Application of Expert System CISOC-SES to the Structure Elucidation of Complex Natural Products

Chen Peng,† Shengang Yuan,*,† Chongzhi Zheng,† Yongzheng Hui,† Houming Wu,‡ and Kan Ma‡

Laboratory of Computer Chemistry and State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 354 Fenglin Lu, Shanghai 200032, People's Republic of China

Xiuwen Han

Dalian Institute of Chemical Physics, Chinese Academy of Sciences, P.O. Box 100, Dalian 116011, People's Republic of China

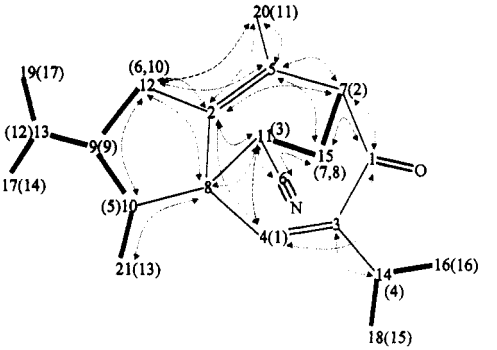
Received October 12, 1993®

This paper demonstrates the application of a newly developed computer-assisted structure elucidation system, CISOC-SES, to the structure determination of complex natural products. Discussed is the structure elucidation of four natural products, including one new compound, with up to 50 non-hydrogen atoms principally from their 2D NMR spectral data. The effectiveness of the novel approaches that exploit the direct and long-range distance constraints is analyzed in detail. The results show that the efficiency of structure elucidation can be substantially improved with the help of the system so that a manageable number of candidate structures, with the correct structure always included, can be obtained in a supportable period of CPU time.

INTRODUCTION

It is one of the most complicated works in the organic structure analysis to elucidate the structures of unknown compounds from their properties, typically from their spectroscopic properties. Modern high-resolution NMR spectroscopy, especially 2D NMR technology, has become an extraordinarily powerful tool for structure elucidation of unknown organic compounds, whether they be simple compounds or complex natural products. However, by now, the approach taken by a chemist to solve the problem of structure elucidation is mainly heuristic with his or her experience and intuition by tedious steps. With the rapid development of computer technology, it is natural that chemists look forward to elucidating the structures of complicated compounds with the help of the capabilities of huge storage and fast logic operation of computers. At the Institute, within the framework of the CISOC (Computerized Information System for Organic Chemistry) project, an expert system SES (Structure Elucidation System), which aims at fully exploiting 2D NMR spectral information, especially the long-range distance constraints (LRDCs) derived from HMBC or COLOC spectra, to determine the constitutional structure of complex natural products, is developed. The methodology used, the logic, and the structure of the system are described in the preceding paper in this issue. The system has been extensively tested by published and unpublished 2D NMR spectral data of natural products, and it is now under routine service within the Institute. By analyzing in detail the results of four typical examples of its application to solving real-world structural problems, this paper demonstrates some of the interesting features of the system, especially its high performance in the exploitation of LRDCs to dramatically reduce the CPU time for structure generation. Some conceivable further improvements of the system are also discussed.

† Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry.
‡ State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry.
® Abstract published in *Advance ACS Abstracts*, May 1, 1994.



1

Figure 1. Structure deduced by CISOC-SES for problem 1. Nonbracketed and bracketed numbers correspond to the numbering of the assigned ^{13}C and ^{1}H peaks (Table 1), respectively. Bold bonds correspond to the one-bond C–C connectivities derived from H,H COSY. Dotted and dashed curves represent long-range connectivities of 1–2 bonds and 2–3 bonds, respectively.

RESULTS AND DISCUSSION

Problem 1. The structure elucidation of **1** (Figure 1) with our system is described here in detail to illustrate the general procedures of using CISOC-SES. The manual elucidation of **1** depended mainly on 2D NMR data, particularly long-range H,C correlations derived from COLOC spectrum.[1] The molecular formula (MF) $C_{21}H_{29}NO$ was determined by elementary analysis and mass spectrometry (MS), while the existence of the carbonyl group, isopropyl group, double bonds, and cyano group was indicated by the IR spectrum. For clarity, the numbering of the carbon atoms and that of the protons (in brackets) in Figure 1 correspond, respectively, to that of their assigned 1D ^{13}C and ^{1}H NMR peaks listed in Table 1.

While CISOC-SES was used to deduce the constitutional structure of this compound from its spectral data listed in Tables 1–4, the MF, 1D ^{1}H, and ^{13}C data were analyzed in the first stage. As shown in Figure 2, one possible element group (EG) set was obtained and the implicit numbering of the carbon atoms was the same as that of the ^{13}C signals. The

0095-2338/94/1634-0814$04.50/0

Table 1. 1D ^1H and ^{13}C Spectral Data of **1**[a]

peak no.[b]	^1H shift (ppm)	^{13}C shift (ppm) (multiplicity)
1	6.59	196.06(s)
2	3.26	145.56(s)
3	3.04	144.65(s)
4	2.77	140.75(d)
5	2.53	123.40(s)
6	2.47	121.57(s)
7	2.22	56.28(d)
8	2.15	53.85(s)
9	2.05	48.16(d)
10	1.90	43.05(d)
11	1.79	34.66(d)
12	1.57	31.08(t)
13	1.11	28.27(d)
14	0.93	28.17(d)
15	0.92	27.78(t)
16	0.88	22.08(q)
17	0.87	21.64(q)
18		21.48(q)
19		21.31(q)
20		18.30(q)
21		11.74(q)

[a] As the they are currently not used in the system, the multiplicities and integrals of the ^1H peaks are not listed. The multiplicities of ^{13}C peaks are derived from the DEPT spectrum. [b] The numbering of ^1H peaks and that of ^{13}C peaks correspond, respectively, to the bracketed and nonbracketed designations of the carbon atoms in Figure 1.

Table 2. COSY Correlations of **1**

^1H shift (no.)	^1H shift (no.)	intensity level[a]	^1H shift (no.)	^1H shift (no.)	intensity level[a]
3.04(3)	6.59(1)	1	1.90(10)	2.05(9)	3
2.22(7)	3.26(2)	1	1.79(11)	2.47(6)	1
2.22(7)	3.04(3)	3	1.79(11)	1.90(10)	1
2.15(8)	3.26(2)	3	1.57(12)	2.05(9)	3
2.15(8)	3.04(3)	1	1.11(13)	2.53(5)	3
2.15(8)	2.22(7)	3	0.93(14)	1.57(12)	3
2.05(9)	2.53(5)	3	0.92(15)	2.77(4)	3
2.05(9)	2.47(6)	3	0.88(16)	2.77(4)	3
1.90(10)	2.47(6)	3	0.87(17)	1.57(12)	3

[a] Here, the "intensity levels" represent semiquantitatively the magnitudes of the H–H *J*-coupling constants, which are divided into three levels, with 1 standing for small (<3.0 Hz), 2 for moderate (3.0–6.0 Hz), and 3 for big ones (>6.0 Hz).

Table 3. HETCOR Correlations of **1**

^{13}C shift (no.)	^1H shift (no.)	intensity level[a]	^{13}C shift (no.)	^1H shift (no.)	intensity level[a]
140.75(4)	6.59(1)	3	27.78(15)	2.22(7)	3
56.28(7)	3.26(2)	3	27.78(15)	2.15(8)	3
48.16(9)	2.05(9)	3	22.08(16)	0.88(16)	3
43.05(10)	2.53(5)	3	21.64(17)	0.93(14)	3
34.66(11)	3.04(3)	3	21.48(18)	0.92(15)	3
31.08(12)	2.47(6)	3	21.31(19)	0.87(17)	3
31.08(12)	1.90(10)	3	18.30(20)	1.79(11)	3
28.27(13)	1.57(12)	3	11.74(21)	1.11(13)	3
28.17(14)	2.77(4)	3			

[a] The intensity levels are estimated from the peak areas in the spectral plot, with 1, 2, or 3 representing weak, moderate, or strong cross peaks, respectively.

total number of free bonds in the EG set was 60, on the basis of which a free-bond connection matrix (FBMX) of 60 × 60 could be constructed, but this was not actually done in this stage.

In the second stage, the 2D COSY, HETCOR, and COLOC cross peaks were interpreted and unified as a set of 41 ^{13}C–^{13}C signal–signal connectivities (SSCNs), which were directly transformed into 41 C–C atom–atom connectivities (AACNs) because the unknown is asymmetric. Note that two pseudo-

Table 4. COLOC Correlations of **1**

^{13}C shift (no.)	^1H shift (no.)	intensity level[a]	^{13}C shift (no.)	^1H shift (no.)	intensity level[a]
196.06(1)	6.59(1)	3	121.57(6)	2.15(8)	3
196.06(1)	3.26(2)	2	56.28(7)	3.26(2)	3
196.06(1)	2.15(8)	3	53.85(8)	3.04(3)	1
145.56(2)	6.59(1)	1	53.85(8)	2.47(6)	3
145.56(2)	3.26(2)	2	53.85(8)	1.11(13)	3
145.56(2)	3.04(3)	1	48.16(9)	2.05(9)	3
145.56(2)	2.47(6)	2	48.16(9)	1.11(13)	3
145.56(2)	1.90(10)	2	48.16(9)	1.11(13)	3
145.56(2)	1.79(11)	3	48.16(9)	0.93(14)	3
144.65(3)	2.77(4)	2	48.16(9)	0.87(17)	3
140.75(4)	6.59(1)	3	43.05(10)	2.53(5)	3
123.40(5)	3.26(2)	3	43.05(10)	2.47(6)	3
123.40(5)	2.22(7)	3	43.05(10)	1.11(13)	3
123.40(5)	2.15(8)	3	28.27(13)	1.57(12)	3
123.40(5)	1.90(10)	3	28.17(14)	2.77(4)	3
123.40(5)	1.79(11)	3	27.78(15)	2.22(7)	3
121.57(6)	3.04(2)	3	27.78(15)	2.15(8)	3
121.57(6)	2.22(7)	3	11.74(21)	1.11(13)	3

[a] The intensity levels are estimated from the peak areas in the spectral plot with 1, 2, or 3 representing weak, moderate, or strong cross peaks, respectively.

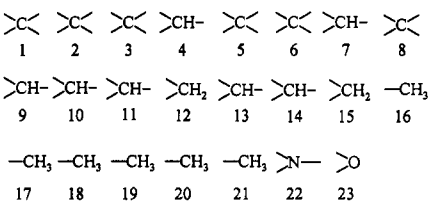


Figure 2. Set of element groups deduced for **1** and their implicit numbering. Note that the numbering of each carbon-centered element group corresponds to that of its corresponding ^{13}C peak (Table 1).

connectivities (with a distance constraint of 3–101 bonds, which would be transformed into a vague C–C distance of 1–99 bonds later) were added by the program because the ^1H chemical shift differences of two ^1H peak pairs (nos. 14 and 15, nos. 16 and 17; see Table 1) were very small (<0.02 ppm). An error message was given in this process because the two ^1H–^1H SSCNs (8–2: 3∼3 0)[2] and (7–2: 4∼5 0) led to two conflicting ^{13}C–^{13}C SSCNs concerning the same signal pair: (7–15: 1∼1 0) and (7–15: 2∼3 0). By editing the data file, we enlarged the distance of the latter ^1H–^1H SSCN to be 3–5 bonds. Then, the program unified them into a ^{13}C–^{13}C SSCN, (7–15: 1∼1 0). The ^1H–^1H SSCN pair (7–3: 3∼3 0) and (8–3: 4∼5 0) were also processed in the same way. At this point, the user can "tell" the program some known structural information by entering AACNs, especially those concerning heteroatoms derived from other spectral information (e.g., IR, UV, etc.). Here, the carbonyl group and the cyano group were added as two user-supplied AACNs in the same format as that of a spectra-derived AACN: (1–23: 1∼1 2) and (6–22: 1∼1 3). Note that here a carbon atom should be represented by its assigned ^{13}C peak number and each heteroatom by its atom number assigned while analyzing the MF (here 23 and 22 are respectively the atom numbers of the oxygen and the nitrogen atoms; see Figure 2). So, the current system uses user-supplied structural information in a limited way, which is in accordance with the initial purpose of the system: the least reliance on user-supplied structural information. In addition to the two user-supplied AACNs, 10 of the spectra-derived AACNs were also direct, on the basis of which 15 fixed bonds, illustrated as bold bonds in Figure 1, were extracted and a reduced FBMX of 30 × 30 was set up. After eliminating those already satisfied by the direct AACNs [e.g., (9–19: 1∼2 0) was satisfied by the two direct AACNs,

#EG		1	2	3	4	5	6	7	8	10	11	12	14	20
	#fb	1-2	3-6	7-10	11-13	14-17	18	19-20	21-24	25	26-27	28	29	30
1	1-2	0	1	1	13	1	1	25	1	1	13	1	1	1
2	3-6	1	0	1	13	1	1	13	1	1	13	13	1	13
3	7-10	1	1	0	1	1	1	1	1	1	1	1	13	1
4	11-13	13	13	1	0	1	1	0	1	0	0	0	0	0
5	14-17	1	1	1	1	0	1	25	1	1	13	13	1	13
6	18	1	1	1	1	1	0	13	1	1	25	1	1	1
7	19-20	25	13	1	0	25	13	0	1	0	0	0	0	0
8	21-24	1	1	1	1	1	1	1	0	13	13	13	1	1
10	25	1	1	1	0	1	1	0	13	0	0	13	0	0
11	26-27	13	13	1	0	13	25	0	13	0	0	0	0	0
12	28	1	13	1	0	13	1	0	13	13	0	0	0	0
14	29	1	1	13	0	1	1	0	1	0	0	0	0	0
20	30	1	13	1	0	13	1	0	1	0	0	0	0	0

Figure 3. Free-bond connection matrix deduced for **1**. For clarity, those element groups having no free bonds are ignored and the entries between the free bonds that belong to the same pair of element groups are condensed to be one, with #EG and #fb representing, respectively, the numbering of the element groups and free bonds.

(9–13: 1∼1 0) and (13–19: 1∼1 0)], 21 significant long-range AACNs remained and were used as long-range distance constraints (shown as curves in Figure 1) in the subsequent structure generation stage. On the basis of these LRDCs, the possible connections (entries of 1) in the FBMX were weighted with $W = 24$ and a weighted FBMX was obtained, as shown in Figure 3.

At the outset of the structure generation stage, the user can modify the value of some parameters to control the generation process in many respects, such as the direction of structure generation (GEN_FLAG), the demanded rate of LRDC satisfaction (K_L), the maximum allowed number of violated LRDCs (MAX_ERR_LRDC), the maximum tolerable deviation of ^{13}C chemical shift (ADD_C13_RNG), etc. Practically, the user can specify a large K_L (e.g., 1.5) and the structure generation will be iteratively carried out, with K_L reduced by ΔK_L (e.g., 0.1) in each cycle, until at least one structure is generated in a certain cycle (this value of K_L is designated as K_L^M) or K_L becomes zero. In this experiment, MAX_ERR_LRDC and ADD_C13_RNG were all set to zero (default values), but some parameters were varied to identify their influences on the efficiency of structure generation. Under these conditions, the single correct structure **1** was always obtained, but the CPU time for structure generation varies considerably (Table 5). It can be seen that the efficiency of structure generation is significantly improved when the search tree is properly weighted and rearranged on the basis of LRDCs (conditions No. 3 and 5 in Table 5).

Moreover, variation of the CPU time for structure generation versus K_L when GEN_FLAG = 2 and $W = 24$ is illustrated in Figure 4 to demonstrate the effectiveness of evaluation of intermediate structures based on the rate of LRDC satisfaction. It can be seen that the CPU time for structure generation decreases exponentially as K_L increases, but when $K_L > K_L^M$, the correct structure is lost. So, K_L^M is essential to the efficiency of structure generation. By the way, if the two functional groups were unknown, the same single structure would be obtained in less than 10 min of CPU time when GEN_FLAG = 1 or 2.

Table 5. CPU Time for Structure Generation of **1** under Different Conditions[a]

no.	GEN_FLAG[b]	W[c]	K_L^M[d]	CPU time (s)
1	0			1,399
2	1	0	0.2	490
3	1	24	0.9	15
4	2	0	0.7	1,226
5	2	24	1.1	42

[a] Except for those explicitly specified in the table, default values are adopted for all parameters. Under the five conditions, the correct structure alone is obtained. [b] GEN_FLAG determines the direction of structure generation, with 0, 1, and 2 for random-direction, CCSS-first, and connectivity-first structure generation, respectively. [c] W is a parameter used to weight the FBMX when GEN_FLAG = 1 or 2. If $W = 0$, then FBMX is actually not weighted. [d] K_L^M is the maximum slope of the cutoff line used to evaluate intermediate structures based on the rate of LRDC satisfaction when GEN_FLAG= 1 or 2.

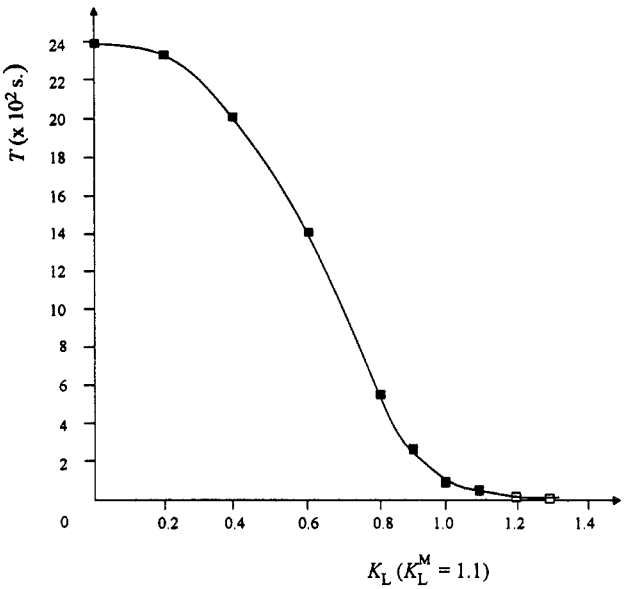
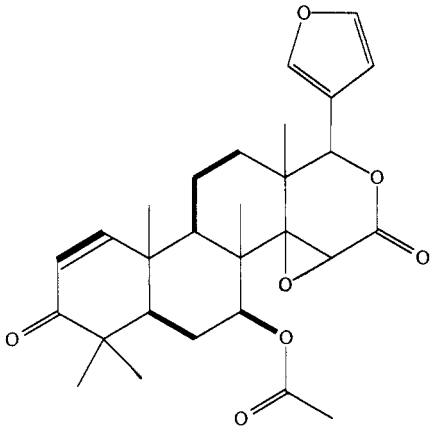


Figure 4. Graph of T, the CPU time for structure generation of **1**, versus K_L, the demanded rate of LRDC satisfaction when GEN_FLAG = 2. Note that structure **1** is obtained uniquely when $K_L \leq 1.1$ (designated as filled squares) and no structure is obtained when $K_L > 1.1$ (designated as empty squares).

The structural diagram of the resulting structure was displayed on a graphics terminal, with the carbon atoms depicted as the numbers of their assigned ^{13}C peaks (like Figure 1). The expected ^{13}C chemical shift ranges of the carbon atoms were listed together with the connection table of the structure. If any LRDCs are violated, they will be prompted in comparison with the actual distances between the concerned atom pairs. Here all the LRDCs were satisfied.

Problem 2. The structure elucidation of natural product **2**, whose MF is $C_{28}H_{34}O_7$, served as an exercise in an NMR workbook.[3] The available NMR spectral data included ^1H, ^{13}C, DEPT, COSY, HETCOR, and COLOC spectra and some NOE difference spectra. Besides, IR and UV spectra indicated the presence of an α,β-unsaturated ketonic functionality and a furane ring. The (manual) deduction of the structure was claimed to be a demonstration that "instead of depending on empirical hints, a complete ^1H and ^{13}C signal assignment of complicated structures can be reliably based on spectroscopic proofs."[3]

While our system was used to solve this problem, the MF and 1D NMR spectral data were first entered and one possible EG set with a total of 92 free bonds was obtained. Compared with problem 1, much less direct AACNs were derived from

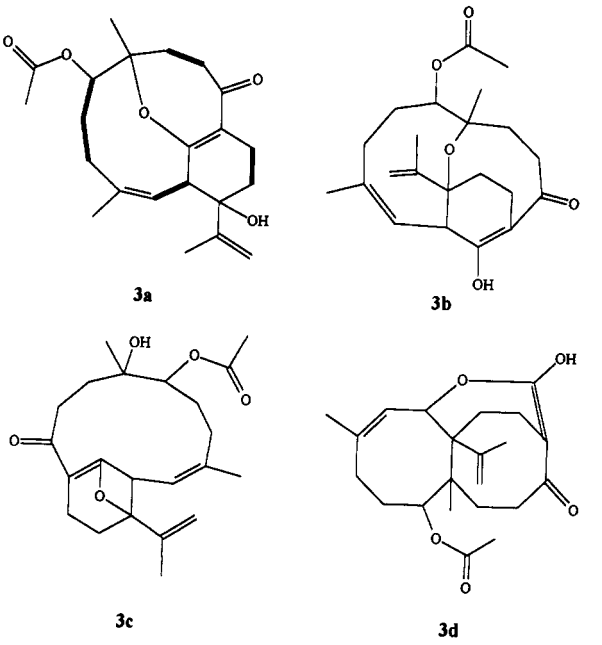


2

Figure 5. Target structure of problem 2.[3] The bold bonds are those that are indicated by H,H COSY correlations.

the COSY spectrum (only the five bold bonds in Figure 5). To further reduce the search space, the following fragments were entered as user-supplied AACNs: a carbonyl group at δ = 204.00 ppm, two carboxy groups at δ = 169.90 and 167.40 ppm, and a furane ring. On the other hand, 30 LRDCs with a distance of 1–2 bonds were derived from the COLOC spectrum, and one LRDC with a distance of exactly two bonds was derived from a weak COSY peak (implying a C–C distance of 2–3 bonds) and an HMBC peak (implying 1–2 bonds) concerning the same carbon atom pair. A reduced FBMX of 52 × 52 was thus obtained. The structure generation gave 10 candidate structures in a CPU time of 455 s when GEN_FLAG = 2 and K_L^M = 1.3. These candidates include the reported correct structure **2** and the other possible alternative that was excluded on the basis of NOE evidence in the literature.[3] For comparison, the structure generation took over 1289 s of CPU time to give 12 candidates when GEN_FLAG = 1 and K_L^M = 1.0. Note that the two additional candidates, which were fully compatible with the LRDCs and chemical shift constraints, were neglected in the former case because of the low rate of LRDC satisfaction in their generation paths. If K_L = 1.2 was used when GEN_FLAG = 2, the same 12 structures were obtained, but the CPU time was much longer (3496 s.).

Problem 3. A new natural product (MF: $C_{22}H_{30}O_5$) was studied independently by hand and by the system. Similar to problem 2, very sparse direct AACNs were available from COSY because of the existence of many quaternary carbon atoms. On the other hand, 55 cross peaks were picked from the three HMBC spectra acquired with different delays for evolution of long-range couplings and were transformed into 30 C–C LRDCs. As for user intervention, a carbonyl group and a carboxy group with very characteristic resonances at δ = 200.14 and 169.82 ppm, respectively, were entered as user-supplied AACNs. These boiled down to an FBMX of 48 × 48. The first run of the structure generator gave no result, and we doubted that some of the HMBC peaks might have been erroneously picked because some HMBC peaks were very difficult to discern, especially in the ^{1}H dimension. By defining MAX_ERR_LRDC = 1, so that violation of one LRDC was permitted for each generated structure, the structure generation gave three candidate structures **3a–c** as shown in Figure 6, each with the same violated LRDC, (1–15: 1~2 0). Candidate **3a** coincides with the manually deduced structure. Further scrutiny of the HMBC spectra found that the used cross peak at (δ_1 = 200.14 ppm, δ_2 = 1.92 ppm),



3a **3b**

3c **3d**

Figure 6. Candidate structures deduced by CISOC-SES for problem 3. The bold bonds in **3a**, which is the correct structure, correspond to the five direct connectivities indicated by H,H COSY correlations.

which gave rise to the violated LRDC, was indeed located at (δ_1 = 200.14 ppm, δ_2 = 1.95 ppm). The error occurred because the two proton multiplets at δ = 1.92 and 1.95 ppm were seriously overlapping and the line width of this cross peak was big (ca. 0.05 ppm) in the ^{1}H dimension. After this cross peak was corrected, 29 LRDCs were obtained because this corrected cross peak gave a redundant C–C AACN, as did the cross peak between this carbon atom and a geminal proton of the one at δ_2 = 1.95 ppm. Then, with MAX_ERR_LRDC = 0, GEN_FLAG = 2, and K_L^M = 1.2, the structure generation gave the same three candidate structures **3a–c** in a CPU time of 1433 s. For comparison, the structure generation took over 159 h of CPU time to give the four structures in Figure 6, including the additional candidate **3d**, if GEN_FLAG = 1, wherein K_L^M is only 0.6.

As illustrated in Figure 4, the magnitude of K_L^M is essential to the efficiency of structure generation, which is commonly the bottleneck of computer-assisted structure elucidation systems. In the case of a simple problem with M, the size of the FBMX, being less than about 30 (e.g., problem 1 and the subsequent problem 4), the CCSS-first structure generation (GEN_FLAG = 1) shows approximately the same efficiency as or a slightly higher efficiency than the connectivity-first structure generation (GEN_FLAG = 2). When M grows bigger (e.g., problems 2 and 3), however, the constraining power of chemical shifts decreases rapidly since there are more carbon atoms with similar structural environments in a larger molecule. In contrast, the number of LRDCs increases in parallel to the size of a molecule. So, LRDCs become the dominating constraints on structure generation when M is big. As connectivity-first structure allows greater K_L^M to be used, it shows significantly higher efficiency. Figure 7 illustrates the variation of L_s, the number of satisfied LRDCs, versus B, the number of chosen connections, in the generation paths of structure **3a** under the two directions. It can be seen that L_s grows quicker when GEN_FLAG = 2 (designated as empty circles) than when GEN_FLAG = 1 (designated as black squares), so a cutoff line with a greater slope (K_L^M = 1.2) can be used in the former case than in the latter case (K_L^M = 0.6). By the way, if the conventional structure generation

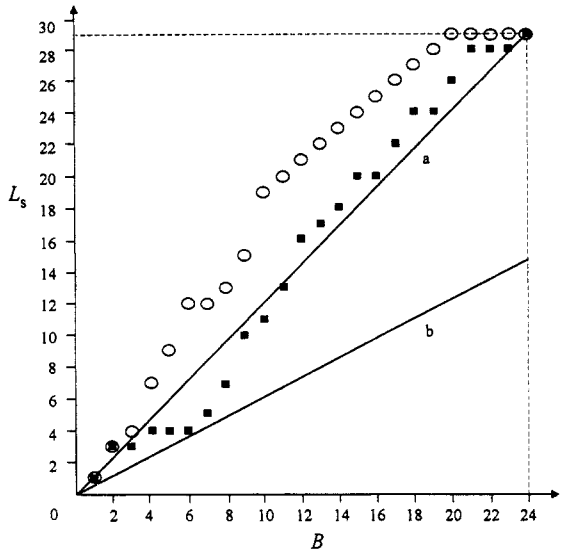


Figure 7. Illustration of L_s, the number of satisfied LRDCs, versus B, the number of chosen connections in the structure generation paths of **3a** when GEN_FLAG = 2 (designated as empty circles) and when GEN_FLAG = 1 (designated as filled squares). Line a ($K_L^M = 1.2$) and line b ($K_L^M = 0.6$) are the cutoff lines with the maximum slopes used under the two conditions, respectively.

(i.e., GEN_FLAG = 0) is used when M is big, the CPU time would be forbiddenly long because it is equivalent to the case where $K_L = 0$.

Another problem associated with the use of K_L^M is whether or not the correct structure can be always produced when K_L^M is used or whether the generation path of the correct structure has the greatest rate of LRDC satisfaction. This has not been mathematically proved yet, but in all the problems coped with by our system so far, the correct structure has not been lost when K_L^M is used, though sometimes more candidate structures are given when $K_L < K_L^M$. Anyway, a conservative user can use a small or even zero K_L to pursue a more or completely exhaustive structure generation.

Problem 4. This example illustrates the system's efficient use of direct AACNs when 2D INADEQUATE correlations are available. The key step in the structure determination of **4** (MF: $C_{40}H_{68}O_8$) was the identification of the cross peaks in the 2D INADEQUATE spectrum by using a spectrum evaluation program because the signal-to-noise ratio of the spectrum was too low to be identified by the naked eyes.[4] The acquisition of the INADEQUATE spectrum was optimized in such a way that only single C–C bonds were manifested in the spectrum. In addition to the complete ^{13}C, DEPT, and 2D INADEQUATE spectra, several COSY, HMBC, and INAPT connectivities were also used in the (manual) determination of the structure.[5]

All of these data, together with three user-supplied carbonyl groups at δ = 198.86, 173.36, and 175.12 ppm, respectively, were submitted to our system, and a small FBMX (30 × 30) was obtained after 44 fixed bonds were extracted. If the single C–C bonds were forbidden to be formed in the structure generation, three candidates, **4a–c** (Figure 8), including the correct one, **4a**, were given out in less than 10 s, irrespective of the direction of structure generation (GEN_FLAG = 1, 2, or 0). If the single C–C bonds were allowed to be formed, the same results were obtained in a CPU time of about 12 s if GEN_FLAG = 1 or 2 ($K_L^M = 1.8$ in both cases), or in a CPU time of 1906 s if GEN_FLAG = 0. Though only five LRDCs were used in this problem, efficient use of them

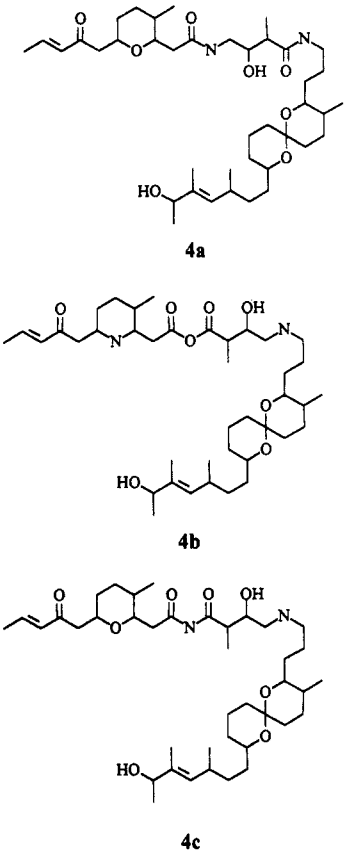


Figure 8. Three candidate structures deduced for problem 4 by CISOC-SES, with **4a** being the correct one.[5]

enhanced the performance of the structure generation significantly.

EXPERIMENTAL SECTION

The 1D and 2D NMR spectra of the new compound **3** were recorded at 600 MHz and processed on a Bruker-AMX 600 spectrometer. $CDCl_3$ and TMS were used as solvent and internal reference, respectively. The 2D correlated spectra, DQF-COSY, HMQC, and HMBC, were recorded with built-in pulse sequences cosydftp, invb, and inv41plrnd, respectively. All of their time dormain data sets were collected by using 1024 complex data points in t_2 dimension and 256 t_1 increments. The data sets were apodized with a sine bell window function along t_2 and t_1 and were zero-filled to 1024 (for DQF-COSY) or 512 (for HMQC and HMBC) data points in t_1 dimension prior to Fourier transformation. Three HMBC spectra were obtained with different delays for evolution of long-range coupling: 129, 81, and 55 ms, respectively.

CISOC-SES is run on a microVAX 3300 computer. The 1D and 2D NMR spectral data of a molecule were entered via a data file prepared after manual extraction of 1D peaks and 2D cross peaks from the spectral plots. In the file each cross peak was represented by its associated 1D peaks in both dimensions and an intensity level, such as those illustrated in Tables 1–4. For a COSY peak where an accurate J-coupling constant was available, the intensity level represents semi-quantitatively the magnitude of the observed J-coupling constant, with 1 standing for a small (e.g., <3.0 Hz), 2 for a moderate (e.g., 3.0–6.0 Hz), and 3 for a large coupling constant (e.g., >6.0 Hz). In other cases where no accurate J-coupling constant was available, on the other hand, the intensity levels were estimated from the peak areas in the spectral plots.

CONCLUSION

We described here the application of CISOC-SES to the structure elucidation of four natural products of practical complexity to demonstrate its current capability to solve real-world problems. It can be concluded from the results that, with our program, the efficiency of the structure elucidation of complex natural products can be significantly improved so that (1) if the available data are free of error, the correct structure can be always guaranteed, and (2) if the conventional 2D NMR spectral data are employed, a manageable number (e.g., less than 50) of candidate structures can be given out in a supportable period of CPU time (e.g., a few minutes to several hours). As computers are nowadays widely used on modern NMR spectrometers for data acquisition and basic data processing such as base line correction and Fourier transformation (FT), CISOC-SES, with its high efficiency in transforming 2D NMR spectral data into plausible chemical structures, can serve as an additional built-in post-FT data processor on an NMR spectrometer or serve as a stand-alone NMR processing software package.

On the other hand, the practical application of the system, particularly to unknown molecules such as problem 3, gave us many hints on the further improvement of the system. First, accurate peak picking is very important for efficient structure elucidation. Even in the spectra of a moderate-sized molecule, peak overlapping may be serious and the peak picking itself demands a lot of chemical reasoning to resolve ambiguities. So, it would be much better to combine with an interactive semiautomatic peak-picking facility to fully use the potential resolution of the spectral data file, instead of measuring on paper plots. Second, more comprehensive use of ^{13}C and ^{1}H chemical shifts is imperative for evaluating the candidate structures. Moreover, combination with some stereoisomer enumerating[6,7] and molecular modeling programs[8] is important to further study the candidate structures based on spatial proximities provided by NOESY spectra. Further improvements of the system in these respects will make it a more versatile and more helpful assistant to chemists in a variety of applications.

MAIN ABBREVIATIONS AND ACRONYMS USED IN THE TEXT

1D	one-dimensional
2D	two-dimensional
AACN	atom–atom connectivity
ADD_C13_RNG	tolerable deviation between observed and predicted ^{13}C chemical shifts
B	number of currently chosen connections at a certain step in the structure generation path
B_0	total number of connections to choose to generate a complete structure
CCSS	carbon-centered single-spherical substructure
CISOC-SES	computerized information system of organic chemistry-structure elucidation subsystem
COSY	correlated spectroscopy
COLOC	correlation via long-range coupling
CPU	central processing unit
DEPT	distortionless enhancement by polarization transfer
EG	element group, i.e., a non-hydrogen atom bearing a definite number of protons
FBMX	free-bond connection matrix
GEN_FLAG	parameter that determines the direction of structure generation (if GEN_FLAG = 0, 1, or 2, the structure generation is carried out in random, CCSS-first, or connectivity-first direction, respectively)
HETCOR	heteronuclear correlation
HMBC	heteronuclear multibond connectivity
HMQC	heteronuclear multiple quantum coherence
INADEQUATE	incredible natural abundance double quantum transfer experiment
LRDC	long-range distance constraints
K_L	demanded rate of LRDC satisfaction
K_L^M	maximal value of K_L, with which at least one candidate structure is generated
L_s	number of actually satisfied LRDCs at a certain step in the structure generation path
M	size of FBMX, which equals $2B_0$
MAX_ERR_LRDC	maximum allowed number of violated LRDCs
MF	molecular formula
NOE	nuclear Overhauser enhancement
SSCN	signal–signal connectivity
W	parameter used to weight FBMX based on LRDCs

REFERENCES AND NOTES

(1) Han, X. W.; Rüegger, H.; Sonderegger, J. The Stereostructural Investigation of 3(R),8-Dimethyl-2(R),5-Bis(1-Methyl-Ethyl)-6-Oxo-2,3,6,7,9,10-Hexahydro-1H-3A,7-Ethanoaulene-10-Carbonitrile by 2D NMR Spectroscopy. *Chin. Sci. Bull.* **1990**, *35*, 997–1002.

(2) From left to right, the five numerals represent, respectively, the numberings of the two atoms (or signals) concerned, the lower and upper limits of the distance (i.e., the minimum and maximum numbers of intervening bonds), and the type of the bond (with 0 representing a meaningless or unknown type).

(3) Duddeck, H.; Dietrich, W. *Structure Elucidation by Modern NMR, A Workbook*; Springer-Verlag: New York, 1989; pp 137–147, 229.

(4) Dunkel, R.; Mayne, C. L.; Pugmire, R. J.; Grant, D. M. Improvements in the Computerized Analysis of 2D INADEQUATE Spectra. *Anal. Chem.* **1992**, *64*, 3133–3149.

(5) Foster, M. P.; Mayne, C. L.; Dunkel, R.; Pugmire, R. J.; Grant, D. M.; Kornprobst, J.-M.; Verbist, J.-F.; Biard, J.-F.; Ireland, C. M. Revised Structure of Bistramide A (Bistratene A): Application of a New Program for the Automated Analysis of 2D INADEQUATE Spectra. *J. Am. Chem. Soc.* **1992**, *114*, 1110–1111.

(6) Nourse, J. G. The Configuration Symmetry Group and Its Application to Stereoisomer Generation, Specification, and Enumeration. *J. Am. Chem. Soc.* **1979**, *101*, 1210–1216.

(7) Nourse, J. G.; Carhart, R. E.; Smith, D. H.; Djerassi, C. Exhaustive Generation of Stereoisomers for Structure Elucidation. *J. Am. Chem. Soc.* **1979**, *101*, 1216–1223.

(8) Ripka, W. C.; Blaney, J. M. Computer Graphics and Molecular Modeling in the Analysis of Synthetic Targets. In *Topics In Stereochemistry*; Eliel, E. L., Wilen, S. H., Eds.; John Wiley & Sons: New York, 1991; Vol. 20, pp 1–85.