Efficient Application of 2D NMR Correlation Information in Computer-Assisted Structure Elucidation of Complex Natural Products

Chen Peng, Shengang Yuan,* Chongzhi Zheng, and Yongzheng Hui

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

Received October 13, 1993*

Efficiency is one of the essential problems concerning computer-assisted structure elucidation (CASE) systems if real-world problems are to be solved by them. On account of this, in our newly developed CASE system, CISOC-SES (Computerized Information System for Organic Chemistry-Structure Elucidation System), 2D NMR correlation information, instead of chemical shifts, is fully exploited through some novel approaches such as reduction of the search space based on one-bond connectivities (typically derived from COSY or INADEQUATE spectra), weighting and rearrangement of the search tree based on long-range connectivities (typically derived from HMBC or COLOC spectra), and prospective evaluation of intermediate structures based on the rate of satisfaction of long-range connectivities. The rationale and implementation of these approaches are described in this paper, while the practical application of the system to the structure elucidation of complex natural products, which gives satisfactory results, is discussed in the following paper in this issue.

INTRODUCTION

Two-dimensional NMR technique are now being routinely used in laboratories as powerful tools for the determination of the solution structure of organic molecules.[1] The advantages of using 2D NMR techniques lie in the fact that 2D correlation spectra, such as COSY and 2D INADEQUATE, provide unequivocal "hard" proof of through-bond atom connectivities. In contrast, traditional spectral features, such as NMR chemical shifts and IR absorptions, provide "soft" hints of local structural features, which can only be evalulated through experience and analogous comparisons.[2] But, even with these new techniques, structure elucidation of complex natural products is still far from trivial work for at least the following reasons.

1. As it was addressed by Rycroft,[3] subjective prejudice usually misleads one's inference in structure elucidation as one tends to consider only those structural types that he knows well. As a result, the correct result may be overlooked and sometimes a structure has to be corrected several times before the real structure is elucidated. To overcome this, some ideas to make the process of structure elucidation more free of prejudice by full use of 2D NMR spectral data, with the least referring to reference structures, have been proposed.[2,3]

2. In practice, 2D INADEQUATE spectrum, which provides direct C–C connectivities, is generally unavailable because of its low sensitivity, especially when only a small amount of sample is isolated. Thus C–C connectivities are more frequently indirectly derived from COSY/HMQC spectral combination. But, this method is usually limited by the existence of quaternary carbon atoms and heteroatoms which makes the vicinal H–H scalar couplings very sparse. Usually, HMBC (or COLOC) provides a large amount of long-range H–C connectivities, which can be transformed into one- or two-bond C–C connectivities when combined with HMQC information. Because of their inherent ambiguities, it is a rather tedious work to construct the molecular skeleton principally based on such spectral information when little background information about the molecule is known.

In both respects, computer assistance is feasible because of a computer's intrinsic capability to enumerate possibilities and to manage large sums of information. Computer-assisted structure elucidation (CASE) of organic compounds has been one of the earliest chemical applications of artificial intelligence (AI) techniques since the early 1960s.[4,5] With the initial aim being to enhance a chemist's speed and thoroughness in solving structural problems, such a system is essentially a structure generator that produces exhaustively and irredundantly all the compatible structures on the basis of some structural fragments, either supplied by the user or deduced from MS, IR, and/or NMR spectral data.[6] Among the various kinds of spectral data, ^{13}C NMR data have been most intensively used before the advent of 2D NMR spectroscopy because the structural information contained in a ^{13}C spectrum is related directly to the molecular carbon skeleton. As a general rule of AI applications, the efficiency of such systems arises as a bottleneck when real-world problems are to be tackled, most of the chemical-shift-based CASE systems, which either use a chemical shift–substructure correlation database (e.g., refs 7, 8) or a more concise chemical shift–substructure correlation model (e.g., ref 9), relies to a large extent on user intervention, without which combinatorial explosion in the structure generation is inevitable save for the case of a very small molecule. In fact, few structures with practical complexity have been reported as results of such a chemical-shift-based CASE system without intensive user-supplied structural information.

The impact of using 2D NMR correlation information on improving the efficiency of CASE had been recognized early.[10] In the example of ref 10, the original over 8000 candidate structures deduced for *iresin* from its ^{13}C, ^{1}H NMR shifts, multiplicities, and ^{1}H *J*-couplings were reduced to the unique correct structure after 2D INADEQUATE information was employed. Later, INADEQUATE spectral data were also introduced into other CASE systems, and results of dramatically decreased number of candidate structures were reported.[11,12] While dealing with the C–C correlation information supplied by 2D spectra, the traditional methods, which were originally sophisticatedly designed for the purpose of interpretation of chemical shifts or other spectral features at multiatom substructure levels, became awkward. In CHEMICS,[12] for example, because of the lack of explicit

* Abstract published in *Advance ACS Abstracts,* May 15, 1994.

0095-2338/94/1634-0805$04.50/0

correspondences between the carbon atoms and the ^{13}C signals, the ^{13}C–^{13}C signal connectivities given by an INADEQUATE spectrum cannot be directly used as bonds in the target structure. Instead, they are used to enhance the correspondences that are summarized in an "NM matrix". As this makes the correspondences less ambiguous, stronger constraints are imposed on the selection of component sets (combinations of substructures), and consequently, fewer candidate structures are generated. In SESAMI,[13] on the other hand, the direct C–C atom connectivities are taken as substructural constraints used in structure generation. Apparently, this has also taken a roundabout course, for, even in manual analysis, a C–C connectivity is implicitly taken as a determined C–C bond without further consideration.

While the one-bond C–C connectivities can be applied to the traditional CASE systems without much modification, efficient use of the ambiguous C–H long-range connectivities derived from HMBC or HETCOR spectra proves to be the biggest challenge to CASE. Such a C–H connectivity actually imposes a long-range (topological) distance constraint (LRDC) on the concerned atom pair. As their number of intervening bonds is ambiguous (two or three bonds), LRDCs could not be used directly in the traditional CASE methods. In several recent improvements,[13,14] LRDCs are used as an additional test criterion on the intermediate structures produced during the structure generation, and successful structure elucidation of complex natural products using such LRDCs were reported.[13] As it will be shown in the subsequent sections and the following paper in this issue, the information-abundant LRDCs can indeed be more efficiently used to prospectively guide the structure generation path and thus substantially improve the efficiency of CASE. LSD[15] is another CASE system that mainly uses 2D NMR spectra data (including LRDCs); but it relies too much upon user intervention, such as user-supplied substructures and atom status.

Other types of programs have also been reported which aim at (semi)automatically extracting the correlation information (i.e., peak picking) from 2D spectral data files.[16,17] Yet, most of such programs are designed for the complicated spectra of biopolymers, which are generally more difficult to discern than those of natural products.[18] The results of these programs can serve as the input of a conventional CASE system.

We present here a newly developed CASE system, CISOC-SES, which deduces the constitutional structures of complex natural products principally from conventional 2D NMR spectral data. It can be seen that, with our novel methods such as prospective reduction of the search space based on 2D NMR correlation information and full use of LRDCs to guide and constrain the structure generation, the problems addressed above can, to a large extent, be solved and thus high efficiency in solving real-world structural problems can be realized.

METHODS

The logical structure of CISOC-SES, which simulates the routines used by NMR spectroscopists to some extent, is illustrated in Figure 1. The elucidation process, from the molecular formula (MF) of the unknown to one or, more generally, a list of candidate structures, consists of three successive phases. The first two aim at establishing and reducing a search space inside which all structures compatible with the original data are generated in the last phase. If no constraint information is used, the system can give an exhaustive and irredundant list of isomers solely from the MF. Practically, however, use of conventional 1D and 2D

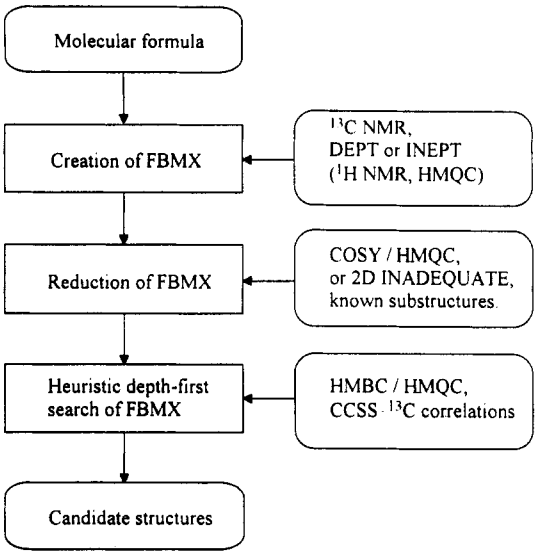

Figure 1. Logical structure of CISOC-SES. FBMX: free-bond connection matrix. CCSS: carbon-centered single-spherical substructure.

NMR spectral data are imperative to improving the efficiency, namely, to diminish both the number of candidate structures and the CPU time needed.

Phase 1. Creation of FBMX(s). In his book (ref 6, pp 192–203), Gray demonstrated that a connectivity matrix of atomic fragments (heavy atoms with definite hybridization and certain numbers of attached protons) was very suited to represent the reduced search space of CASE using INADEQUATE C–C connectivities. Similar fragment matrices were also used in the structure reduction approach proposed by Christie and Munk[8] and in the structure generation approach proposed by Bohanec and Zupan,[19] though both approaches were originally designed for the manipulation of chemical-shift-derived substructures. Sharing some similarities with these approaches, we use a *free-bond connection matrix* (FBMX) to represent the bonding possibilties between all the constituent heavy atoms (non-hydrogen atoms) of the unknown. A heavy atom with a certain number of attached protons and at least one unsaturated valence (called *free bond*) is referred to as an *element group* (EG), which serves as the basic structural unit for the subsequent structre assembly. Note that the "bond-types" of the free bonds are not distinguished in our method; namely, an EG with two free bonds can form two single bonds with other two EGs or form one double bond with another EG. This is more realistic because the hybridization states of the carbon atoms cannot be unambiguously determined from NMR spectral data. If an EG has more than one free bond, the free bonds are equivalent to each other and are called *equivalent free bonds*.

If only the MF is known, all the possible ways of allocation of the constituent protons to the heavy atoms can be enumerated by use of a simple combinatorial algorithm based on the valences of the atoms. Each proton allocation leads to an EG set, and a matrix

$$\mathrm{FBMX} = [c(i,j)], \quad 1 \le i,j \le M \tag{1}$$

is created based on the total number of free bonds, M, with each element $c(i,j)$ in FBMX representing the possibility of bond formation between the ith and the jth free bond: if $c(i,j) = 0$, the two free bonds are forbidden to connect. Otherwise, they are connectable; namely, there is a *possible connection* between the incident EGs, and a bond (of a type not yet determined) may be formed between them in the structure

generation phase. Initially, the diagonal elements and elements between equivalent free bonds are set to 0, and if two EGs have both a single free bond, the element between the two free bonds are also set to 0 because otherwise it would lead to a disconnected structure.

In practice, the total number and multiplicites of ^{13}C peaks derived from ^{13}C and DEPT (or INEPT) spectra can effectively limit the number of possible proton allocations and thus the number of FBMXs. As it is done by Christie and Munk,[13] we assumed that all ^{13}C resonances are resolved unless molecular symmetry exists in the unknown structure, which results in fewer ^{13}C resonances than the number of constiutent carbon atoms. This is realistic since a high-resolution ^{13}C NMR spectrum normally gives well-resolved lines. If no symmetry exists in the structure, i.e., the number of ^{13}C resonances equals that of carbon atoms, a unique correspondence between the carbon atoms and ^{13}C signals is built up and the number of attached protons of each carbon atom can thus be determined by the multiplicity of the corresponding ^{13}C signal. The remaining problem is to allocate residual protons, if any, to the heteroatoms. If symmetry does exist in the unknown structure, the problem is complicated because no unique C–^{13}C correspondence can be obtained. Again, a combinatorial algorithm is used to enumerate the permutation of allocations of carbon atoms to the ^{13}C signals. To alleviate such permutation, ^{1}H peaks are assigned to ^{13}C peaks by the one-bond ^{1}H–^{13}C correlations from an HMQC (or HETCOR) spectrum and the integrals of the ^{1}H peaks are then used as a criterion to eliminate some inappropriate C–^{13}C allocations.[20] On the basis of each C–^{13}C allocation, every carbon atom can again be assigned a certain number of attached protons, and, if necessary for heteroatoms, proton allocations can be enumerated and FBMXs subsequently constructed. The procedure is illustrated in Figure 2.

Before the ^{13}C NMR spectrum is used, the identical EGs are viewed as equivalent and are assigned to the same *symmetry class*. After the carbon atoms are allocated in ^{13}C signals, the symmetry classes are further partitioned according to the associated ^{13}C signals of the EGs. This actually creates an explicit correspondence between the carbon atoms and the ^{13}C signals, which serves as the fundamental basis for the subsequent efficient applications of 2D NMR correlation information.

It should be pointed out that the FBMX is never actually produced in the program; instead, a logically equivalent *atom* (or more strictly, *element group*) *connection matrix* (ACMX) is used because the latter one is commonly smaller in size. An ACMX is defined as

$$\mathrm{ACMX} = [C(i,j)], \quad 1 \le i,j \le N \qquad (2)$$

where N is the number of EGs and each element $C(i,j)$ represents the bonding possibility between the ith and the jth EGs. As the equivalent free bonds are not distinguished in our method, a connection between a pair of EGs is equivalent to all the connections between the free bonds on both the EGs, or a submatrix of FBMX, namely,

$$C(i,j) \equiv [c(k,l)], \quad i_0 \le k < i_0 + n, \quad j_0 \le l < j_0 + m \qquad (3)$$

where i_0 and j_0 are respectively the lowest subscripts of the free bonds of the ith and the jth EGs and, n and m are respectively the numbers of free bonds of the ith and the jth EGs. So, in the following sections, most of the operations are carried out at the level of EGs in ACMX (where C is used to represent connections), rather than at the level of free bonds

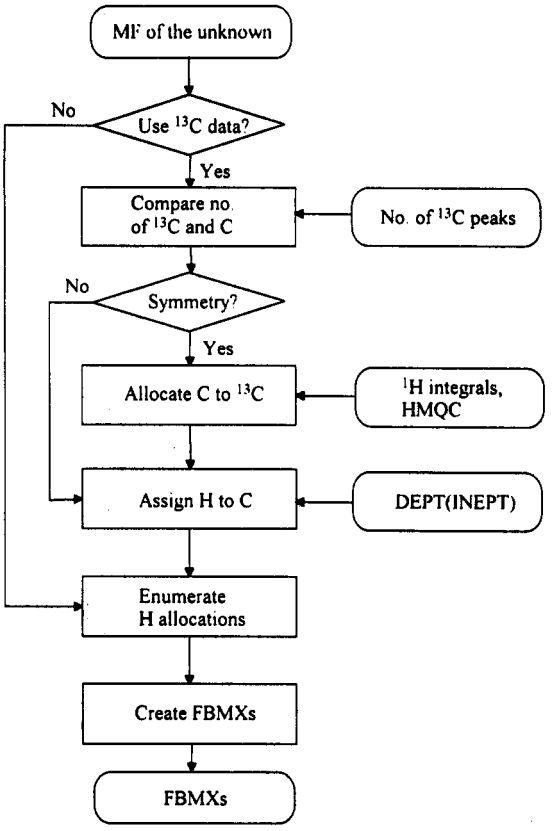

Figure 2. Flow chart of the construction of free-bond connection matrices (FBMXs). H and C stand for hydrogen and carbon atoms, respectively, while ^{1}H and ^{13}C represent ^{1}H and ^{13}C NMR signals, respectively.

in the FBMX (except when pursuing the depth-first search of FBMX, where c is used to represent a connection), though the notation of FBMX is always used.

Phase 2. Reduction of FBMX. In this phase, an FBMX is reduced on the basis of the spectra derived and, if avaialble, user-supplied atom–atom distance constraints through the following steps.

1. Interpretation of 2D Spectra To Obtain ^{13}C–^{13}C Signal Connectivities. In our method, the through-bond connectivities, or distance constraints, are strictly distinguished as two classes, namely, *signal–signal connectivities* (SSCNs) and *atom–atom connectivities* (AACNs). The connectivities derived from the 2D spectra are SSCNs that can be transformed into AACNs based on the atom–signal correspondence created in phase 1. In the system, both SSCNs and AACNs are represented in a uniform format, with a sample connectivity shown as follows:

(2–10: 1∼2 0)

where, from left to right, the first two numbers represent the ID numbers of the correlated signal (or atom) pair, the next two numbers represent the distance, i.e., range of intervening bonds between the two signals (or atoms), and the last number represents the bond type. For a long-range connectivity or a one-bond connectivity derived from most of the 2D spectra, the bond type is meaningless or unknown and is set to 0. For some user-specified one-bond connectivities or, sometimes, those derived from INADEQUATE spectrum where the experiment can be optimized in such a way that only single-bond-connected ^{13}C pairs give rise to cross peaks (e.g., ref 21), the bond type may be explicitly specified by a number 1, 2, or 3 to represent a single, double, or triple bond,

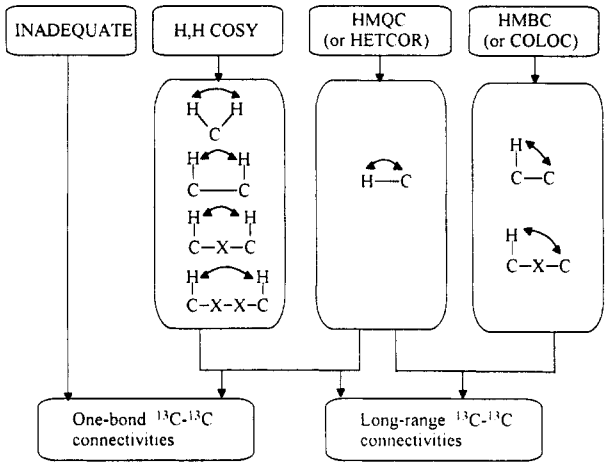

Figure 3. Information flow in the 2D spectra interpretation step. X stands for an arbitrary atom. Note that geminal ^1H–^1H connectivities are automatically discarded when transformed into ^{13}C–^{13}C connectivities.

respectively (aromatic bonds are not specifically considered; they are taken as alternating single and double bonds). Besides, a user-specified long-range connectivity can also be given a nonzero "bond type" to identify its higher priority so that it can override other inconsistent spectra-derived connectivities between the same signal (or atom) pair.

Instead of processing directly the raw spectral data files (produced on an NMR spectrometer after Fourier transformation), the 1D and 2D spectral features are first extracted from these files (either by hand or by use of some semiautomatic peak-picking programs) and represented in the computer as abstract correlation information, which is much more compact in size and more suitable for subsequent automatic analysis. This novel representation of 2D NMR correlation information, which is based on the concept of chromatic graph,[22] will be described elsewhere. Transformation of such correlation information into AACNs is generally trivial work as the rationale has been well-defined;[1,15] the information flow of this step is merely illustrated here (Figure 3). What is worth mentioning is that in our method, the following two additional practical factors are considered in this step: (1) the dependence of the distance and if a one-bond connectivity, the bond type, on the intensity level of the cross peak and (2) the uncertainty of the existence of a near-diagonal cross peak in a homonuclear spectrum. One example is the interpretation of COSY spectra. Generally, a COSY cross peak with a large coupling constant ($J > 5$ Hz) is regarded as proof for the existence of a vicinal or geminal proton pair. Yet, in a favorable case a proton pair separated by four or even five bonds may also give rise to a long-range coupling, which is usually weak. By default, for a cross peak arsing from a coupling with a small (e.g., $J < 3$ Hz), medium (e.g., 3 Hz $\leq J \leq$ 6 Hz), or large J constant (e.g., $J > 6$ Hz), a ^1H–^1H connectivity with distance of 4–5, 3–3, or 3–3 bonds is assigned, respectively. (Here the possible distance of two bonds is not considered for a medium or large J constant because a geminal ^1H–^1H connectivity can be readily recognized and thus omitted when transformed into a connectivity of the same ^{13}C signal.) On the other hand, if two ^1H signals are very close (difference of chemical shfit is below a defined threshold, e.g., 0.02 ppm) and the cross peak between them is absent, it is possible that the cross peak is hidden by diagonal peaks and thus lost while picking peaks. To prevent the error that might arise, a *pseudoconnectivity* with a vague distance of 3–101 bonds (which will be later transformed into a characteristic ^{13}C–^{13}C distance of 1–99 bonds) is assigned to the ^1H pair. This is necessary because, by default, two proton-bearing carbons will be prohibited to be connected in the subsequent structure generation if no COSY peak exists between them. These criteria are represented by some parameters stored in a file, by editing the file, the user can modify them conveniently.

The use of the various kinds of spectra is optional. The only limitation is that if COSY, HMBC, or COLOC spectra are used, the use of HMQC (or HETCOR) is imperative. This is because the protons are not explicitly represented in the system, so the connectivities involving protons should be mapped to their directly connected carbon atoms to be useful.

The obtained ^1H–^1H, ^1H–^{13}C, and ^{13}C–^{13}C SSCNs from various spectra are then cross-checked and transformed into a unified set of ^{13}C–^{13}C SSCNs. Again, some practical factors, such as the existence of proton resonance degeneracy (complete overlap) and inconsistent SSCNs derived from different spectra, are considered.

2. Transformation of ^{13}C–^{13}C Signal Connectivities into C–C Atom Connectivities. The task is now to transform the ^{13}C–^{13}C SSCNs into C–C AACNs that can be subsequently used to reduce the FBMX. If there is a one-to-one correspondence between the ^{13}C signals and the carbon atoms, the transformation is straightforward. On the other hand, if molecular symmetry exists in the unknown structure, a ^{13}C signal may correspond to several carbon atoms; the problem may be very complicated. In the current implementation, if symmetry exists, a ^{13}C–^{13}C SSCN is simply mapped to all the combinations of the corresponding carbon atom pairs. So, when both of the concerned signals correspond to multiple carbon atoms, the user is warned to remove the resulting surplus C–C AACNs. If available, the user-supplied AACNs are also processed in the same manner except that they have higher priorities and are always retained even if inconsistent spectra-derived connectivities exist.

3. Reduction of FBMX. On the basis of the AACNs, both the size of the FBMX and the number of possible connections in the FBMX are reduced in this step. First, the total number of free bonds, M, which determines the size of FBMX, is reduced on the basis of the one-bond AACNs by extracting them as *fixed connections*. For a one-bond AACN with a definite bond type, n (n = 1, 2, or 3 for a single, double, or triple bond, respectively), the n-fixed connections are set in a connection table (CT), the free valences of the concerned atom pair are both reduced by n, and M is thus decreased by $2n$. If the bond type is indefinite, however, one fixed connection is set in the CT and M is reduced by 2. The latter is feasible because, even if it is actually a multiple bond, further formation of some additional connections between this same atom pair in the subsequent generation phase will complement the bond. The CT represents the current state of the structure being generated and is characterized by its unique representation of multiple bonds by multiple single connections. The reduction of the size of FBMX is essential to the efficiency of structure generation because the cost of the search of the FBMX increases exponentially with the size of the FBMX. With the supposition that, in a favorable case, an INADEQUATE spectrum manifests all the C–C bonds of an alkane, M will be then reduced to zero; i.e., the complete structure will be set in the CT, and no further structure generation is necessary. Ordinarily, however, the possible connections in the FBMX should be further reduced through the following steps, which are listed in the descending order of their priorities.

1. If there is a one-bond AACN with a definite bond type between two EGs, the remaining free bonds of them are set mutually unconnectable. On the other hand, if there is a one-bond connectivity with an indefinite bond type between two EGs, the remaining free bonds of them are set mutually connectable.

2. If there is a long-range AACN with a distance excluding the possibility of one bond, the remaining free bonds are set mutually unconnectable.

3. If the unknown structure is symmetric, the carbon atoms in the same symmetry class are set mutually connectable because connectivities between them cannot be detected by spectral methods.

4. As a default option, when COSY spectral data are used, the free bonds of two proton-bearing carbons are set unconnectable if no J-coupling between their protons is detected. Note that there is a risk involved here because in some conformation $^3J_{H-H}$ may approach zero; the user is supposed to modify the FBMX or supply a pseudoconnectivity between the carbon pair under such a case. Of course, a careful user may turn off this option at the price of more CPU time in the subsequent structure generation phase.

5. The free bonds of heteroatoms are set mutually unconnectable by default. Again, this can be changed by the user if he or she thinks that bonds might be formed between heteroatoms.

6. If some specific types of bonds are prohibited between carbon atoms, the connection that will undoubtedly lead to such bonds is eliminated. For example, in the case in which all single C–C bonds have been obtained from an INADEQUATE spectrum, the user may instruct that only multiple bonds be formed between carbon atoms in the structure generation phase. Then, a carbon atom with a single free bond is prohibited to connect to all other carbon atoms because such connections will inevitably lead to surplus C–C single bonds. Again, this rule should be used with care and can be rejected by the user.

In a word, the reduction of the size of FBMX and the number of possible connections in the FBMX reduce in effect the number of levels and the average number of nodes at each level of the subsequent search tree and can significantly narrow the search space for the subsequent structure generation.

Phase 3. Heuristic Depth-First Search of FBMX. In the above two phases the problem of structure generation is transformed into a reduced FBMX, wherein the possible connections between the EGs are represented as entries of value 1, while some fixed connections between the EGs are precluded from the FBMX and are stored in a CT. The total number of free bonds are denoted as M, which is bound to be even; then the task is to choose from the FBMX $M/2$ possible connections in such a way that each column and each row contributes one and only one possible connection and to add them into the CT to form a connected structure. By use of a simple recursive depth-first search method such as that adopted by Bohanec and Zupan[19] and together with some intermediate evaluation procedures, it is possible to generate all the compatible structures. In fact, we have first implemented such a structure generator, which works well when M is small (typically, less than 20) because, for example, the molecular size is small or an INADEQUATE spectrum is used. But when M grows larger, the time of structure generation tends to be unbearably long. To overcome this, we devised some novel heuristic approaches, most of which use the abundant LRDCs derived from an HMBC spectrum, to further enhance the generation efficiency. These approaches are described as follows.

1. Weighting of FBMX. First, the possible connections in the FBMX are weighted on the basis of the LRDCs before the structure generation begins. As most of the LRDCs, which are derived from an HMBC/HMQC spectra combination, have a distance range of one to two bonds, it is implied that the existence of a direct bond between such a carbon pair has a probability of 0.5. Generally, if there exists an LRDC of $1–n$ bonds on an atom pair, the probability of the bond formation between the two atoms is $1/n$. In this way, all the possible connections, $C(i,j)$, between the atom pair are weighted by adding an increment

$$\Delta C(i,j) = W/n \tag{4}$$

where W is a weighting parameter whose value is empirically set as 24 and can be modified by the user. In this way, the FBMX represents not only the possibilities but also the probabilities of bond formation between the EGs.

During the generation process, the FBMX is also dynamcially updated when a connection is chosen (i.e., a bond is formed). The rationale is that when a bond is formed between EGs i and j, the LRDCs concerning j (or i) may give some hints on the neighboring pattern of i (or j). For example, if j has an LRDC of one to three bonds to another EG, k, among all the possible arrangements that the shortest path from j to k covers i, namely, path $j–i–k$ or $j–i–X–k$ (X is an arbitrary EG), the probability that i and k is directly connected (i.e., the first path) is 0.5. Generally, if an unsatisfied LRDC on EGs j and k is *min* to *max* bonds and min $\leq 2 \leq$ max, the possible connections $C(i,k)$ can be incremented as

$$\Delta C(i,k) = W'/(\max - \min + 1) \tag{5}$$

where W' is a weighting parameter that should be smaller than W in (4) because the paths from j to k may have up to four possible directions. For our purpose of qualitative comparison, however, the equal value of W is adopted for W'.

Note that the dynamic weighting occurs only when it is the first connection between i and j that is chosen and this process is iterated for each of the concerned EGs over all the appropriate LRDCs involving the other one. On the other hand, if the last connection between i and j is rejected when backtracking takes place, the above process is reversed to recover $C(i,j)$. In this way, the FBMX remains unchanged after the structure generation is completed as compared with the static weighting results.

The current state of the FBMX is stored in an array, paw[i] ($1 \leq i \leq N$). Before weighting the FBMX, the basic value of paw[i] for each EG, i, is calculated as follows:

if i has no free bonds

$$\text{paw}[i] = 0$$

otherwise

$$\text{paw}[i] = N - \sum_{j=1}^{N} C(i,j) \tag{6}$$

where N is the total number of EGs. Whenever FBMX is weighted, both paw[i] and paw[j] are incremented (or decreased) simultaneously by $\Delta C(i,j)$ if the connection $C(i,j)$ is incremented (or decreased). So, the fewer possible connections an EG has and the more its connections are weighted, the greater the value of *paw* of the EG.

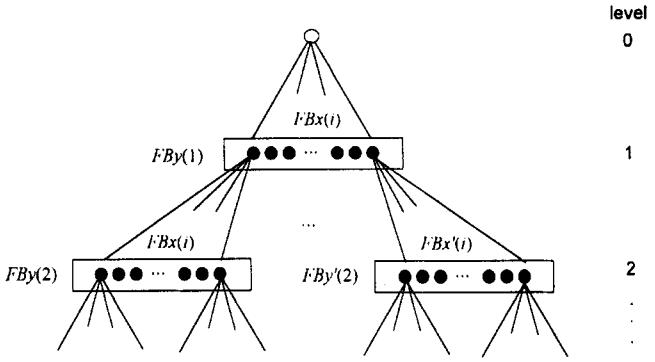

Figure 4. Search tree for the structure generation. For each level, FBx's (represented as the nodes) are the alternative free bonds to connect FBy, a free bond specific to this level.

2. Depth-First Traversal of Search Tree. The structure generation problem can be represented by a depth-first traversal of the search tree illustrated in Figure 4. The top level, designated as level 0, contains only the root, which is a dummy node that represents the initial state of the structure generation. Each of the other levels that descends from a parent node represents all the alternative free bonds (noted as FBx) to connect a free bond (noted as FBy) specific to this level. Actually, these nodes correspond to some or all of the possible connections in the row corresponding to FBy in the FBMX. If the total number of free bonds is M, then the height of the tree should be $M/2$ and a path from the root to a leaf (a node in bottom level) represents the $M/2$ possible connections picked that form a complete and connected structure.

The search tree is implicitly constructed from FBMX as the tree is being traversed. In conventional implementations, the search tree can be arranged in the same order as the FBMX; i.e., take the first row as the first level, then the next row wherein no connection has been selected as the second level, and so on.[19] In our program, however, the search tree is arranged on the basis of the weights of the possible connections so that the levels consisting of the most possible connections with the highest weights are nearest the top level. Here, the array *paw*[ ] serves as the basis on which the hierarchy of the search tree is arranged: the first unsatisfied free bond of the EG that has the greatest *paw* value is selected as FBy, and its corresponding row of possible connections in the FBMX is then screened and reordered according to the following rules to form the new level.

1. Let the incident EGs of FBy and FBx be EGy and EGx, respectively, if FBy is not the first free bond of EGy; i.e., another equivalent free bond of FBy, FBy_old, has been taken as FBy in a previous level, and the selection of FBx's (or equivalent ones of them) that have already been tried to connect FBy_old in that level may cause generation of duplicated bonds so they should not be taken as the candidate FBx's of this FBy. For this purpose, a matrix tried_atm [ ][ ] is used to dynamically record those *tried connections*. The original entries of tried_atm[ ][ ] are set 0, if the connection between an EG pair has been tried in a level, the corresponding entry tried_atm[i][j] and its symmetrical entry tried_atm[j][i] are both incremented by 1. So if the corresponding tried_atm[ ][ ] of EGy and EGx is nonzero, FBx should be discarded.

2. Among the remaining FBx's, only the one with the lowest numbering is retained for each equivalent free bond set. If two EGs are equivalent, i.e., they are of the same symmetry class and of the same neighboring pattern, their free bonds are also viewed as equivalent ones and are processed similarly.

3. The remaining FBx's are sorted in the descending order of the weights of their associated possible connections to FBy and are then taken as a new level of the search tree.

The first and the second rules eliminate the potential duplication resulting from the permutation among the equivalent free bonds of the same EG pair. The third rule puts the most plausible connections in the front of a level.

Starting with the root, the depth-first traversal of the tree is carried forward by building up a new level as described above and, in the new level, exploring a proper node (i.e., a connection between FBy and FBx) according to the predefined order of the nodes. The choice of each connection is evalulated by a series of tests as described in the subsequent sections. If the tests are passed, the connection is added to CT and a descendant level is built up for this node and the traversal goes one level downward. When a proper node is chosen at the bottom level, a complete structure is produced and the structure is canonicalized[23] and compared with the previously produced ones to detect duplication. If it is a fresh one, the structure is stored as a candidate, otherwise it is rejected. In either case, no descendant level is built up any more; instead, the traversal is carried on to the next nodes at the same level. If all nodes at a certain level have been explored, the search goes up one level (i.e., to the parent node) in the tree, and the traversal continues at that level. If the backtracking takes place from the first level to the root, the traversal is completed; i.e., the structure generation process is finished.

3. Evaluation of the Intermediate Structures. Whenever a possible connection is chosen in the generation process, a series of tests are pursued to evalulate the plausibility of the currently generated structure, which is an intermediate one except when a connection at the bottom level is chosen. Each of the tests is responsible for returning an error message whenever a contradiction is detected, and if so, the chosen connection is rejected without further consideration. Most of the tests are optional and simple ones such as checking the connectedness of the intermediate structure and detecting the existence of some very simple "bad substructures" (e.g., cumulated double bonds and cyclopropane rings), only the three most important tests, which use the two kinds of the most important spectral information, LRDCs and chemical shifts, are discussed in detail below.

Test 1. The currently unsatisfied LRDCs are checked. For each of them, the distance between the concerned atom pair is calculated by a breadth-first search method. As the intermediate structure is incomplete, three kinds of results are possible; namely, (1) an indefinite distance, (2) a definite distance, or (3) a lower limit of the distance. In either of the latter two cases, the result is compared with the LRDC and, if contradictory, an error message is returned. If, on the other hand, a definite distance is obtained and is in accordance with the LRDC, this LRDC is marked as satisfied.

Test 2. The "rate" of LRDC satisfaction is checked. Generally, as the generation path extends, the number of satisfied LRDCs should also increase. Suppose that L_s, the number of satisfied LRDCs, increases linearly versus B, the number of connections that have currently been chosen, then the minimal number of LRDCs, L_{min}, that must be satisfied at a certain level can be calculated as

$$L_{min} = K_L L_0 B / B_0 \qquad (7)$$

where K_L is a user-defined slope of the line, or a minimum rate of the satisfaction of LRDCs in the generation path of the intermediate structure, L_0 is the total number of LRDCs,

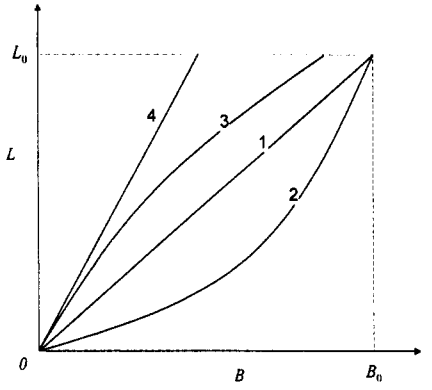

Figure 5. Illustration of the possible traces of L_s, the number of satisfied long-range distance constraints (LRDCs), as a function of B, the number of chosen connections along the structure generation path. L_0: the total number of LRDCs. B_0: the total number of connections to choose to form a complete structure.

and B_0 is the total number of connections to choose. If $L_s < L_{min}$, an error message is returned.

Test 3. If EGx (or EGy), the incident EG of the current FBy (or FBx), consists of a carbon atom and all its free valences have been satisfied, the ^{13}C chemical shift of the central carbon atom is simulated and compared with its observed value. For this purpose, a simple substructure–^{13}C chemical shift range correlation table is used. Such a substructure covers the central carbon being considered, the incident bonds, and the first layer of the neighboring atoms (the outward bonds of the atoms in the first layer are generic) and is referred to as a *carbon-centered single-spherical substructure* (CCSS). Currently, the ^{13}C chemical shift ranges of less than 100 common CCSSs composed of C, N, O, S, Cl, and P have been adapted from literature.[24] If the CCSS centered by EGx (or EGy) is found in the table, the corresponding chemical shift range is compared with the observed ^{13}C chemical shift of the central carbon. If inconsistency is detected, an error message is returned. If the CCSS has not been defined in the table, the user is signaled and the test is assumed to be passed by default.

4. "Direction" of Structure Generation. The evaluations of the intermediate structures actually prune those fruitless branches of the search tree as early as possible. Generally, the earlier the branches are pruned, namely, the earlier errors are detected in these tests, the more efficient the structure generation would be. For tests 1 and 2, it is demanded that the LRDC-involved bonds be generated in the early steps in the generation path so that a violation of LRDCs can be detected as early as possible and a bigger K_L can be used without losing the correct structure (the maximal K_L, with which at least one candidate structure is generated, is designated as K_L^M). As illustrated in Figure 5, the variation of L_s versus B should normally have a trace in the shape of line 1, or more unfavorably, because of the formation of multiple bonds, in the shape of curve 2. In our program, however, with the novel approach of rearrangement of the search tree which makes the more LRDC-involved bonds be formed in the early steps in the generation path, the trace of the function may be in the shape cf curve 3 or line 4, whereby a bigger K_L^M can be used in test 2. Similarly, for test 3, the carbon atoms are always arranged in the front of the EG set, which guarantees that the complete CCSSs are generated first, so that violations of chemical shift constraints could be detected as early as possible. In a word, the efficient application of these tests depends strongly on the arrangement of the search tree, as it is demonstrated in the accompanying paper.

While organizing the search tree, the sequence of formation of the multiple bonds gives rise the problem of "direction" of structure generation. To gain the greatest rate of LRDC satisfaction, it is demanded that the skeleton of the molecule be generated frist, i.e., only one connection be chosen for each EG pair until all the LRDCs are satisfied because any connections other than the first one between an EG pair do not increase L_s. We call this *connectivity-first structure generation.* On the other hand, test 3 demands that the complete CCSSs be formed first, namely, that not only the connectivities but also the types of the bonds be determined for the first layer of a central carbon atom. So the latter, called *CCSS-first structure generation,* is to some extent contradictory to the connectivity-first structure generation. In our program, the direction of structure generation is controlled by a parameter GEN_FLAG. If GEN_FLAG = 1, structure generation is carried out in the CCSS-first direction, i.e., when an EG (with the currently greatest value of *paw*) is selected, all its free bonds are consecutively taken as FBy's without interruption. If GEN_FLAG = 2, on the other hand, connectivity-first structure generation is adopted. Now, whenever the first connection is chosen for an EG pair, i and j, the weight of the connection between them is modified to a negative, i.e., $C(i,j) = -C(i,j)$, which in effect decreases the values of both *paw*[i] and *paw*[j] and therefore postpones the selection of the remaining free bonds of i (and also those of j) as FBy's. This makes, in most cases, that the skeleton of the molecule be generated first, leaving the multiple bonds to be complemented after all the LRDCs have been satisfied. In contrast, conventional structure generation strategies can be viewed as proceeding in a "random direction"; i.e., the arrangement of the search tree depends solely on the original arrangement of the FBMX, and the multiple bonds are generated sequentially as CCSS-first structure generation does. For the purpose of comparison, the user can invoke such a kind of structure generation by defining GEN_FLAG = 0. So, by specifying the value of GEN_FLAG, the user can choose the direction of the structure generation to emphasize either the constraint of ^{13}C chemical shifts or that of the LRDCs, or neither of them.

IMPLEMENTATION OF THE SYSTEM

CISOC-SES was developed on a microVAX 3300 computer under VMS V5.3. The source codes consist mainly of about 150 functions written in C, together with about 80 FORTRAN subroutines for the display of 2D molecular structures.[25] The size of the executive program is 225 kB. The program is run in interactive command-driven mode, and the results of analysis in each step are written in a common data file, which can be conveniently inspected and edited by the user and serves as the input of the subsequent steps. The structure generation step, which may take more CPU time, can be executed in batch mode after the data file is prepared. All parameters are stored in a parameter file which can be edited by the user at any step. Similarly, the knowledge base of CCSS/$\delta^{13}C$ range correlation table is stroed in a text file, which can also be edited by the user. Moreover, a routine that extracts such knowledge from a large ^{13}C spectral database is under development, which will be used to automatically expand and update the knowledge base. The schematic diagrams of 2D NMR spectra and the 2D diagrams of the deduced structures are displayed on VT330 or VT340 graphics terminals by using VT125 ReGIS commands.

CONCLUSION

Application of artificial intelligence in structure elucidation has been a topic for over 2 decades, yet most of the reported programs are limited in their capabilities to solve real-world problems. On account of this, while designing CISOC-SES, two essential problems have been addressed: (1) the 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 realize high-efficient structure generation. The high performance of the system in structure elucidation of some natural products of practical complexity is described and discussed in detail in the following paper in this issue.

The systematic and sophisticated exploitation of all available spectral information in the system can be summarized as follows:

1. Unambiguous direct AACNs, 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 structural fragments, are extracted as fixed bonds in the preparation stage, i.e., before structure generation. The search space is thus effectively reduced.

2. Ambiguous long-range AACNs, such as those 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 FBMX to reorganize the search tree and evaluation of intermediate structures based on the rate of LRDC satisfaction.

3. ^{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.

In contrast to the intensive use of 2D NMR correlation information, the use of ^{13}C chemical shifts in the present system is rather superficial because a CCSS covers very limited structural information. We think it better to use ^{13}C chemical shifts in a more comprehensive manner to further rank the candidate structures after the structure generation process (e.g., ref 26), and work in this respect is in progress. In addition, many reported heuristic rules, such as a prospective test of symmetry for symmetric structures[8,11] and more flexible use of user-supplied substructures,[8] can be readily adapted for our system. A further development of the system is hopefully to enable it to further discriminate candidate structures by using NOE information in a stereochemical context.

ACKNOWLEDGMENT

We thank Prof. Yuehua Xu, Prof. Weiming Chen, and Ms. Cuidi Zhu at our laboratory for supplying us the chemical structure display modules. Our thanks also go to Prof. Houming Wu, Prof. Yiwen Wang, Prof. Xinmao Zhong, Guanghong Weng, Jian Zhuang, and Ming Ye in the Institute and Prof. Xiuwen Han in the Dalian Institute of Chemical Physics for their helpful advice and generously offering us 2D NMR spectra for testing.

MAIN ABBREVIATIONS AND ACRONYMS USED IN THE TEXT

1D	one-dimensional
2D	two-dimensional
AACN	atom–atom connectivity
ACMX	atom (or more strictly, element group) connection matrix
AI	artificial intelligence
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
$c(i,j)$	element in FBMX, representing the probability of bond formation between the ith and jth free bonds
$C(i,j)$	element in ACMX, representing the probability of bond formation between the ith and jth EG
CASE	computer-assisted structure elucidation
CCSS	carbon-centered single-spherical substructure
CISOC-SES	computerized information system of organic chemistry-structure elucidation subsystem
COLOC	correlation via long-range coupling
COSY	correlated spectroscopy
CPU	central processing unit
CT	connection table (of a chemical structure)
DEPT	distortionless enhancement by polarization transfer
EG	element group, i.e., a non-hydrogen atom bearing a definite number of protons
FBx	candidate free bond to connect FBy of a level in the search tree
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 a 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
INEPT	insensitive nuclei enhanced by polarization transfer
LRDC	long-range distance constraints
K_L	demanded rate of LRDC satisfaction in the structure generation path of each candidate structure
K_L^M	maximal value of K_L, with which at least one candidate structure is generated
L_0	total number of LRDCs
L_{min}	minimal number of LRDCs that must be satisfied at a certain step in the structure generation path
L_s	number of actually satisfied LRDCs at a certain step in the structure generation path
M	size of FBMX or the total number of free bonds
MF	molecular formula
N	total number of EGs
NOE	nuclear Overhauser enhancement
SSCN	signal–signal connectivity
W	parameter used to weight FBMX based on LRDCs

REFERENCES AND NOTES

(1) Atta-ur-Rahman; Choudhary, M. I.; Pervin, A. Principles and Applications of Modern 2D NMR Techniques in Structure Elucidation of Complex Natural Products. In *Studies in Natural Products Chemistry*; Atta-ur-Rahman, Ed.; Elsevier Science Publishers B.V.: Amsterdam, 1991; Vol. 9, pp 127–161.

(2) Duddeck, H.; Dietrich, W. *Structure Elucidation by Modern NMR, A Workbook*; Springer-Verlag: New York, 1989; p 149.

(3) Rycroft, D. S. NMR in the Elucidation of Complex Structures Application to Two Novel Heptanortriterpenoid Derivatives from the

Stem Bank of *Entandrophragma utile* (Melianceae). In *Studies in Natural Products Chemistry*; Atta-ur-Rahman, Ed.; Elsevier Science Publishers B.V.: Amsterdam, 1991; Vol. 9, pp 93–107.
(4) Smith, D. H., Ed. *Computer-Assisted Structure Elucidation*; ACS Symposium Series 54; American Chemical Society: Washington, D.C., 1977.
(5) Pierce, T. H., Hohne, B. A., Eds. *Artificial Intelligence Applications in Chemistry*; ACS Symposium Series 306; American Chemical Society: Washington, D.C., 1986.
(6) For an overview of computer-assisted structure elucidation, see: Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Wiley & Sons: New York, 1986.
(7) Sasaki, S.-I. Structure Elucidation System Using Structural Information from Multisources. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 252–257.
(8) Christie, B. D.; Munk, M. E. Structure Elucidation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87–93.
(9) Yuan, S. G.; Peng, C.; Zheng, C. Z. Application of A C-13 NMR Topological Model to the Structure Elucidation of Organic Compounds. *Sci. China (Ser. A)* **1992**, *35*, 1136–1143.
(10) Lindley, M. R.; Shoolery, J. N.; Smith, D. H.; Djerassi, C. Application of Artificial Intelligence for Chemical Inference. 43. Applications of the Program GENOA and 2-Dimensional NMR Spectroscopy to Structure Elucidation. *Org. Magn. Reson.* **1983**, *21*, 405–411.
(11) Christie, B. D.; Munk, M. E. The Application of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Assisted Structure Elucidation. *Anal. Chem. Acta* **1987**, *200*, 347–361.
(12) Funatsu, K.; Susuta, Y.; Sasaki, S.-I. Introduction of Two-Dimensional NMR Spectral Information to an Automatic Structure Elucidation System, CHEMICS. Utilization of 2D-INADEQUATE Information. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 6–11.
(13) 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.
(14) Hass, P.; Robien, W. Automatic Interpretation of 2D-NMR-Spectra. In *Software Development in Chemistry 4, Proceedings of the Workshop "Computers in Chemistry"*; Gasteiger, J., Ed.; Springer-Verlag: Berlin–Heideiberg, 1990, pp 157–163.
(15) Nuzillard, J.-M.; Massiot, G. Logic for Structure Determination. *Tetrahedron* **1991**, *47*, 3655–3664.
(16) Matsuura, T.; Suzuki, H.; Abe, A.; Inagaki, F. Automatic Molecular Structure Analysis by Computer Manipulation of Two-dimensional Nuclear Magnetic Resonance Spectra. *Comput. Enhanced Spectrosc.* **1986**, *3*, 185–188.
(17) 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.
(18) For a review of computer-assisted biopolymer NMR data reduction, see: Hoch, J. C.; Redfield, C.; Stern, A. S. Computer-Aided Analysis of Protein NMR Spectra. *Curr. Opinion Struct. Biol.* **1991**, *1*, 1036–1041.
(19) Bohanec, S.; Zupan, J. Structure Generation of Constitutional Isomers from Structural Fragments. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 531–540.
(20) Actually, the method of Christie and Munk[11] can be employed to further eliminate the C–^{13}C allocations, but it is not implemented in our system yet.
(21) 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.
(22) Dubois, J. E. Ordered Chromatic Graph and Limited Environment Concept. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press London, 1976; pp 333–370.
(23) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.
(24) Bremser, W. Chemical Shift Ranges in Carbon-13 NMR Spectroscopy: Verlag Chemie: Weinheim, Germany, 1982.
(25) Xu, Y. H. *Chemical Structure Input Output System*. M. Sc. Dissertation, Shanghai Institute of Organic Chemistry, 1985.
(26) Gray, N. A. B.; Crandell, C. W.; Nourse, J. G.; Smith, D. H.; Dageforde, M. L.; Djerassi, C. Computer-Assisted Structural Interpretation of Carbon-13 Spectral Data. *J. Org. Chem.* **1981**, *46*, 703–715.