J. Chem. Inf. Comput. Sci. 1995, 35, 539−546

Practical Computer-Assisted Structure Elucidation for Complex Natural Products: Efficient Use of Ambiguous 2D NMR Correlation Information

Chen Peng, Shengang Yuan,* and Chongzhi Zheng

Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 354 Fenglin Lu, Shanghai 200032, People's Republic of China

Zhengshuang Shi and Houming Wu

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

Received December 16, 1994[⊗]

If a computer-assisted primary structure elucidation expert system is to tackle structural problems of real-world complexity, it must be able to cope with the ubiquitous ambiguous 2D NMR correlation information, especially the ambiguous-node(signal or atom)-involved (topological) distance constraints (ANDCs). Such ambiguities of the distance constraints result typically from spectral imperfections and molecular symmetry. As a significant improvement of CISOC-SES, an expert system that automatically deduces the constitutional structure for unknown organic compounds from their molecular formula and conventional one- and two-dimensional NMR spectral information, this paper reports our novel approaches to the representation and use of such ambiguous distance constraints for this purpose. The high performance of the improved CISOC-SES is demonstrated by the structure elucidation of alborixin from its molecular formula ($C_{58}H_{84}O_{14}$) and ^1H, ^{13}C, COSY, HMQC, HMBC, and NOESY spectral data.

INTRODUCTION

Computer-assisted determination of the constitutional structure of organic compounds, which is generally referred to as computer-assisted structure elucidation (CASE) of organic compounds, has been one of the earliest applications of artificial intelligence in chemistry.[1] A CASE expert system is supposed to help a chemist interpret spectral and/or chemical data more thoroughly and more efficiently, but for unknown compounds with real-world complexity, practical performance of such a system has been achieved only when 2D NMR correlation information was employed.[2-4] The major reason is that the core of a CASE system, the structure generation algorithm, is inevitably of exponential time complexity, thus easily leading to combinatorial explosion if the molecule is big. Compared with 1D NMR spectra, 2D NMR experiments provide more abundant and more unambiguous structural information, typically in the form of through-bond or through-space atom−atom connectivities (often referred to as "2D correlation information").[5] As such correlation information can be used to effectively narrow the search space and to guide the search in structure generation, the employment of (through-bond) correlation information usually dramatically increases the efficiency of CASE. In the recently reported method of Bangov et al., the 2D correlation information is used to limit the number of extension nodes at each level of the search tree, thus pruning the fruitless branches and reaching the target structure more efficiently.[6] It seems to us, however, that the interactive selection of the extensions is impractical, because this would be too tedious to be coped with by a human save for the simplest cases where sufficient direct C−C connectivities (e.g., those from COSY or INADEQUATE spectra) are available to exclusively guide each bond-formation step. Our recently developed expert system CISOC-SES uses 2D NMR correlation information in a more systematic fashion and exhibits elegant performance in the structure elucidation of several natural products of real-world complexity.[3,4] In CISOC-SES, the 2D NMR through-bond connectivities are first interpreted as (topological) distance[7] constraints (DCs) on the associated 1D signals, which are finally interpreted as DCs on atoms. One of the features of CISOC-SES is that it can efficiently use DCs of an ambiguous number of bonds to achieve high efficiency in the (constitutional) structure elucidation of complex natural products. Such ambiguous-distance-involved DCs (ADDCs) are typically derived from HMBC/HMQC spectral combination and imply C−C distances of either one or two bonds, so we also refer to them as "long-range distance constraints" (LRDCs). Like all other similar systems, CISOC-SES implicitly assumed that all the 2D cross peaks are well-resolved, and each of the derived DCs corresponds unambiguously to two associated atoms. In practice, however, spectral imperfections (e.g., low-digital resolution, artifacts, resonance degeneracy, and peak missing) and molecular symmetry make it impossible to reach this requirement. As a result, the structure elucidation of complex molecules like alborixin (1) proved to be a big challenge to CISOC-SES not only because of the big molecular size but also because the severe ^1H resonance degeneracy leads to a lot of ambiguous DCs not usable to it. This paper presents a general solution to the representation and efficient use of such ambiguous 2D NMR correlation information for automated spectral interpretation. The impressively high performance of the improved CISOC-SES is demonstrated by its application to the structure elucidation of alborixin on the basis of its conventionally used NMR spectral data and MF.

[⊗] Abstract published in *Advance ACS Abstracts*, April 15, 1995.

0095-2338/95/1635-0539$09.00/0

Chart 1. Main Abbreviations and Acronyms Used in the Text

AADC	atom−atom distance constraint
ADC	ambiguous distance constraint, including ANDC and ADDC
ADDC	ambiguous-distance-involved distance constraint
ANDC	ambiguous-node-involved distance constraint where a node represent a signal or an atom
CASE	computer-assisted structure elucidation
CISOC-SES	computerized information system of organic chemistry-structure elucidation subsystem
COSY	correlated spectroscopy
CPU	central processing unit
DC	(topological) distance constraint
DEPT	distortionless enhancement by polarization transfer
FBMX	free-bond connection matrix
FBx	a candidate free bond to connect *FBy* of a level in the search tree.
FBy	a free bond specific to a level in the search tree.
HMBC	heteronuclear multi-bond connectivity
HMQC	heteronuclear multiple quantum coherence
INADEQUATE	incredible natural abundance double quantum transfer experiment
K_A	rate of ADC-checking of a structure generation path
LRDC	long-range distance constraints, a typical kind of ADDC
MF	molecular formular
N_FBX_STEP	maximum average number of (carbon-incident) FBx's to be searched at each level in the search tree of structure generation
sDC	"single distance constraint", i.e., one of the concerned node pair(s) of a DC
SSDC	signal-signal distance constraint.

METHODS

Main Features of CISOC-SES. The main features of CISOC-SES, which have been discussed in detail in previous papers,[3,4] are summarized here to help the reader better understand the present improvements. With the aim of truly assisting chemists in their structure elucidation of real-world complex natural products, the design of CISOC-SES have addressed the two essential problems: (1) least reliance upon user-supplied structural information to avoid subjective biases and (2) efficient use of conventional 2D NMR data, especially the information-rich HMBC data (i.e., LRDCs), to achieve high-efficient structure generation. The structure elucidation process with CISOC-SES consists of the following three logical phases. In the first phase, on the basis of the MF (and ^{13}C and DEPT spectral data if available), one or more initial free-bond connection matrices (FBMXs) are constructed to represent the bonding possibilities between the unsatisfied valences (called *free bonds*) of the non-hydrogen atoms. The FBMX actually forms the search space of the subsequent structure generation. The use of the FBMX which is characterized by its simplicity and flexibility in many aspects originates from the idea of Munk[8] and has been incorporated in GEN of Bohanec and Zupan.[9] For example, the hybridization states of the constituent atoms, which generally cannot be unambiguously determined from NMR spectral data, need not be explicitly specified. Exhaustive generation of structures, both acyclic and cyclic, can be achieved simply by a bond-searching algorithm which chooses exactly one nonzero element (standing for a *possible connection* between two associated free bonds) from each row and each column. What is more, the FBMX is most suited for the use of 2D NMR-derived correlation information, because it is possible to construct an explicit correspondence between the constituent carbon atoms and the observed ^{13}C resonances. In the second phase, the various kinds of 2D NMR spectral data are systematically interpreted and transformed into a unified set of topological distance constraints on ^{13}C signal pairs (called signal−signal distance constraints or SSDCs). Based on the molecular symmetry, the SSDCs are finally transformed into atom−atom distance constraints (AADCs). Among the AADCs, those with one-bond distance, such as the C−H connectivities derived from DEPT/^{13}C and the C−C ones from COSY/HMQC or 2D INADEQUATE spectra, and if available, user-supplied AADCs, are extracted as fixed bonds, thus reducing the number of free bonds. Meanwhile, some of the nonzero elements in the FBMX are also eliminated by cross checking with all the distance constraints. All this in effect reduces the search space of the subsequent structure generation. In the final structure generation phase, the AADCs with ambiguous distances (i.e., LRDCs), most of which are derived from HMBC/HMQC spectra, are used prospectively to guide and constrain the structure generation. This is realized through some novel approaches such as weighting the FBMX to reorganize the search tree and evaluation of intermediate structures based on the "rate of LRDC-satisfaction". The ^{13}C chemical shifts as well as some other simple heuristics are used as supplementary criteria to evaluate the intermediate structures produced during the structure generation. The heuristics that use DCs with both direct and ambiguous distances prove to be very effective in reducing the CPU time for structure generation, which is generally the bottleneck of CASE.

The previous implementation of CISOC-SES accepted a 2D NMR cross peak as the pair of associated 1D signals (together with a semiquantitative description of the magnitude of the *J*-coupling constant or the intensity of the cross peak) and used DCs on exactly two atoms. In practice, however, we often find that the ubiquitous spectral imperfections and, sometimes, the molecular symmetry make it impossible for us to assign unambiguously two nodes (1D signals or atoms) to a 2D-spectrum-derived DC. In such cases, the ambiguous correlation information had to be resolved manually by the user (with some reasoning or by trial and error) or were simply ignored. Apparently, this is of low efficiency and may introduce subjective biases.

Ambiguous-Node-Involved Distance Constraints. Most spectral imperfections and molecular symmetry result in ambiguous-node-involved distance constraints (ANDCs). For example, because of low digital resolution or resonance degeneracy (fortuitous peak overlapping), it is very common that a cross peak in a 2D spectrum cannot be precisely located to a 1D signal in either dimensions. Moreover, molecular symmetry also introduces ambiguities to the interpretation of 2D correlation information. This is schematically illustrated in Figure 1 with the hypothetical HMQC and COSY spectra of a symmetrical phenyl moiety. Suppose that the resonances of H_2 and H_3 degenerate (giving an envelope $^1H_{2+3}$), then their two corresponding HMQC peaks cannot be resolved in the ^1H dimension, and a COSY peak between 1D signals $^1H_1$ and $^1H_{2+3}$ can only be interpreted as a one-bond DC on either $^{13}C_1$ and $^{13}C_2$, or $^{13}C_1$ and $^{13}C_3$, or both. When the DCs on the ^{13}C signals are mapped to the carbon atoms, the molecular symmetry further complicates the situation. As $^{13}C_1$ and $^{13}C_2$ each corresponds to two symmetrical carbon atoms, alternative interpretations of the $^{13}C-^{13}C$ DCs are possible, and ambiguity is introduced again. To our knowledge, symmetry caused ambiguity has

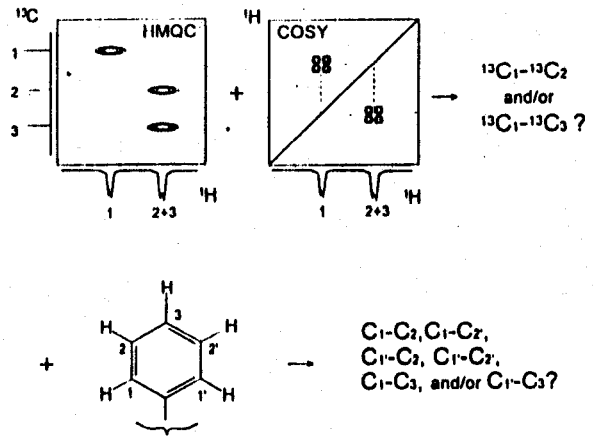


Figure 1. Schematic illustration of the ANDCs (ambiguous-node-involved distance constraints) derived from the hypothetical HMQC and COSY spectra of a phenyl moiety. Upper part: Because of the ^1H resonance degeneracy ($^1H_2$ and $^1H_3$), two HMBC peaks cannot be resolved in the ^1H dimension, so the COSY peak can only be interpreted as a one-bond distance constraint (DC) on either signal pair $^{13}C_1$ and $^{13}C_2$, or $^{13}C_1$ and $^{13}C_3$, or both. Lower part: As $^{13}C_1$ and $^{13}C_2$ each corresponds to two symmetrical carbons, the interpretation of these signal−signal DCs as atom−atom ones is further complicated and ANDCs are introduced again.

been considered in Munk's system;[10] however, there has been no report on using other ANDCs for the purpose of automated structure elucidation of natural products.

Representation of Ambiguous DCs. In order to use DCs with the various ambiguities for automated analysis, the following scheme is devised to represent a general distance constraint

$$(n_1^1 \cdots n_x^1 - n_1^2 \cdots n_y^2; d_1 - d_2; b; s_1 - s_2) S_1 \cdots S_m \quad (1)$$

where $n_1^1 \cdots n_x^1$ are the x ($x \geq 1$) possible nodes associated on one side of the DC, $n_1^2 \cdots n_y^2$ are the y ($y \geq 1$) possible nodes associated on the other side. d_1 and d_2 denote, respectively, the lower and upper limits of the distance, while b is the type of bond ($b = 0$, 1, 2, or 3 for unknown, single, double, or triple bond, respectively).[11] As the $(x + y)$ possible nodes imply xy possible pairs of nodes (called *single distance constraints* or sDCs), the lower and upper limits of the number of sDCs that must be satisfied in the target structure are represented as s_1 and s_2, respectively. Finally, the letters $S_1 \cdots S_m$ represent the m ($m \geq 1$) source spectra from which this DC is derived (e.g., "C" for COSY, "Q" for HMQC, and "B" for HMBC).

This scheme can be very flexibly used to represent a variety of ambiguous 2D NMR-derived DCs besides the previous LRDCs. For example, the distance constraint interpretation of the COSY/HMQC peaks in Figure 1 can be conveniently represented as a $^{13}C-^{13}C$ ANDC: (1−2 3: 1−1; 0; 1−2)CQ. The last two numerals specify that either one or both sDCs (i.e., 1−2 and 1−3) must be satisfied. The user can define it as 1−1 if one believes that only one of the two sDCs must be satisfied. Then, based on the molecular symmetry, this $^{13}C-^{13}C$ SSDC is mapped to a C−C AADC of (1 1′−2 2′ 3: 1−1; 0; 1−6)CQ. In addition to the ANDCs, it is interesting to note that scheme 1 can be used to represent another kind of common spectral ambiguity, i.e., ambiguity that lies in the existence of a cross peak itself. For example, a cross peak may be suspected to be an artifact. This can be conveniently represented as a DC with an sDC range of 0−1.

Efficient Use of Ambiguous DCs. As mentioned above, the various 2D cross peaks (e.g. COSY, HMQC, HMBC, and/or INADEQUATE ones) are first systematically interpreted as their corresponding SSDCs. If the picking of a cross peak is of ambiguous nature as described above, the user can modify its corresponding SSDC (so that x or $y >$ 1, or $s_1 = 0$) to accommodate this ambiguous nature. Then, these various SSDCs are transformed into a unified set of $^{13}C-^{13}C$ SSDCs on the basis of the one-bond $^1H-^{13}C$ connectivities derived from HMQC. This step is now very simply completed by replacing each ^1H signal in the SSDC with its directly connected ^{13}C signal(s). As shown in Figure 1, a ^1H signal may correspond to more than one ^{13}C signal in the case of ^1H resonance degeneracy and thus a $^{13}C-^{13}C$ ANDC is obtained. The set of $^{13}C-^{13}C$ SSDCs are finally mapped to atom−atom (C−C) ones on the basis of the molecular symmetry (deduced from ^{13}C peak number, ^1H integrals, and HMQC connectivities). In the simple case of an asymmetrical molecule, this step is straightforward, as each ^{13}C signal is assigned to a carbon atom with an identical index. In the case of a symmetrical molecule, however, a ^{13}C may be assigned to several symmetrical carbon atoms. Similar to the $^1H-^{13}C$ mapping, the $^{13}C-^{13}C$ mapping is simply achieved by replacing each ^{13}C signal with its corresponding carbon atoms. In this whole spectral interpretation process, a cross peak may be initially incompletely resolved (thus having ambiguously associated 1D signals in F1 and/or F2 dimensions), or be transformed into a $^{13}C-^{13}C$ ANDC when a degenerated ^1H signal corresponds to several ^{13}C signals in HMQC spectrum, or be finally transformed into a C−C ANDC when a ^{13}C signal is mapped to several carbon atoms because of molecular symmetry. When an ANDC is obtained, the user is signaled to supply the range of sDCs that must be satisfied. (The default range of sDCs is 1 to xy, but generally we choose 1 to 2.)

Prior to the structure generation process, the DCs that have definite nodes ($x = y = 1$), one-bond distance ($d_1 = d_2 = 1$), and exactly one sDC that must be satisfied ($s_1 = s_2 = 1$) are extracted as *fixed bonds*. The fixed bonds constitute the partially elucidated structure of the unknown molecule. The remaining DCs, with ambiguous nodes, distances, and/or other characters, are called *ambiguous distance constraints* (ADCs). Some of the ADCs may be already satisfied by the partial structure. During the structure generation process, which is actually a depth-first bond searching process, the presently generated (sub)structure is checked against all the presently unsatisfied ADCs whenever a bond is searched. For each ADC represented as (1), the actual distances between all the sDCs are calculated. If the distance between an sDC can be determined and the distance falls between d_1 and d_2, the sDC is marked as satisfied, otherwise as violated. When all its sDCs have been checked, the ADC is satisfied if the number of satisfied sDCs falls between s_1 and s_2; otherwise it is violated. If too many ADCs are violated, the present substructure is rejected. Depending on the quality of peakpicking, the user can define the minimum number of ADCs that can be violated.

In addition to this simple generate-and-test approach, CISOC-SES has been featured by some heuristics that use LRDCs to prospectively prune the branches of the search

tree and thus achieve high efficiency of structure generation. We have extended and generalized these approaches so that the various ADCs are systematically and uniformly used for this purpose. These renovated approaches are described as follows.

(1) Weighting of FBMX Based on ADCs. First, the possible connections (i.e., elements that equal 1) in the FBMX are weighted based on the unsatisfied ADCs before the structure generation begins. If there exists an ADC represented as (1) on an atom pair, all the possible connections, $c(i, j)$, between the atom pair, are weighted by adding an increment

$$\Delta c(i, j) = W/(d_2 - d_1 + x + y - 1) \qquad (2)$$

where W is a weighting parameter whose value is empirically set as 24 and can be modified by the user. Typically, for an HMBC-derived DC with unambiguous nodes, or a COSY-derived DC with two alternative nodes at one side, the increment is $W/2$. The greater the ambiguity is involved, the smaller the increment will be. In this way, an element $c(i, j)$ of FBMX has the following physical meanings:

1. The possibility of bond formation between the ith and jth free bonds: if $c(i, j) = 0$, no bond can be formed; otherwise a bond might be formed (i.e., a possible connection).

2. The probability of bond formation between the ith and jth free bonds: the bigger $c(i, j)$ is, the more probable that a bond may exist between them.

3. The amount of ADCs that are imposed on this pair of atoms: the bigger $c(i, j)$ is, the more ADCs may be checked (either satisfied or violated) if the local structure of the two atoms are generated.

These are the fundamentals of the subsequent rearrangement of the search tree in order to enhance the search efficiency. But it must be noted that these meanings are incomplete in the sense that 2D NMR correlation information is generally only given between H−H, H−C, or C−C atom pairs and not all such atom pairs that theoretically give signals may be observed in a 2D spectrum. These are specially considered in our subsequent heuristics that are based on the FBMX weighting.

In analogue to the above static weighting of FBMX, the dynamic weighting of FBMX during the structure generation process is also generalized to use all the ADCs: when the first bond is formed between two atoms i and j and, if there is an unsatisfied ADC on atom j and another atom k with $d_1 \le 2 \le d_2$ and $c(j, k) \neq 0$, $c(j, k)$ is then incremented in the same fashion as (2). The rationale was described in the previous paper.[3]

(2) Rearrangement and Prospective Limit of the Search Space Based on ADCs. The structure generation process of CISOC-SES is actually a depth-first search of all the valid combinations of the possible connections in FBMX. The search tree is implicitly constructed from FBMX as the tree is being traversed. Starting from the root, which is a dummy node that represents the initial state of the structure generation, each of the extension levels consists of all the alternative free bonds (denoted as $FBx(i)$) that are to be connected with a free bond (denoted as FBy) specific to this level. Actually, these FBxs, or the extension nodes, correspond to some or all of the possible connections in the row of the FBMX that corresponds to FBy. If the total number of free bonds is M, then the height of the search tree should be $M/2$, and a path from the root to a leaf (a node at the bottom level) represents the $M/2$ possible connections chosen to form a complete and connected structure. This path is called a *structure generation path*, whose maximum length is $M/2$.

The key point of our generator is that the search tree is organized on the basis of the weights of the possible connections so that the levels consisting of the most probable connections with the highest weights are nearest to the top level. Moreover, in each level, the nodes are also sorted in the descending order of their weights so that the most promising connections are searched first. The initial aim of this tree rearrangement was to satisfy the LRDC's as early as possible in the generation path, so that the fruitless paths are detected as early as possible and a greater K_L^M can be used.[3,4] On the other hand, in the extensive application of CISOC-SES we observed that, in most cases when sufficient LRDCs are available, this approach guarantees the target structure to be generated as long as a small portion of all the possibilities have been searched. This suggested to us to limit the search scope to the most promising portion of the reordered search space, i.e., only part of the extension nodes at each level of the search tree are to be visited. So we defined a parameter, N_FBX_STEP, to allow the user to limit the average number of carbon-incident FBxs to be searched at each level. Note that it is only the search of carbon-incident FBxs that are limited, not that of the heteroatom-incident FBxs, because heteroatoms are seldom involved in ADCs.

So far in all the test cases of CISOC-SES (where sufficient unsatisfied DC's are available), N_FBX_STEP $\le$ 3.0 has been successfully used.[12] In effect, with this heuristic, the original time complexity of the structure generation

$$O(n^n) \qquad (3)$$

is thus reduced to

$$O(3.0^n) \qquad (4)$$

where $n = M/2$, i.e., the maximum length of the generation paths.

(3) Prospective Termination of the Structure Generation Paths Based on ADCs. As described in the previous papers, our experience showed that a structure generation path which leads to the correct structure usually has the greatest rate of LRDC-satisfaction (K_L), so structure generation paths with lower K_L than a user-required value, SAT_LRDC_RATE, is prospectively terminated.[3,4] In the present version, the rate of LRDC-satisfaction is generalized to be the "rate of sDC-checking" (denoted as K_A), which is calculated as follows

$$K_A = D_c B_0 / D_0 B \qquad (5)$$

where D_c is the current number of checked (i.e., either satisfied or violated) sDCs, B_0 is the total number of bonds that must be searched to generate a complete structure (i.e., the maximum length of a generation path), D_0 is the total number of sDC's that must be checked, and B is the current number of chosen bonds or the length of the present generation path.

This heuristic, together with other test criteria such as those based on the prediction of ^{13}C chemical shifts and the

violation of ADCs, actually reduces the average length of the structure generation paths. Let that the average length of the paths be reduced to $n - x$, and the time complexity of the structure generation is now reduced to

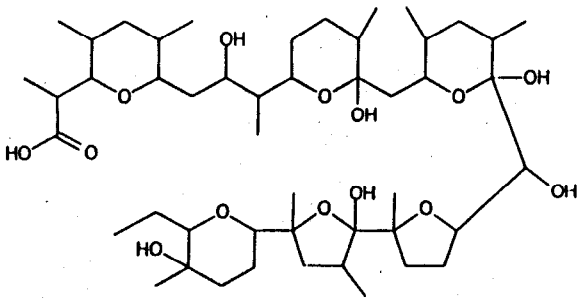
$$O(3.0^{n-x}) \quad (6)$$

As the heuristic based on K_A terminates the structure generation paths prospectively, i.e., before any contradiction with the spectral data is detected, it reduces the value of $n - x$ more significantly than the other tests.

RESULTS AND DISCUSSION

In our experience, even with a 600 MHz NMR spectrometer, the 2D NMR spectra of a quite simple natural product molecule may be complicated by peak overlapping, particularly in the ^1H dimension. The new capability of CISOC-SES to cope with such spectral ambiguity substantially facilitates a chemist's structure elucidation process with CISOC-SES. This is demonstrated by the structure elucidation of alborixin with CISOC-SES.

The structure of alborixin (1), an antibiotic isolated from cultures of a strain of *Steptomyces albus*, was first determined by an X-ray method[13] and was later studied with a 2D NMR spectroscopy.[14] As a test case, the reported ^1H, ^{13}C (Table 1), COSY (Table 2), HMQC (Table 3), HMBC (Table 4), and NOESY (Table 5) spectral data of alborixin were entered



1

to CISOC-SES together with its molecular formula, $C_{48}H_{84}O_{14}$. Note that in order to gather more long-range C−H connectivities, two HMBC spectra with the delays for evolution of long-range coupling selected as 50 and 80 ms, respectively, were used here and 93 cross peaks were picked. Note that these HMBC cross peaks covers only 34.7% of the 239 theoretical HMBC peaks. Besides the big molecular size, the main difficulty encountered in the manual analysis was caused by the five degenerated ^1H resonances (four of which arising from two spins and one from three spins, see Table 1), which cannot be resolved even with the help of HMQC spectrum (see Table 3). Moreover, the absence of COSY cross peaks between two vicinal proton pairs (δ^1_H: 4.450−3.400, 2.213−1.245 ppm, the attached protons of C10, C13, C14, and C17, respectively) also complicated the manual spectral interpretation process, and this was solved with the help of NOESY connectivities.[14]

With CISOC-SES, the total 244 2D cross peaks were interpreted and cross checked as a unified set of 147 ^{13}C−^{13}C DCs. This process was carried out in an interactive mode and was completed in a few minutes. As a result of the ^1H resonance degeneracy, 39 of the ^{13}C−^{13}C DCs were ANDCs. Note that we had to manually link each of the three observed hydroxyl proton resonances (δ^1_H: 6.444, 5.997, and 2.416 ppm, which show no cross peak in the HMQC spectrum) to an oxygen atom so that any involved ^1H−^1H or ^1H−^{13}C connectivities can be properly transformed into AADCs. For all the ANDCs, we specified that the number of sDCs that must be satisfied be 1 or 2. Among the 39 ANDCs, the COSY-derived ones are illustrated as wavy bonds in Figure 2a. Twenty-two unambiguous direct DCs, together with two user-supplied connectivities concerning the carboxylic group (δ^{13}_C = 176.40 ppm) were extracted as fixed bonds (denoted as bold bonds in Figure 2a). The remaining ADCs, except

Table 1. 1D ^1H and ^{13}C Spectral Data of Alborixin (^1H, 600 MHz; ^{13}C, 125 MHz)

peak no.[a]	δ^1H (ppm)	multiplicity[b]	$\delta^{13}C$ (ppm)	multiplicity[b]
1	6.444	s	176.40	s
2	5.997	s	108.31	s
3	4.450	m	99.51	s
4	4.337	m	97.79	s
5	4.119	m	89.33	s
6	3.610	m	86.81	d
7	3.513	m	83.91	d
8	3.400[c]	m	83.73	d
9	3.380	m	82.74	s
10	3.075	m	77.00	d
11	2.725	m	72.42	d
12	2.416	d	71.74	d
13	2.271	m	70.24	d
14	2.213	m	70.00	d
15	2.160	m	69.59	s
16	2.134	m	68.78	d
17	2.039	m	46.38	d
18	1.811	m	42.26	t
19	1.789	m	41.59	t
20	1.723	m	40.44	d
21	1.718[c]	m	39.81	t
22	1.683	m	39.13	d
23	1.618[d]	m	38.80	d
24	1.601	m	38.21	t
25	1.579[c]	m	36.06	d
26	1.540	m	35.93	t
27	1.522	m	34.67	d
28	1.507	m	33.31	d
29	1.500	m	32.77	t
30	1.476	m	31.78	d
31	1.440	m	31.23	t
32	1.410	m	31.18	t
33	1.386	m	29.01	t
34	1.373	m	27.21	t
35	1.367[c]	m	25.45	q
36	1.354	m	24.90	q
37	1.319	s	24.23	q
38	1.245	m	21.47	t
39	1.166	s	19.94	t
40	1.072	d	18.28	q
41	1.047	m	17.66	q
42	1.036	s	16.85	q
43	1.018	d	16.72	q
44	0.950	t	16.28	q
45	0.932	m	14.72	q
46	0.878[c]	d	10.56	q
47	0.824	d	8.97	q
48	0.809	d	8.57	q
49	0.768	d		
50	0.751	d		

[a] The numbering of the peaks are used as their indices in Tables 2−6. [b] Only the simplest splitting patterns, i.e., singlet (s), doublet (d), triplet (t), and quartet (q), are descriminated; others are all specified as multiplets (m) or unknown (u). [c] Degenerated resonances of two spins. [d] Degenerated resonances of three spins.

Table 2. $^1H-^1H$ COSY Correlation Data of Alborixin[a]

(11−7)	(12−8)	(16−3)	(16−14)	(17−2)[b]
(17−13)	(18−5)	(19−6)	(20−9)	(21−3)
(21−14)	(21−16)	(22−18)	(23−8)	(23−19)
(24−13)	(24−17)	(25−7)	(26−4)	(26−10)
(29−14)	(29−16)	(31−23)	(32−15)	(33−5)
(33−30)	(34−10)	(34−21)	(35−9)	(35−20)
(36−1)[b]	(36−23)	(38−6)	(40−11)	(41−6)
(41−19)	(41−23)	(41−31)	(43−17)	(44−20)
(44−35)	(45−25)	(46−15)	(46−36)	(47−33)
(48−25)	(49−38)	(50−34)		

[a] Each pair of numbers enclosed in parentheses are the indices of the associated two ^1H peaks of a cross peak. Only the cross peaks above the diagonal are listed. For the corresponding ^1H chemical shifts, see Table 1. [b] Weak cross peaks, which are interpreted as $^1H-^1H$ distance constraint of four to five bonds. The other cross peaks, either strong or medium, are interpreted as $^1H-^1H$ distance of two or three bonds.

Table 3. HMQC ($^{13}C-^1H$) Correlation Data of Alborixin[a]

(6−10)	(7−7)	(8−9)	(10−3)	(11−5)
(12−8)	(13−8)	(14−4)	(16−6)	(17−38)
(18−13)	(18−24)	(19−21)	(19−45)	(20−11)
(21−18)	(21−22)	(22−36)	(23−17)	(24−26)
(24−28)	(25−34)	(26−30)	(26−32)	(27−33)
(28−15)	(29−14)	(29−29)	(30−25)	(31−19)
(31−41)	(32−23)	(32−25)	(33−16)	(33−21)
(34−23)	(34−31)	(35−37)	(36−42)	(37−39)
(38−23)	(38−27)	(39−20)	(39−35)	(40−47)
(41−50)	(42−48)	(43−46)	(44−46)	(45−43)
(46−44)	(47−49)	(48−40)	(56−1)[b]	(57−2)[b]
(58−12)[b]				

[a] Except the last three ones, each pair of numbers enclosed in parentheses are the indices of the associated HD ^{13}C and ^1H resonances of a cross peak. Each cross peak is interpreted as a direct $^{13}C-^1H$ connectivity regardless of its intensity. For the corresponding ^{13}C and ^1H chemical shifts, see Table 1. [b] Fabricated $O-^1H$ connectivities for hydroxyl proton resonances, the two numbers being the indices of the associated oxygen atom and ^1H resonance, respectively.

Table 4. HMBC ($^{13}C-^1H$) Correlation Data of Alborixin[a]

(1−11)	(1−40)	(2−2)	(2−13)	(2−14)
(2−17)	(2−39)	(2−43)	(3−15)	(3−46)
(4−1)	(4−22)	(4−46)	(5−14)	(5−21)
(5−39)	(6−4)	(6−26)	(6−28)	(6−34)
(6−50)	(7−21)	(7−40)	(7−45)	(7−48)
(8−20)	(8−42)	(8−44)	(9−23)	(9−37)
(10−14)	(10−16)	(10−29)	(11−18)	(11−47)
(12−9)	(12−13)	(12−23)	(12−24)	(12−25)
(12−27)	(12−37)	(13−3)	(13−15)	(13−16)
(14−49)	(15−25)	(15−27)	(15−42)	(16−4)
(16−49)	(17−49)	(18−17)	(18−37)	(18−43)
(19−7)	(19−25)	(19−34)	(19−48)	(19−50)
(20−40)	(21−1)	(22−1)	(22−23)	(22−41)
(22−46)	(23−2)	(23−43)	(25−45)	(25−50)
(26−46)	(26−47)	(27−30)	(27−32)	(27−47)
(28−46)	(29−39)	(30−48)	(32−42)	(33−14)
(33−29)	(34−46)	(35−24)	(37−14)	(37−29)
(38−23)	(39−44)	(44−15)	(45−24)	(46−20)
(47−4)	(48−7)	(48−11)		

[a] Two HMBC spectra with the delays for evolution of long-range couplings selected as 50 and 80 ms, respectively, were used. Each pair of numbers enclosed in parentheses are, respectively, the associated HD ^{13}C and ^1H resonances of a cross peak. Each cross peak is interpreted as a two- or three-bond $^{13}C-^1H$ connectivity regardless of its intensity (but the redundant peaks as in the HMQC spectrum are discarded later). For the corresponding ^{13}C and ^1H chemical shifts, see Table 1.

Table 5. NOESY ($^1H-^1H$) Correlation Data of Alborixin[a]

(2−1)	(5−1)	(5−4)	(6−1)	(8−3)
(10−4)	(10−7)	(11−7)	(18−1)	(19−6)
(20−8)	(21−16)	(22−18)	(23−2)	(24−8)
(24−13)	(28−10)	(29−14)	(31−1)	(31−23)
(33−5)	(35−20)	(37−6)	(37−8)	(37−24)
(38−4)	(39−2)	(40−11)	(41−19)	(42−9)
(43−39)	(44−8)	(44−42)	(45−21)	(46−1)
(46−2)	(46−8)	(46−36)	(47−5)	(47−22)
(48−7)	(48−11)	(48−25)	(49−6)	(49−26)
(49−38)	(50−10)			

[a] Mixing time: 400 ms. Each pair of numbers enclosed in parentheses are the indices of the associated HD ^1H resonances of a NOESY cross peak. Each cross peak is interpreted as a two- to six-bond $^1H-^1H$ connectivity regardless of its intensity. For the corresponding ^1H chemical shifts, see Table 1.

the NOESY-derived ones, were checked against the fixed bonds and finally 65 unsatisfied ADCs, of which 23 were ANDCs, remained to be used as constraints on the subsequent structure generation (Table 6). The NOESY-derived DCs were not used in the subsequent structure generation process.[15]

While constructing the FBMX, we chose to use COSY and NOESY connectivities in combination to determine the disconnectivities between the proton-bearing carbon atoms. This means that two proton-bearing carbon atoms are forbidden to be connected if they manifest neither COSY nor NOESY connectivities. This successfully avoided the potential problem that might arise from the absence of two COSY peaks between two vicinal proton pairs. The total number of free bonds were 88, thus an FBMX of 88 × 88 was set up. The FBMX was further reduced on the basis of all the DCs as well as the ^1H multiplicities, leaving 5387, or 70.27%, of the off-diagonal elements in the FBMX being nonzero. Finally the nonzero elements in the FBMX were weighted on the basis of all the unsatisfied ADCs.

According to the time complexity of (3), a huge number of combinations are inevitable in the structure generation based on this FBMX ($n = 44$). With CISOC-SES, however, the structure generation took a CPU time of only ca. 1752 s on a DEC AXP 4610 computer with $K_\Lambda = 1.1$ and N_FBX_STEP = 2.0, when the evident carboxylic group. ($\delta^{13}C = 176.40$ ppm) was entered as two user-known C−O connectivities and the generation of structures with component ring size greater than 7 was prohibited. The correct structure **1** was given out together with 89 other candidates. CISOC-SES was so designed that, principally, no user input save the molecular formula and NMR spectral data is needed, and it will give all the structures consistent with these data. However, additional structural information of the unknown, which usually is available from user's experience and background of the unknown, is also expected, because it will significantly improve the structure generation efficiency. CISOC-SES provides some flexible approaches for the user to enter their knowledge about the unknown. In addition to atom−atom connectivities, the user can also limit the neighboring pattern of some carbons to complement the present superficial use of ^{13}C chemical shifts in CISOC-SES. In the case of alborixin, according to the knowledge on HD chemical shift-structural feature relationship, we forced the three carbon atoms with $\delta^{13}C$ falling between 108.31 to 97.79 ppm to have exactly two neighboring oxygen atoms and 12 carbon atoms with $\delta^{13}C$ falling between 89.33 to 69.59 ppm to have exactly one neighboring oxygen. This user-known information, input as user-intervention, was used by the

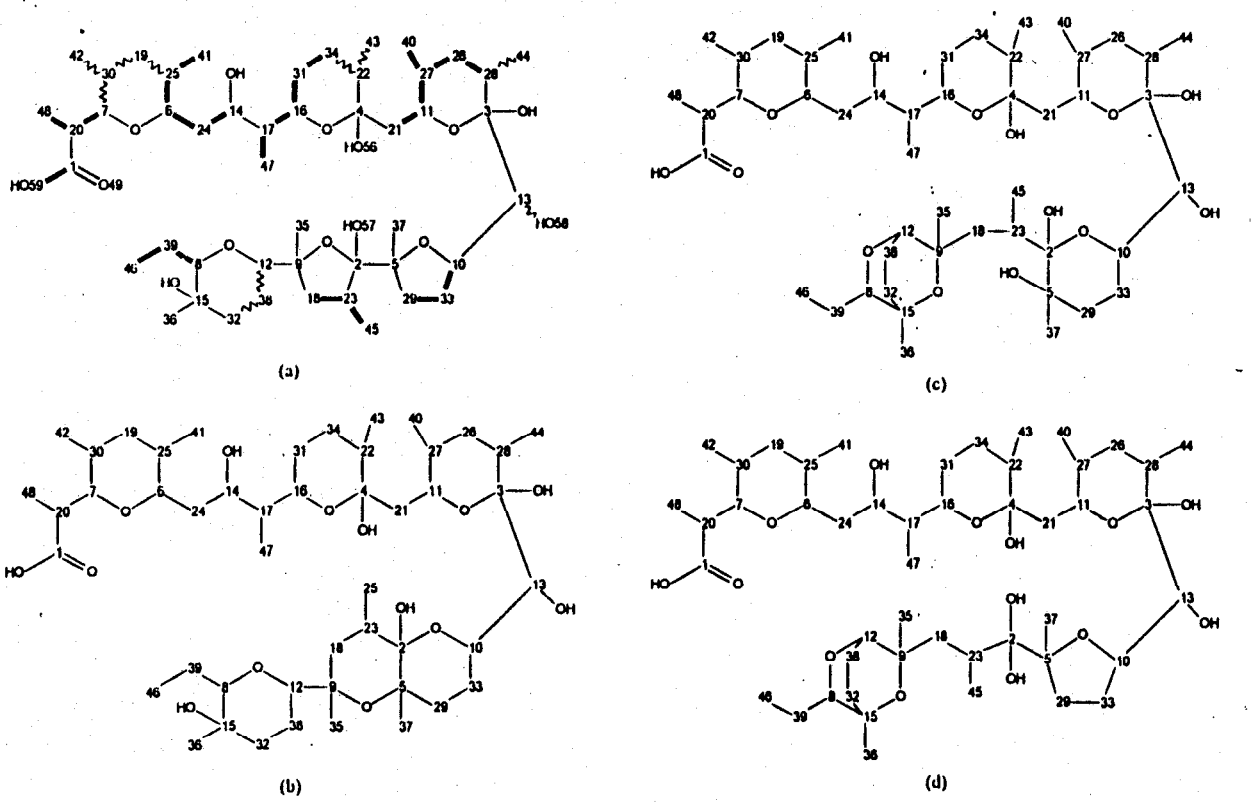


Figure 2. The four candidate structures given by CISOC-SES for alborixin. Numbering of the carbon atoms corresponds to their assigned ^{13}C peaks (see Table 1): (a) the correct structure of alborixin. The bold bonds denote the unambiguous COSY-derived and user-supplied connectivities, whereas the wavy bonds denote the COSY-derived ambiguous DCs resulting from ^{1}H resonance degeneracy.

system as environment constraints which together with the allowed component ring size being set at 5 or 6 led to an enhancement of the efficiency. With these two kinds of constraints and the sDC range of each ADC being specified as 1−1, the CPU time for structure generation was ca. 105 s and only four result structures (Figure 2) were obtained, but one ADC must be allowed to be violated otherwise no candidate structure would be obtained. (Two sDCs were satisfied for an ADC, thus violating the upper limit of satisfied sDCs.)

For this specific (or other large) molecule, the user-supplied upper limit of the component ring size is crucial, without it the number of generated candidate structures (most of which have at least one component ring size greater than 10) soon reaches 1000, the actually allowed upmost number of candidate structures to be generated, and the generation process is thus automatically aborted before the correct structure is generated. The rationale of restricting the component ring size, in general sense, is based on CAS' statistics[16] that the 20 most frequently occurring ring graphs, among which no one consists of component rings whose size is bigger than 7, account for over 90% of the total ring occurrences. In the practical use of CISOC-SES, the generation of structures with component ring size greater than 7 is prohibited at the first structure elucidation process. (Other restrictive parameters and user-supplied structural information may also be used.) Based on the primary result obtained and CPU time needed, the structure generation process can be repeated with relieved (or tightened, as in the case of alborixin) restrictions and more general (or specific) results will be obtained. This is done in order to avoid unsupportable CPU time, insignificant and unmanageable results, and the loss of correct structure.

These results demonstrate the high performance of CISOC-SES in the management of a large sum of (possibly ambiguous) spectral information and in the transformation of the spectral information into structural results for complex natural products. But it is important to consider the difference between this process and an *ab initio* use of CISOC-SES. The peakpicking of some of the cross peaks in Tables 2−5 were quite marginal, some could only be confirmed *after* the structure was determined. As described above, CISOC-SES *can* cope well with such ambiguities; but the increased ambiguity would inevitably lead to (possibly much) longer CPU time for structure generation. Also, it takes time before arriving at the optimal parameters for structure generation as listed above. On the other hand, it can be argued that, in real-world structure elucidation, the chemist can use various methods (e.g., use of different solvents, chemical shift reagent) to resolve the spectral ambiguities to some extent and generally more background information than that was used here is available. So we believe that these results are close to those of a real-life application of CISOC-SES.

Finally, for each candidate structure, CISOC-SES offers a facility that lists the actual topological distance between each NOESY-active proton pair. This helps the chemist recognize spatially significant NOE connectivities for the purpose of stereochemical and conformation study.

Table 6. Unsatisfied Ambiguous Atom−Atom Distance Constraints for Alborixin[a]

(58−13 12: 1−1; 0; 1−2)CQ	(23−57: 2−2; 0; 1−1)CQB
(33 19−10: 1−1; 0; 1−2)CQ	(33 19−29: 1−1; 0; 1−2)CQ
(38 34 32−13 12: 1−1; 0; 1−2)CQ	(32 30−7: 1−1; 0; 1−2)CQ
(34−38 34 32: 0−1; 0; 1−2)CQ	(25−33 19: 1−1; 0; 1−2)CQ
(22−56: 2−2; 0; 1−1)CQB	(22−38 34 32: 1−1; 0; 1−2)CQB
(19−32 30: 1−1; 0; 1−2)CQB	(44 43−28: 1−1; 0; 1−2)CQB
(44 43−22: 1−1; 0; 1−2)CQB	(42−32 30: 1−1; 0; 1−2)CQ
(1−20: 1−2; 0; 1−1)BQ	(1−48: 1−2; 0; 1−1)BQ
(2−57: 1−2; 0; 1−1)BQ	(2−18: 1−2; 0; 1−1)BQ
(2−29: 1−2; 0; 1−1)BQ	(2−23: 1−2; 0; 1−1)BQ
(2−37: 1−2; 0; 1−1)BQ	(2−45: 1−2; 0; 1−1)BQ
(3−28: 1−2; 0; 1−1)BQ	(3−44 43: 1−2; 0; 1−2)BQ
(4−56: 1−2; 0; 1−1)BQ	(4−21: 1−2; 0; 1−1)BQ
(4−44 43: 1−2; 0; 1−2)BQ	(5−29: 1−2; 0; 1−1)BQ
(5−33 19: 1−2; 0; 1−2)BQ	(5−37: 1−2; 0; 1−1)BQ
(6−14: 1−2; 0; 1−1)BQ	(7−33 19: 1−2; 0; 1−2)BQ
(7−19: 1−2; 0; 1−1)BQ	(7−42: 1−2; 0; 1−1)BQ
(8−36: 1−2; 0; 1−1)BQ	(9−38 34 32: 1−2; 0; 1−2)BQ
(9−35: 1−2; 0; 1−1)BQ	(10−29: 1−2; 0; 1−1)BQ
(12−8: 1−2; 0; 1−1)BQ	(12−18: 1−2; 0; 1−1)BQ
(12−38 34 32: 1−2; 0; 1−2)BQ	(12−32 30: 1−2; 0; 1−2)BQ
(12−38: 1−2; 0; 1−1)BQ	(12−35: 1−2; 0; 1−1)BQ
(13−10: 1−2; 0; 1−1)BQ	(13−28: 1−2; 0; 1−1)BQ
(13−33: 1−2; 0; 1−1)BQ	(14−47: 1−2; 0; 1−1)BQ
(15−32 30: 1−2; 0; 1−2)BQ	(15−38: 1−2; 0; 1−1)BQ
(15−36: 1−2; 0; 1−1)BQ	(16−14: 1−2; 0; 1−1)BQ
(18−35: 1−2; 0; 1−1)BQ	(19−25: 1−2; 0; 1−1)BQ
(19−42: 1−2; 0; 1−1)BQ	(19−41: 1−2; 0; 1−1)BQ
(21−56: 1−2; 0; 1−1)BQ	(22−31: 1−2; 0; 1−1)BQ
(26−44 43: 1−2; 0; 1−2)BQ	(29−37: 1−2; 0; 1−1)BQ
(30−42: 1−2; 0; 1−1)BQ	(32−36: 1−2; 0; 1−1)BQ
(34−44 43: 1−2; 0; 1−2)BQ	(38−38 34 32: 1−2; 0; 1−2)BQ
(44−28: 1−2; 0; 1−1)BQ	

[a] See the text for explanation of representation of a distance constraint. The numbering of the atoms corresponds to that in Figure 2a.

CONCLUSION

Efficient use of 2D NMR correlation information in computer-assisted structure elucidation has forwarded the application realm of such systems from simple molecules to real-world complex organic compounds and natural products. Yet, for complex molecules, the ubiquity of spectral imperfections (such as peak degeneracy, low-digital resolution, artifacts, and noises) as well as the existence of molecular symmetry force us to consider the use of their resulting ambiguous correlation information. As an improvement of CISOC-SES in this aspect, we presented in this article a general scheme to represent such ambiguous correlation information in the form of ambiguous-node-and/or-distance-involved distance constraints and some novel approaches to the efficient use of them to achieve high-efficient structure generation. What is worth mentioning are the heuristics that prospectively limit the search scope of structure generation and terminate the structure generation paths, which significantly cut down the CPU time for structure generation without ignoring the correct structure. In our view, without such powerful heuristics, it is unimaginable to complete the structure generation process of alborixin in a predictable length of CPU time. The other new facilities, which enable it to tolerate the missing of COSY peaks and erroneous picking of HMBC peaks, are also essential to guarantee the generation of the correct structure. In conclusion, we believe that this improvement of CISOC-SES is a giant step toward our final goal of practical computer-assisted structure elucidation for complex natural products.

CISOC-SES now still is an in-house program, if you are interested in it, a copy of the executable file on VAX/VMS platform is available upon request from author SGY.

REFERENCES AND NOTES

(1) *Artificial Intelligence Applications in Chemistry, ACS Symposium Series 306*; American Chemistry Society: Pierce, T. H., Hohne, B. A., Ed. Washington DC, 1986; Preface.

(2) Christie, B. D.; Munk, M. E. *J. Am. Chem. Soc.* **1991**, *113*, 3750−3757.

(3) Peng, C.; Yuan, S.-G.; Zheng, C.-Z.; Hui, Y.-Z. Efficient Application of 2D NMR Correlation Information in Computer-Assisted Structure Elucidation of Complex Natural Products. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 805−813.

(4) Peng, C.; Yuan, S.-G.; Zheng, C.-Z.; Hui, Y.-Z.; Wu, H.-M.; Ma, K; Han, X.-W. Application of Expert System CISOC-SES in Structure Elucidation of Complex Natural Products. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 814−819.

(5) Duddeck, H.; Dietrich, W. *Structure Elucidation by Modern NMR, A Workbook*; Springer-Verlag: New York, 1989.

(6) Bangov, I. P.; Simova, S. Computer-Assisted Structure Generation from a Gross Formula. 6. Reducing the Structural Redundancy by the Employment of 2D NMR Spectral Information. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 546−557.

(7) Throughout this paper, the term "distance" refers to the number of intervening bonds between two correlated atoms (or signals), so it is of topological nature.

(8) 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.

(9) Bohanec, S.; Zupan, J. Structure Generation of Constitutional Isomers from Structural Fragments. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 531−540.

(10) Munk, M. E.; Farkas, M.; Lipkis, A. H.; Christie, B. D. Computer-Assisted Chemical Structure Analysis. *Mikrochim. Acta* **1986 II**, 199−215.

(11) The aromatic bonds are represented as alternating single and double bonds.

(12) *Test Report of CISOC-SES*; Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry, Shanghai, 1994.

(13) Gachon, P.; Fauges, C.; Kergomard, A. Alborixin, a New Antibiotic Ionophore: Isolation, Structure, Physical and Chemical Properties. *J. Antibiotics* **1976**, *29*, 603−610.

(14) Tao, X.-L. M. Sc. Dissertation, Shanghai Institute of Organic Chemistry, 1989.

(15) An NOESY cross peak is interpreted as a $^1H-^1H$ distance of two to six bonds and the user can choose whether or not to use the NOESY-derived DCs in the way as other DCs in the structure generation process. In our experience, use of NOESY-derived DCs does not increase the efficiency of structure generation as some DCs must be allowed to be violated.

(16) Stobaugh, R. E. Chemical Abstracts Service Chemical Repistry System. 11. Substructure-Related Statistics: Update and Additions. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 180−187.

CI940135M