their presence is applied during structure generation, but the internal data indicate that it led to no pruning of the search space, i.e., no speedup of the run was achieved by this constraint. In run 3, that test was not applied—saving a small amount of time—and, the size ACF Shortlist increased only slightly. The result was a net decrease in execution time compared to run 2. In constrast, in the stenocarpine and demethylwinkleridine problems, although the tests applied in run 2 did not result in pruning of the problem state, relaxation of the three-membered ring constraint increased the size of the ACF Shortlist. In these two cases, the impact of that increase is an increase in the space needed to be searched by several orders of magnitude, accounting for the observed net increase in execution time.

The study of additional structure problems taken from the primary literature led to trends comparable to those observed in the DDP problem. Table 1 provides pertinent information about the literature-derived input to SESAMI for each of the six additional problems. Table 5 lists execution times and other features of the output for both SESAMI C and SESAMI H.

The stenocarpine,[8] 1-demethylwinkleridine,[7] and guttiferone A[9] problems were each run in automated mode using the same four variations of initially applied constraints as in the DDP problem: i.e., standard edit; standard edit, but with the constraint restricting sp atoms disabled; standard edit, but with the constraints restricting three-membered rings and sp atoms disabled; and standard edit with the constraints restricting sp atoms, three-membered rings, and chemical shift ranges (constraints 5−9) disabled.

The only input for the stenocarpine and 1-demethylwinkleridine problems was the molecular formula of the compound and the spectral data reported in the original papers: 1D ^{1}H NMR and ^{13}C NMR signals and the results of HMQC, HMBC, and COSY experiments (Table 1). In both problems, significantly faster execution times were observed for the HOUDINI-based system than the COCOA-based system in all comparative runs (Table 5, numbers 2 and 3). Once again, the performance of SESAMI C also revealed greater sensitivity to constraint relaxation. Although disabling of the sp atom constraint did not measurably increase execution time, disabling the sp atom constraint and the three-membered ring constraint or disabling constraints 5−9 did do so. In contrast, execution times revealed little sensitivity of SESAMI H to constraint relaxation. Note that a single structure was produced in the case of stenocarpine (Figure 3b) but two for 1-demethylwinkleridine (Figure 3, parts c and d, the latter of which was reported to be correct).

In addition to spectral data—1D ^{1}H NMR and ^{13}C NMR signals, and the results of only HMQC and HMBC experiments—the input for the guttiferone problem included two user-entered substructures identified by the authors of the original paper: a 3,4-dihydroxybenzoyl group with four of its carbon atoms assigned to observed ^{13}C NMR signals, and an enolic 1,3 dicarbonyl unit (O═C−C═C−OH) all of whose carbon atoms are also assigned. Such chemical shift assignments in user-entered, required substructures increase the efficiency of structure generation in both COCOA and HOUDINI.

From the 28 HMQC and 88 HMBC correlations reported for guttiferone (Table 1), INFER2D produced 55 paired structural inferences (C−C or C−A−C). As in an earlier case, a striking difference in performance between the two programs was observed. Whereas SESAMI H demonstrated very practical execution times, SESAMI C failed to complete a solution after several days of computational time (Table 5, numbers 4). Once again, little sensitivity to relaxation of constraints is apparent in SESAMI H. In contrast to the stenocarpine and 1-demethylwinkleridine problems, the absence of COSY data in this problem precludes the generation of explicit carbon connectivity information by INFER2D. The four structures generated by HOUDINI in each of the four runs are two pairs of tautomers. Only one of each pair is shown in Figure 3, parts e and f, of which the latter is correct. The structure is the one of the enolic forms of a tricarbonyl compound. Although three, rather than two tautomers are possible for that structure, one is precluded by the entry of the benzoyl group with its carbonyl group. Since that is entered as a required substructure, all generated structures must possess that carbonyl group. It is interesting to note that if guttiferone were an unknown, its ^{1}H NMR spectrum offers a means to rank the two alternative structures. The observed coupling constants (information not currently utililized by SESAMI) favor structure f (Figure 3), the same as that assigned by the original investigators. Specifically, the H_b vinyl hydrogen of the five-carbon isoprene unit (Figure 3f) is reported as a double doublets of quartets.[9] Two of the coupling constants (J^3_{HH} = 8.0 and 6.6 Hz) are consistent with coupling to the two adjacent diastereotopic methylene hydrogens. Allylic coupling to one of the two terminal allylic methyl groups (J^3_{HH} = 1.2 Hz) completes the observed pattern. The vinyl hydrogen of the one isomeric, five-carbon unit of the alternative structure (H_a, Figure 3e) would not be expected to display such a pattern based on coupling constants of 8.0 and 6.6 Hz.

The input for the lacinan-8-ol problem included the full complement of NMR spectral data (Table 1).[10] However, the original paper describing the elucidation of the structure of lacinan-8-ol contained some conflicting information. The presence of two long-range COSY correlations was reported (W coupling; four-bond correlations), but the text and a table described different signal assignments for one of them. Applying a general treatment to a situation where such uncertainty exists, the bond range of *all* 22 reported COSY correlations was expanded from three intervening bonds to "three or four" intervening bonds, and the problem was run with two variations of applied constraints: all constraints enabled (standard edit) and five constraints disabled (5−9). In automated mode, a single structure was obtained in both runs (Figure 3g). SESAMI C and SESAMI H each arrived at the solution very quickly; however, the latter was faster by almost 2 orders of magnitude (Table 5, number 5, runs 1 and 2). For comparison, the problem was rerun with only the two correct long-range COSYs increased to four bonds (Table 5, number 5, runs 3 and 4). All runs completed in less than 0.1 s and produced the same structure.

The difference in the output of INFER2D between the runs in which the bond range of all COSYs is set to three or four (runs 1 and 2), and those in which only the two actual long-range COSYs are so set (runs 3 and 4) is significant and highlights the importance of unambiguous COSY data. With all COSYs ambiguous, no explicitly defined substructures are produced, and the only output of INFER2D is 40 inferences of the form C−C/C−A−C which are derived from COSY and HMBC data. In the runs in which only the two actual long-range COSYs are given ambiguous values (three or four), INFER2D gives rise to a *single* "substructure", actually, the complete carbon skeleton of lacinan-8-ol. INTERPRET has done all of the work, and this is reflected in the less than 0.1 s execution time for both SESAMI C and H. Furthermore, the ACF Shortlist has been reduced to a single ACF entry in each family by INTERPRET. However, in runs 1 and 2, a normal ambiguous ACF Shortlist and a set of 40 ambiguous inferences must be processed by the structure generator. Here, the recorded execution times again reveal the greater efficiency of HOUDINI.

One of the objectives in developing a new structure generator was the creation of a version of SESAMI more

routinely applicable to compounds of higher molecular weight. α-Botryoxanthin[11] ($C_{74}H_{112}O_2$) with a molecular weight of 1032 daltons and extensive reported 2D NMR data—largely HMQC and HMBC correlations (Table 1)—represented an ideal test case. The input to the two SESAMI programs consisted of the molecular formula and all of the reported NMR spectral data: the chemical shift and multiplicity of 74 ^{13}C NMR signals; the chemical shift and integral of 72 ^{1}H NMR signals (which account for all 112 hydrogen atoms none of which are exchangeable); 72 one-bond HMQC correlations; 227 two- or three-bond HMBC correlations; but only 4 three-bond COSY correlations. The latter give rise to three explicit (i.e., chemical-shift labeled), two carbon fragments (three carbon−carbon connectivities).

The problem was run in automated mode using the usual four variations of initially applied constraints (Table 5, number 6). With the standard edit set of constraints, an ACF Shortlist of 74 families of carbon-centered ACFs and one family of oxygen-centered ACFs (to serve as the ACF resource for the two oxygen atoms) resulted. Five of the carbon-centered families consisted of a single ACF and only three contained over 100 ACFs. The total ACF count was 2125. The interpretation of the collective 2D NMR data was expressed by INFER2D as a set of 133 inferences of which two are explicitly defined fragments derived from the limited COSY data. All of the rest are HMBC inferences, only two of which have been reduced to a single inference (C−A−C).

The difference in performance between SESAMI C and SESAMI H in the α-botryoxanthin problem is dramatic. In each of the four runs, SESAMI C was terminated after several days of computer time failed to yield a solution. In contrast, SESAMI H provided a solution in all runs in less than 10 s (Table 5, number 6). A single structure was generated in each of the four runs—the same as that reported by the original investigators[11] (Figure 3h)—again indicating that no equally compatible structure had been overlooked.

An examination of the number of different initial problem states that COCOA must treat in the α-botryoxanthin problem is revealing of the limitations of this structure generator. In the standard edit run (SE, Table 5, number 6), COCOA has to treat 192 initial problem states but failed to do so in 48 h of computer time. Relaxing five initial constraints (run AA, Table 5, number 6), which would be expected to increase the ambiguity of the hybridization states of carbon atoms and therefore the number of initial problem states, again led to an aborted failed run. In an AA run, but using COCOA *only* to generate initial problem states, over seven billion were generated when the program was aborted after 36 h. Clearly, COCOA's requirement of explicit designation of atom hybridization in each initial problem state is a major impediment to the efficient solution of unknowns with large number of atoms. In constrast, HOUDINI deals with such ambiguity far more efficiently.

Syringolide was isolated from the culture of a Pseudomonas bacterial strain by Professor James Sims at the University of California, Riverside. In this case, the elucidation of the structure of the unknown was a collaboration between SESAMI C and Professor Sims. The presence of the four substructures shown in Figure 4 was determined in his laboratory. Only one carbon atom—the hemiketal carbon atom (Figure 4d)—was assigned to a specific ^{13}C NMR

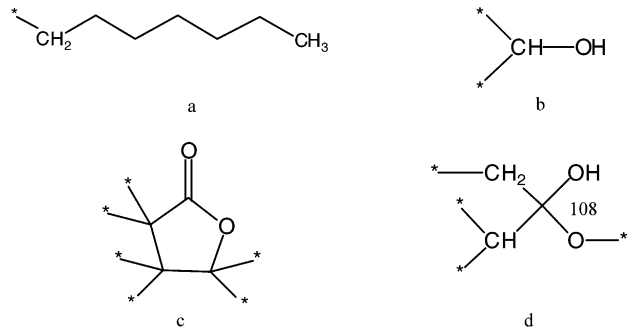

Figure 4. The four user-entered substructures in the Syringolide problem.

signal. These four substructures (including the one ^{13}C NMR signal assignment) were entered as input along with the molecular formula ($C_{l5}H_{24}O_6$) and the 1D and 2D NMR data (Table 1) derived from a study carried out at Arizona State University. However, the procedure for the HMBC experiment was modified to use four different time delays in the pulsed sequence. This permitted the assignment of each of the 21 observed HMBC correlations to one of two classes: those correlations limited to two or three intervening bonds between the coupled hydrogen and carbon atoms (16 correlations), and those correlations where the possibility of four intervening bonds could not be excluded (five correlations, set to two, three, or four intervening bonds). The assignment of two oxygen atoms as hydroxyl groups by INTERPRET was made possible by the identification of two rapidly exchangeable hydrogen atoms during collection of the ^{1}H NMR data. INFER2D produced 18 inferences: 2 explicitly defined fragments, 14 paired inferences (C−C or C−A−C), and 2 inferences (of the form C−C or C−A− or C−A−A−C) derived from the extended range HMBC correlations (Table 5, number 7).

SESAMI C and SESAMI H each rapidly generated two structures (Figure 3i,j) in automated mode with standard edit constraints and the four user-entered substructures (Table 5, number 7, run 1). Syringolide was assigned the latter structure by Professor Sims. Although SESAMI C and SESAMI H each solve the problem very rapidly (2.4 and 10 s, respectively), the former program is faster, the only such occurrence in the set of problems studied. This is because the initial version of HOUDINI has not as yet been optimized for efficient processing of user-entered required substructures without carbon chemical shift labels. Syringolide is the only problem in the set with a significant number of unlabeled carbon atoms in the user-entered substructures.

As expected, the deletion the four user-entered substructures from the input (run 2) reversed the order; SESAMI H's execution time was faster than that of SESAMI C by an order of magnitude. Put differently, deletion of the user-entered substructures adversely impacted the performance of COCOA (by a factor close to 300) significantly more than that of HOUDINI (a factor of 4). It should be noted that the number of generated structures is increased to 13 in the absence of the user-entered information.

SUMMARY AND CONCLUSIONS

Overall, the comparative tests revealed significant differences in performance between the HOUDINI and COCOA based systems. In most cases, significantly faster execution

times were observed for the former system. However, of greater significance is the observation that in every problem studied, a decrease in the amount of information and/or an increase in the ambiguitiy of the information provided to the structure generator—that is, increasing the size of the search space in a given problem—decreases the efficiency of structure generation to a far greater extent in COCOA than in HOUDINI. The greater susceptibility of COCOA to factors that can increase the size of the search space—e.g., the number of atoms in the unknown, amount and explicitness of information—led to a failure to produce a solution after several days of computational time in several cases. Thus, even though not fully developed and optimized, HOUDINI appears to demonstrate a greater capacity for efficiently solving the structures of large, complex organic compounds than COCOA. The results suggest that convergent structure generation does offer considerable promise in extending the range and scope of structure problems amenable to solution by SESAMI.

EXPERIMENTAL SECTION

All timings are based on an 1.4 GHz AMD Athlon with 1 GiB DDR266 memory running Windows 2000. The computer contained one 80 GiB Seagate Barracuda (ATA-IV) hard disk. Program COCOA (version 1.1.7) is written in ANSI C and compiled by gcc (cygwin32). Program HOUDINI is coded in C++ using Borland C/C++ 5.0 as compiler. Both compilations applied standard optimizations of the corresponding compiler without removing debugability. The sizes of the executables are 519 kB (COCOA) and 20 MB (HOUDINI).

ACKNOWLEDGMENT

The financial support of this research by the National Institutes of Health (Grant GM62457) is gratefully acknowledged. The authors express their appreciation to Professor James Sims at the University of California, Riverside for initiating a fruitful collaboration on the structure elucidation of Syringolide and to Dr. Ronald Nieman and Dr. Scott Smith of the Nuclear Magnetic Resonance Facility at Arizona State University for the quality of the critically important NMR spectra of this compound.

REFERENCES AND NOTES

(1) Korytko, A.; Schulz, K.-P.; Madison, M. S.; Munk, M. E. HOUDINI: A New Approach to Computer-Based Structure Generation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1434−1446.
(2) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87−93.
(3) Munk, M. E.; Velu, V. K.; Madison, M. S.; Robb, E. W.; Baderstscher, M.; Christie, B. D.; Razinger, M. Chemical Information Processing in Structure Elucidation. *Recent Advances in Chemical Information II*; Collier, H., Ed.; Royal Society of Chemistry: Cambridge, U.K., 1993; pp 247−263.
(4) Christie, B. D.; Munk, M. E. The Role of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Enhanced Structure Elucidation. *J. Am. Chem. Soc.* **1991**, *113*, 3750−3757
(5) Wasserman, H. H.; Ennis, D. S.; Vu, C. B.; Schulte, G.; Munk, M. E.; Madison, M. S.; Velusamy, K. V. Synthesis and Characterization of Pyrrolinecarboxylates by Reaction of Vicinal Tricarbonyl Derivatives with Aldehyde Schiff Bases. *Heterocycles* **1993**, *35*, 975−995.
(6) Munk, M. E. Computer-Based Structure Determination: Then and Now. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 997−1009.
(7) Almanza, G.; Bastida, J.; Codina, C.; de la Fuente, G. Norterpenoid Alkaloids from *Aconitella Hohenackeri*. *Phytochem.* **1997**, *45*, 1079−1085.
(8) Martin, G.; Mesia, L. R. Stenocarpine, A Diterpenoid Alkaloid from *Aconitella Stenocarpa*. *Phytochem.* **1997**, *46*, 1087−1090.
(9) Gustafson, K. R.; Blunt, J. W.; Munro, M. H. G.; Fuller, R. W.; McKee, T. C.; Cardellina II, J. H.; McMahon, J. B.; Cragg, G. M.; Boyd, M. R. The Guttiferones, HIV-Inhibitory Benzophenones. *Tetrahedron* **1992**, *48*, 10093−10102.
(10) Fukushi, Y.; Yajima, C.; Mizutani, J.; Tahara, S. Tricyclic Sequiterpenes from *Rudbeckia Laciniata*. *Phytochem.* **1998**, *49*, 593−600.
(11) Okada, S.; Tonegawa, I.; Matsuda, H.; Murakami, M.; Yamaguchi, K. Botryoxanthin B and α Botryoxanthin A from the Green Microalga *Botryococcus Braunii* Kawaguchi-1. *Phytochem.* **1998**, *47*, 1111−1115.

CI034058J