Computer-assisted structure elucidation of natural products with limited 2D NMR data: Application of the StrucEluc system

ARTICLE *in* MAGNETIC RESONANCE IN CHEMISTRY · MAY 2003

Impact Factor: 1.18 · DOI: 10.1002/mrc.1187

MAGNETIC RESONANCE IN CHEMISTRY
Magn. Reson. Chem. 2003; **41**: 359–372
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/mrc.1187

MRC

Computer-assisted structure elucidation of natural products with limited 2D NMR data: application of the StrucEluc system†

Kirill A. Blinov,[1] Dean Carlson,[2] Mikhail E. Elyashberg,[1] Gary E. Martin,[3] Eduard R. Martirosian,[1] Sergey Molodtsov[4] and Antony J. Williams[2]*

[1] Advanced Chemistry Development, Moscow Department, 6 Akademik Bakulev St., Moscow 117513, Russia
[2] Advanced Chemistry Development Inc., 90 Adelaide Street, Suite 702, Toronto, Ontario, Canada, M5H 3V9
[3] Rapid Structure Characterization Group, Global Pharmaceutical Sciences, Pharmacia Corporation, 7000 Portage Road, Kalamazoo, Michigan 49001-0199, USA
[4] Novosibirsk Institute of Organic Chemistry, Siberian Branch of the Russian Academy of Science, Lavrentiev Avenue 9, Novosibirsk 630090, Russia

Received 1 December 2002; Accepted 23 January 2003

This paper considers the strategy of the StrucEluc expert system application for structure elucidation of new natural products when there is a lack of connectivity information that is characteristic of proton-deficient molecules. It is shown that in this case, a database search for fragments using a ^{13}C NMR spectrum as input allows an investigator to fill gaps in the recorded data. Algorithms and programs have been developed that allow fragments found in the library and/or proposed by the user to be embedded in the molecular connectivities diagrams built on the basis of 2D NMR data analysis. We demonstrate the structure determination of three alkaloids from the cryptolepine series using the principles of construction and application of a user fragment library. The approach described appears to be the most efficient means of structure elucidation for natural products with 2D NMR spectra characterized by sparse responses. Copyright © 2003 John Wiley & Sons, Ltd.

KEYWORDS: NMR; ^{1}H NMR; ^{13}C NMR; ^{15}N NMR; expert system; structure elucidation; spectral analysis

INTRODUCTION

Natural product structure elucidation has been profoundly changed by the advent of 2D NMR methods.[1,2] The development of robust, heteronucleus-detected chemical shift correlation methods in the early 1980s, first by direct [^{1}J(C,H)] correlation and a few years later by long-range [^{n}J(C,H), $n = 2$–4 but more generally 2, 3][3] has made the elucidation of complex structures, regardless of their source, a more facile process. Heteronucleus-detected methods were supplanted in the late 1980's by the development of proton- or 'inverse'-detected heteronuclear chemical shift correlation methods, with a considerable improvement in sensitivity, thereby reducing sample sizes required for structural characterization to a milligram or less of material.[4–7] The experiments *de jour* in 1986 were HMQC and HMBC, for direct and long-range correlation, respectively.[6,7]

Improvements in NMR sensitivity during the late 1980s were based largely on improvements in the NMR methods being utilized, the development of so-called 'inverse' geometry or inverse-detection NMR probes, and on increases in observation frequency coupled with improvements in instrument design and stability. The early 1990s brought the first reports of 3 mm or 'micro' inverse NMR probes, heralding another substantial reduction in sample requirements to considerably less than 1 μmol of material for direct and long-range heteronuclear shift correlation experiments.[8,9] More recently, 1.7[10–12] and 1.0 mm[13,14] conventional small-volume NMR probes and both 3 and 5 mm cryogenic NMR probes[15–17] have also been developed that further enhance experiment sensitivity with concomitant reductions in sample size. The ongoing advances in both hardware and pulse sequence technologies facilitate the generation of high-quality, high-content data and offer a very high probability of generating the appropriate data for analysis. However, the primary bottleneck remains the analysis of the data itself to elucidate the chemical structure.

The development of expert systems (ES) or CASE applications (computer-assisted structure elucidation) for the elucidation of structures of organic compounds has been an ongoing effort for over 30 years. Initially a purely academic exercise, these efforts have led to the development of a number of commercial applications that have been tried and tested. The problem, of course, is extremely complex, and numerous approaches have been taken. At present,

†Dedicated to Professor William F. Reynolds on the occasion of his 65th birthday.

*Correspondence to: Antony J. Williams, Advanced Chemistry Development Inc., 90 Adelaide Street, Suite 702, Toronto, Ontario, Canada, M5H 3V9. E-mail: tony@acdlabs.com there are approximately 20 applications available that can be applied to spectral data analysis to elucidate a molecular structure. In recent years the following applications have been reported in the literature: RASTR (STREC),[18] X-PERT,[19] SESAMI,[20,21] CHEMICS,[22] SpecSolv,[23] CISOC-SES[24]/NMR-SAMS,[25] COCON,[26,27] LSD,[28] LUCY,[29] SENECA[30] and others.

For the modern CASE systems, 1D ^{13}C, ^{1}H and 2D NMR spectra, COSY, HMQC/HSQC and HMBC and the molecular formula defined on the basis of high-resolution mass spectrometry (HRMS) usually serve as the raw data inputs. The output is one or more molecular structures consistent with the data and any input constraints imposed by the user. The software parameters are usually adjusted so that the default separation between correlated resonances in a COSY spectrum is via $^{2,3}J$(H,H) and in an HMBC spectrum, $^{2,3}J$(X,H). For purposes of discussion in this report, these correlations will be defined as 'standard.'

A review of the literature of CASE systems shows a series of specific commonalities:

- Published work describing various CASE algorithms are usually illustrated by one or more examples of solved structures. The chemical structures elucidated generally range from minor to moderate complexity. In general, the work reported has not considered the general applicability of the algorithms to diverse structural space, providing instead only distinct examples of success.
- The various algorithms described utilize 2D NMR correlations that determine connectivities between individual carbon atoms. The ability to use substructural fragments as inputs to enable the elucidation process does not exist. Since a spectroscopist can commonly hypothesize certain substructural moieties contributing to the target structure such a capability would be a natural capability for a CASE system.
- Previous work has not provided algorithms for the analysis of 2D NMR spectroscopic data allowing for the detection and removal of contradictions relating to unidentified long-range couplings [^{n}J(H,H), ^{n}J(C,H), $n \geq 4$].

The presence of long-range couplings producing 2D NMR data contradictions can be a serious issue impacting the successful application of CASE programs and cannot be ignored. The examples reported by Munk and co-workers[20,21] demonstrate that the allowed distance between the coupled nuclei can be defined by the user. However, the CASE system described here has no ability to detect non-standard correlations.

To avoid the inclusion of possible long-range correlations [^{n}J(H,H), ^{n}J(C,H), $n \geq 4$] into the analysis of 2D NMR data, it has been suggested[26,27] that structure generation can be performed in an iterative manner. For the analysis reported, the first attempt to elucidate the structure excluded low-intensity HMBC peaks. When these correlations were included in the elucidation process, no resultant structure was generated by the CASE system. It was concluded that these correlations corresponded to long-range couplings ($n \geq 4$). This approach to the detection of contradictions can be ineffective since the exclusion of some correlations may dramatically increase the time required for the automated elucidation of the final structures. Moreover, this approach is dangerous since the computer program may actually interpret long-range couplings as standard ones. This can occur if the signals in a ^{13}C NMR spectrum are assigned to a potential chemical structure in such a way that the order of some long-range correlations does not contradict the *standard* default values. The result is that the correct structure is commonly absent from the resulting solution set. This situation has been considered in detail previously.[31] The necessity to identify and remove potential contradictions is obvious. There is not an exclusive relationship between peak intensities and the value of a coupling constant and there are many cases when intense signals are observed for couplings separated by four or more bonds, or conversely, there can also be unduly large two-bond correlations that can be misinterpreted as three-bond correlations that usually predominate.

For proton-deficient molecules, the small number of observable 2D NMR correlations can be the primary hurdle to elucidating the structure. This means the set of constraints that can be imposed to determine the molecular skeleton is insufficient to allow a bounded structural file to be generated in a reasonable time. This issue has been considered by authors of the COCON program.[26,27] To reduce the number of possible structures, the authors propose the use of 1,1-ADEQUATE experimental data. This approach does not resolve the general problem of structure determination for newly isolated natural products that are proton deficient because the 1,1-ADEQUATE experiment is very time consuming and requires a high concentration of sample or access to state-of-the-art probe technology such as cryoprobes.[17]

To allow the automated elucidation of larger molecular structures and take into account longer range correlations such as ^{4}J(C,H), which may be commonly detected in HMBC data, Steinbeck[30] included a stochastic structure generation algorithm as part of the SENECA system and commented that the largest molecule that a deterministic system is capable of processing will likely not exceed 30 skeletal atoms. Contrary to this premise, the StrucEluc system,[31–34] a deterministic expert system, has been able to elucidate chemical structures with 30–90 heavy atoms. In our opinion, the limitation of a deterministic system depends to a large extent on the number of HMBC and COSY correlations detected in the 2D NMR spectra. This is obviously highly dependent on the number of hydrogen atoms in a molecule rather than on the size of the structure under study coupled with the sensitivity of the NMR probe technology in use to acquire the requisite data. There is also the expectation that the data are recorded at an observation frequency sufficient for all of the resonances to be resolved.

Among the molecules examined by Steinbeck,[30] the largest was the polycarpol structure ($C_{30}H_{48}O_2$, 32 heavy atoms). To elucidate this structure the SENECA system performed 350 000 iterations within 12 min using eight parallel processors running at a processor speed of 600 MHz. Using the same 2D NMR data as input the StrucEluc software program, running on a single processor PC with a clock speed of 500 MHz, generated six structures within 1 s in the *fully automatic* mode. In this mode the hybridization of the carbon atoms and the nature of their neighboring heteroatoms were automatically determined by the program. The correct structure was identified utilizing the ^{13}C NMR prediction capability to rank order candidate structures. It is difficult to comment on the ability of the SENECA system to generate large structures involving several ^4J(C,H) and $^{4,5}J$(H,H) correlations on the basis of the data reported since only ^4J(C,H) correlations are discussed in that report.[30] Our experience to date indicates that the total number of long-range ('non-standard') COSY and HMBC correlations in typical 2D NMR spectra for fairly complicated chemical structures might exceed 10, thereby complicating the task considerably. For instance, it was shown that application of the ACCORD-HMBC techniques allowed authors[35] to identify 18 ^4J(C,H) responses for strychnine. It is also worth noting that the operation of the SENECA system appears to require a close familiarity with the stochastic algorithm used for structure generation.

It should be evident that proton deficiency is directly associated with a reduction in the amount of structural information available from some 2D NMR experiments, although some of the more recent pulse sequences, specifically the accordion-optimized sequences reported by Martin,[36] allow a potentially higher number of correlations to be extracted from a single 2D NMR experiment. Without this information it becomes necessary to overcome the limitations of the data by other means, for example introducing additional structural information that could better define some of the relationships between the skeletal atoms. As previously reported,[31–34] a valuable approach to this problem requires the introduction of fragments known to be present in the molecule under examination. These data can be available based on expert interpretation allowing partial elucidation of the structure, or via the availability of alternative analytical data, or simply due to prior information regarding the starting materials for a particular synthesis. To this end, it is necessary to invoke large and diverse libraries[33] containing molecular fragments and their associated assigned subspectra. We have previously shown that an approach whereby the scientist suggests user fragments can greatly alleviate issues regarding the automated elucidation problem. Structural hypotheses of this kind are fairly common in chemical research since it is much easier to validate them rather than to elucidate the proposed structure completely.

To implement the approach described, algorithms have been elaborated[31] that allow the fragments to be embedded into the network of connectivities identified via the 2D NMR spectra or proposed by the investigator based on prior or ancillary knowledge. Generally this approach is simpler and more efficient in determining the structures of large molecules than, for example, the stochastic algorithm.[30] It is particularly applicable to the reduction of the number of isomers that can correspond to the experimental 2D NMR data.

To solve the problem of contradictions in a 2D NMR dataset, we suggested[31,32,34] an heuristic algorithm for searching for skeletal atoms displaying connectivities of non-standard lengths. When a particular atom is identified, the program tries to remove contradictions in connectivities related to this atom by lengthening the corresponding number of bonds associated with the connectivities. Experience has shown that typically the program does reveal the presence of the contradictions. The automatic removal of these contradictions is not always achieved owing to the heuristic character of the algorithm. Therefore, to allow removal of the contradictions by the user directly, empirical rules have been suggested.[32]

The algorithms described previously[31–34] have been implemented into StrucEluc. The system has been demonstrated to be generally applicable to the structural determination of organic molecules. In order to challenge the system with structural diversity, we have applied the system in particular to a series of natural products using their 1D ^{1}H and ^{13}C and 2D NMR spectra as inputs. A structure library of about 200 000 structures and a fragment library of about 1 000 000 fragments have been used as the basis of the program. ^{13}C NMR chemical shifts that have been assigned to their respective carbon atoms provide a foundation knowledge base for the system. Both the efficiency and reliability of the system have been confirmed by solving over 100 natural product molecular structures. These challenges were approached using two modes: one based on the utilization of 2D NMR connectivities only and the other based on the use of the fragment library.

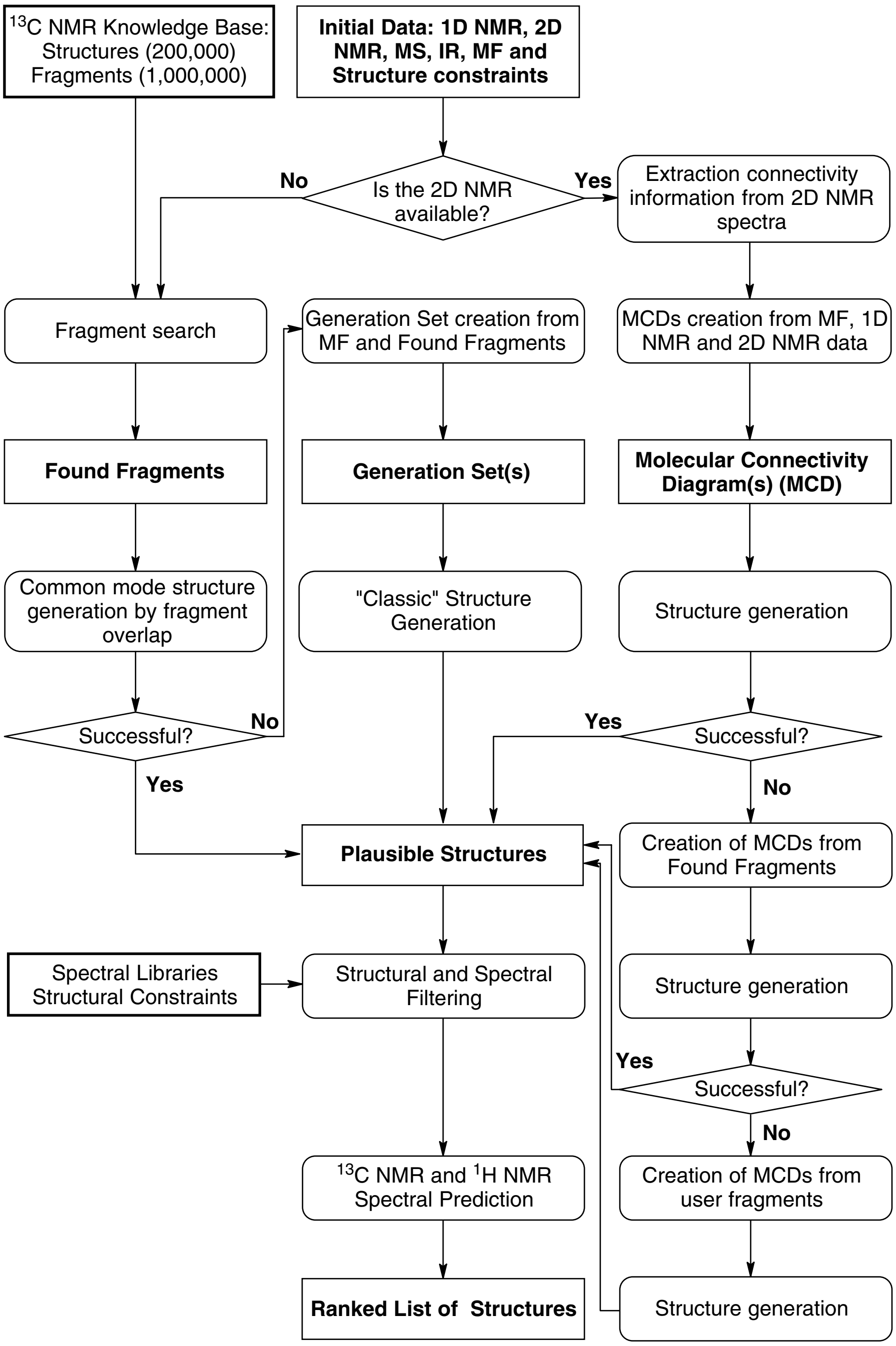
This paper describes the utilization of fragments from the system's knowledge base, and also the use of a user fragment library, to allow the elucidation of natural product chemical structures on the basis of 2D NMR spectra. Particularly we focus on the identification of three indoloquinoline-based alkaloids from the cryptolepine series[37–39] that could not be solved in either the *conventional* mode or via application of the system knowledge base. Rather, in order to solve the problems we had to create a specialized user library of fragments related to the natural products in the series. The ability to create the user knowledge base and its application to elucidate the structures for complex natural products are described.

RESULTS AND DISCUSSION

The methodology of molecular structure determination with the aid of the StrucEluc system

In the initial version of the StrucEluc system,[33] the 1D ^{13}C NMR spectrum, the molecular mass and the molecular formula were the initial data inputs. When available, both IR and ^{1}H NMR spectra could be appended as additional data for use as structural filters. With ongoing development, the latest version of the system[32,34] has been expanded to allow 2D NMR spectral data to be utilized. This has been shown to be critical for the structure elucidation of novel natural products. The basic process describing the StrucEluc system is indicated in Fig. 1. The strategy of structure elucidation adopted is dependent on the nature of the experimental data and prior information available.

If the user has access to only 1D NMR spectra, the system knowledge base is used and an attempt to identify the structure is performed using two possible modes. The first mode approaches structure generation from fragments found in the



^{13}C NMR Knowledge Base: Structures (200,000) Fragments (1,000,000)
Initial Data: 1D NMR, 2D NMR, MS, IR, MF and Structure constraints
Is the 2D NMR available?
No
Yes
Extraction connectivity information from 2D NMR spectra
Fragment search
Generation Set creation from MF and Found Fragments
MCDs creation from MF, 1D NMR and 2D NMR data
Found Fragments
Generation Set(s)
Molecular Connectivity Diagram(s) (MCD)
Common mode structure generation by fragment overlap
"Classic" Structure Generation
Structure generation
Successful?
No
Yes
Yes
Successful?
No
Plausible Structures
Creation of MCDs from Found Fragments
Spectral Libraries Structural Constraints
Structural and Spectral Filtering
Structure generation
Yes
Successful?
No
^{13}C NMR and ^{1}H NMR Spectral Prediction
Creation of MCDs from user fragments
Ranked List of Structures
Structure generation

Figure 1. General flow diagram of the StrucEluc system.

system library (found fragments, FF). In this case, the program will attempt to select fragments with common atoms and the generation process subsequently attempts to assemble a molecule out of these fragments. In this mode, the so-called *common mode*, the program can work even without a molecular formula. If the program does not succeed in assembling one or more molecules, it automatically switches to the second mode of structure generation, which is referred to as the '*classic*' mode. In this mode, the program uses molecular formula input or a molecular formula determined from the experimental molecular mass to build up a set of fragments corresponding to the formula. The user has the ability to manage the generation process and edit the resulting output sets. For ease of use the program displays a 'generalized portrait' of the molecule under investigation. Typical functional groups are sorted by their frequency of occurrence in the series of fragments selected from the library in the first stage. A special library of typical functional groups is used to produce this general portrait of the molecular structure. The more frequent the incidence of a particular functional group fragment in this process, the greater is the likelihood that the fragment in question is a part of the molecular structure. Functional groups whose frequency of occurrence in the found fragments list exceeds a threshold number set by the user can automatically be included into a GOODLIST for use in the final elucidation stage. The functional groups which are absent from the list, but present in the functional group library are used to create a BADLIST.

The next stage of the elucidation process creates structural formulae on the basis of fragment sets. The final output file contains only structures that have persisted through a filtering process in which spectral–structural correlation libraries are established using NMR and IR spectra. The output file also passes structural constraints imposed by the user. As is customary for all expert systems, the most probable structure can be identified from the resulting library of structures using ^{13}C NMR spectral prediction.

The StrucEluc system provides a two-step procedure for the most probable structure identification in the output file. During the first step, ^{13}C NMR spectra are predicted for all structures using an incremental method, our so-called 'Fast' method, and the structures are ranked by the d_F value, the average deviation of the experimental versus calculated chemical shifts, sorted in ascending order. The smallest d_F value indicates the best match between the experimental and calculated spectra and this structure will be the first in the list. During the second stage, more accurate ^{13}C NMR spectra are calculated for the first 10–25 structures of the ranked file. These accurate predictions are performed using a database of fragments with the corresponding assigned subspectra. The description of each nuclear environment is defined using the HOSE code approach[40] (hierarchical ordering of spherical environments). The average deviation values between the experimental and calculated values (d_A) are found and the structures are again rank ordered. Subsequent ranking dramatically increases the probability of moving the correct structure to the first position in the list. For additional control over the correct choice of the output structure, the proton chemical shifts can be predicted and displayed together with the corresponding deviation value, d_H. For proton NMR prediction the predicted proton–proton couplings can be enhanced by three-dimensional optimization of the structure. In ambiguous cases, it may be useful to display the calculated ^{1}H NMR spectra because of the complexity of some of the multiplets. To facilitate structure analysis in the output file, the StrucEluc system is augmented with a feature that calculates structural similarity coefficients. In this way if the investigator has an idea of the class of structure under investigation he can use this structure as an input to allow rank ordering relative to the structural similarity of the results file.

Assuming the availability of 2D NMR spectra, it is appropriate to initiate structure elucidation from these data since they have much richer information content relative to 1D NMR spectra. Commonly the input 2D NMR spectral dataset includes HSQC (HMQC), COSY and HMBC. With this approach there are two modes of identification possible.

The first mode is referred to in this paper as the 'common' or '*ab initio*' mode. Tables of peaks and correlation cross peaks are input into the program either via manual data entry or by direct extraction via the data processing components of the software. On the basis of these inputs, the program creates molecular connectivities from the relationship between the 1D chemical shift data and the extracted cross peaks. A library of spectral–structural correlations for both ^{1}H and ^{13}C NMR spectra allows valence and molecular environment properties to be assigned to each carbon atom. These properties include the atom hybridization type (sp, not sp, sp^2, sp^3) and the identification of potential neighborhoods around a nucleus with one or more heteroatoms (*obligatory*, *forbidden*, *at least one*, *at least two*, *at least three*, *or four heteroatoms*). If the program does not succeed in assigning such a property unambiguously, it is marked as *not defined* (*nd*). All atoms from the molecular formula along with their connectivities are displayed in a graphical format which is referred to as a molecular connectivity diagram (MCD). At this stage, users are able to analyze visually the connectivity diagram (atom-to-atom correlations) and edit these connectivities based on personal judgments and biases if they so choose. Then the program attempts to determine whether the connectivities correspond to typical default values describing the number of bonds between coupled spins; the default values within the program are ^nJ(H,H), ^nJ(C,H), $n = 2$–3 for COSY and HMBC correlations, respectively. If one or more nuclei are found to display at least one connectivity of a non-standard length, the program attempts, based on user permissions, to remove the contradiction. This process proceeds by extending all connectivities originating from this atom. The number of non-standard HMBC and COSY connectivities may be large, up to 10 or even more based on our experience, and the program by necessity has to remove the contradictions in an iterative mode. As the connectivity correction is completed, a new MCD is displayed with all changes made by the procedure shown highlighted with color. Our experience to date shows that, for the most part, the program correctly reveals contradictions. However, the detection and removal of all contradictions is not always possible owing to the heuristic nature of the algorithm. If the program fails, a message alerts the investigator that the contradictions were not removed. At this stage the user can attempt to edit the data manually to remove the contradiction or generate fresh data that allow the correlations lengths to be defined more accurately.

In a recent paper,[31] we showed that if the program fails to detect implicit contradictions, in some cases, structures similar to the correct structure or, at least containing the same large fragments, can be generated on the basis of the 2D NMR dataset, thereby yielding at least an approximate solution. We emphasize that the main objective criterion for considering the first-ranked structure to be correct is the similarity between the calculated and experimental spectra. The evaluation of the structure by chemists using information *a priori* can, of course, serve as an additional criterion.

The detection of contradictions in 2D NMR data is not the only issue that can hamper automated determination of the molecular structure from the input data. If the unknown compound is highly unsaturated and therefore proton deficient, the number of COSY and HMBC correlations may be insufficient to build up an efficient constraint system. Such molecules can contain rather large fragments with skeletal atoms devoid of hydrogen. These 'silent' fragments can result in both the COSY and HMBC experiments showing no correlations into a given region of the molecular structure. The result is greater freedom for the combination of both quaternary and heteroatoms from these fragments and more opportunity for generating potentially consistent structures. This is just as much an issue for a scientist interpreting the data as it is for a software program. To some extent, this problem can be addressed by using one of the new accordion-based long-range heteronuclear shift correlation experiments[36] that have the potential to provide more efficient experimental access to 'longer' range, e.g. $^{n}J(\mathrm{X,H})$, $n \geq 4$.

If the molecule being analyzed is comparatively small, the StrucEluc software program may be able to accomplish structure generation in a reasonable time, but one can anticipate that the number of generated structures can become significant and time consuming to filter and examine. If the molecule is large enough, a situation which is common for natural products, the generation time may amount to many tens of hours and the number of suggested structures to hundreds of thousands. However, if at least a part of the 'silent' fragment can be incorporated into the MCD as a defined substructure, then the solution is partially constrained and the problem of elucidation is correspondingly reduced. This approach is typical of the thought process that spectroscopists use when elucidating a structure.

Difficulties may also arise if the number of observed 2D NMR correlations is much smaller than the number of responses expected by visual inspection of the molecule. This can happen for a variety or reasons that include stereochemical features, low signal-to-noise ratios in the 2D NMR spectra due to limitations in sample size or hardware limitations, low observation frequencies or poorly performing probe technologies. Even though significant leaps in technology[15] can dramatically impact the quality of data which can be fed to StrucEluc, such equipment is currently available in only a limited number of laboratories. Occasionally the molecule under study is so large that even information-rich 2D NMR spectra can result in an extended time for structure generation. This is especially common if the molecule contains a lot of carbon atoms in the sp^2 hybridization state, commonly associated with a number of fused aromatic ring systems. Accidental spectral degeneracy can significantly increase ambiguity, with a corresponding increase in calculation time. We have previously suggested[31] algorithmic and programmatic approaches for using input fragments to aid in the determination of the molecular structure to assist the 2D NMR spectral analysis. For this purpose, fragments from a fragment library containing about 1 000 000 fragments, as well as user-defined fragments introduced by chemists on the basis of prior knowledge, can be utilized. It is possible to embed these fragments directly into the MCDs if they include carbon centers with the appropriate chemical shifts. The fragments from the library include assigned carbon atoms transferred from the molecules used in the construction of the structure library databases.

The process of fragment inclusion has several stages. First, fragments are selected from the library by searching for subsets of the resonances contained within the ^{13}C NMR spectra. The number of found fragments L can, in practice, vary from several hundred to several thousand. The next step is the creation of MCDs using the found fragments (FF) as the basis. All FF or those selected at the users discretion can be selected to create the MCDs. The basic idea of the algorithm implementing this procedure is as follows:

A value for an error E, to define the maximum allowed deviation between a chemical shift of a carbon atom in a found fragment and the corresponding experimental value, is set by the user. ^{13}C NMR subspectra of each fragment are compared with all the experimental chemical shifts including the multiplicity information. Consider a fragment that contains f carbon atoms and an atom C_i of a fragment has a chemical shift $\delta_i (i = 1\text{–}f)$ and multiplicity m_i. Assume that for a series of experimental chemical shifts $\delta_{i1}, \delta_{i2}, \ldots, \delta_{iq}, \ldots, \delta_{ip}$ the following condition is true: $|\delta_i - \delta_{iq}| \leq E$ and $m_i = m_{iq}$. Then all possible ways of substituting $\delta_{i1}, \delta_{i2}, \ldots, \delta_{iq}, \ldots, \delta_{ip}$ for δ_i should be checked. If the conditions $|\delta_i - \delta_{iq}| \leq E$ and $m_i = m_i$ are true for all f carbon atoms, the given fragment is regarded as a candidate fragment for building up further connectivities and otherwise the fragment is excluded from the analysis. There are cases when experimental chemical shifts δ_{iq} can be related to the chemical shifts of several carbon atoms in a fragment. In this case, all arrangements of the experimental shift δ_{iq} on the corresponding carbon atoms in the fragment are considered, and each distribution of chemical shifts producing a new assignment of carbon atoms in the fragment, should be verified. The program checks if the assignment of the carbon atoms can correspond to the correlations between the experimental chemical shifts for the fragment skeletal atoms. The fragments which survive the algorithmic examination are retained in the list of possible fragments.

Obviously, the greater the number of skeletal atoms identified in the selected fragments that can be present in the molecule, the greater is the opportunity to resolve the correct structure and the faster the process will be completed. An algorithm to combine possible fragments within one molecular connectivity diagram has been developed and in this process, all combinations of possible fragments are searched and only combinations that correspond with the experimental 2D NMR correlations are selected. Fragment combinations consistent with the experimental 2D NMR correlations make up additional potential fragments to be used by the program. Along with remaining free atoms that have not been identified during the fragment search, these fragments are 'projected' on to the MCDs, allowing the user to analyze the diagrams visually and again introduce human intervention if necessary to influence the elucidation. In practice, the number of MCDs built up out of FFs varies greatly and in some cases can produce several thousand MCD outputs.

The error value, E, is an important value for the success of the process and it is chosen empirically by increasing it from 0.5 ppm until a value is found that allows the creation of at least one MCD. The set of MCDs delivered is then checked for contradictions, and structure generation is performed. Occasionally, an iteration may yield a set of MCDs that deliver an incorrect solution. As previously stated, a solution is correct only if it contains the correct structure. Experience suggests that commonly an incorrect solution can be identified when the minimal d_A value exceeds 5 ppm, although this is not absolute. As E is further increased in size the number of MCDs also increases and the new MCDs may generate the correct solution.

System efficiency

In a previous paper,[32] we demonstrated for the first time the systematic evaluation of the scope and efficiency of an expert system by the application of StrucEluc to a large number of natural products. Specifically the evaluation of the 2D NMR analysis system efficiency as a routine tool to allow the structure determination of newly isolated natural products was examined.

The chemical structures and associated 2D NMR data of the natural products that we have examined were, for the most part, extracted from articles in the *Journal of Natural Products* published in the period 2000–2002. Another series of elucidation problems were obtained from collaborative efforts with scientists who provided us raw spectral data. To date more than 120 natural product structures have been determined using the StrucEluc system. These have included relatively large and complex molecules with 20–90 skeletal atoms and molecular masses ranging from 200 to 1285 u.

Each problem set was originally examined with the standard default connectivity lengths as defined previously. Approximately 70% of the problems could be solved in this way. The program did find controversial connectivities caused by ^{4-5}J long-range couplings in those cases when the authors indicated in tables all the correlations observable in spectra. If the program did not reveal any contradictions, the correct solution was usually generated and the output file contained the correct structure. In some cases, though rare, the program failed to reveal any contradictions in the datasets.

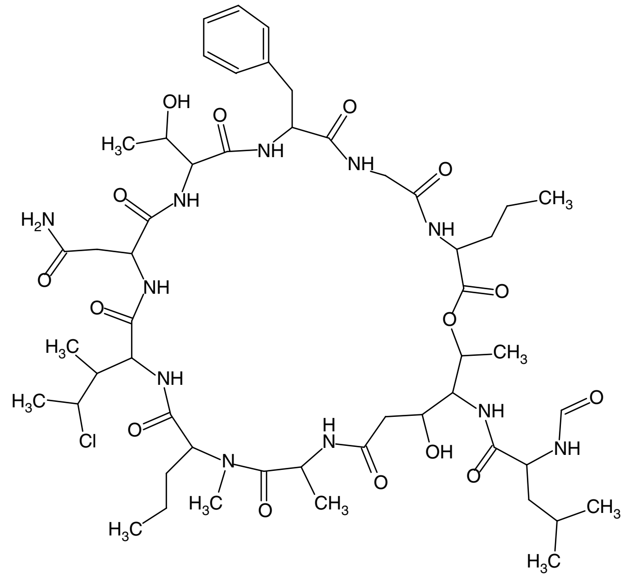
Our work indicates that one formal indication of a possibly incorrect solution is a large $d_A(1)$ value, generally >5 ppm calculated for the first structure in the priority ranked file. The $d_A(1)$ deviation value did not exceed 4 ppm in 85% of the problems studied. In a few cases the deviation value was greater, therefore necessitating thorough verification of the solutions. One important criterion for confirming the validity of a solution is evaluation by the chemist on the basis of prior information and/or supplementary experiments.

If the attempt to remove contradictions automatically fails, then the user can attempt to remove them by making a particular hypothesis and verify it experimentally by solving the problem. Investigations have shown that extending all methyl group-related correlations by one bond length provided a solution for almost half of the problems where contradictions were found. Using this approach increases the amount of time required for structure generation with a corresponding increase in the number of structures in the output file. Lengthening correlations involving $=CH_2$ and $=C—H$ structural fragments has also been shown to help solve the problem. Extending correlation lengths to these fragments seems to have an effect only if the number of these fragments in the analyzed molecule is relatively small.

The results of many challenges to the StrucEluc system have shown that the correct structure is generally the first-ranked structure in an output file even if this file contains thousands of structures. This underscores the validity of our suggested structure determination strategy. Simultaneously we have determined that large output files are relatively rare: the number of structures in the output file did not exceed 10 in about 70% of cases, thereby confirming the highly selective capability of the software system.

It should be noted that a large output file containing hundreds or thousands of structures is not a hindrance for determining the correct structure using the suggested methods outlined here. In 90% of the problems examined to date the correct structure was ranked first in the structural file ranked according to the d_A value. It was also found that even preliminary ranking of structures on the basis of increment-based 'fast' prediction of ^{13}C chemical shifts ranked the correct structure in the first position in 80% of cases. For this work it was shown that it took less than 1 min to find a solution for 75% of cases, and less than half an hour in 95% of cases.

A particular set of problems was identified that could not be solved for a number of reasons. Particularly challenging were large cyclopeptide molecules, for example, phoriospongin A[41] as illustrated.



This molecule has the formula $C_{52}H_{82}N_{11}O_{15}Cl$ with 79 heavy atoms. The spectral data[41] produced over 300 000 structures in just over 17 h. Of these structures, 624 were non-isomorphic. After ranking according to the d_A value, the correct structure was number 51. This observation suggests that the larger the molecule being studied, the less useful d_A becomes for identifying the correct structure. The reason for this is probably the leveling of the deviation values due to the presence of a large number of carbon atoms having similar properties. In this case, spectra of the top-ranked structures have very similar calculated spectra that are characterized by regions containing many tightly situated shifts. To solve such problems and determine the actual structure, additional information is generally required regarding molecular topology. For example, when a constraint was imposed assuming the presence of a 30-membered cycle in the molecule, the actual structure was ranked in eighth position. The determination of the structures of some other peptide molecules failed owing to the huge amount of time required for structure generation. It is likely that these problems will be solved in the future using a specialized user database of fragments focused on the determination of such structures. In contrast, when the structure of durhamycin A[42] ($C_{62}H_{92}O_{28}$, 90 skeletal atoms) was elucidated the correct structure was distinguished in the output file of 66 structures both by the 'fast' and accurate methods of ^{13}C NMR spectrum prediction.

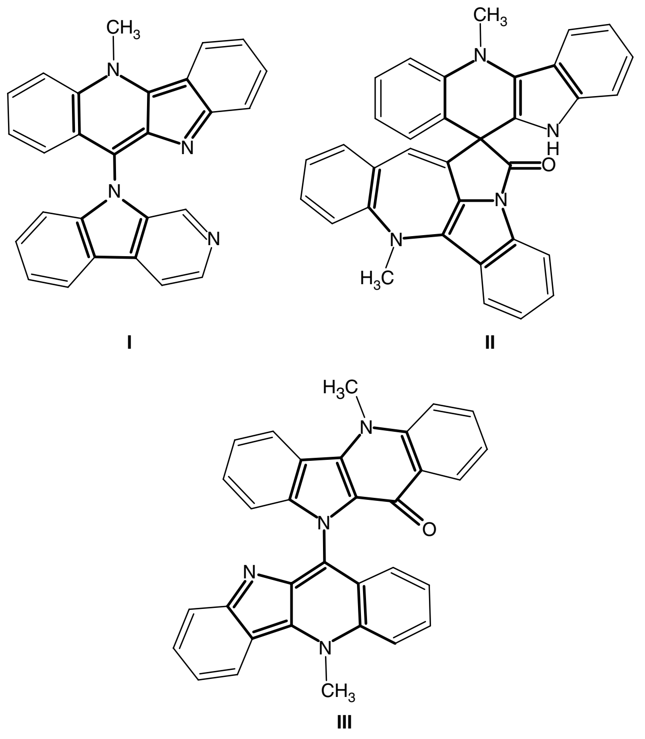
Challenging problems

During the course of this work, a number of complicated and challenging structure problems were studied. Typically in these problems when attempting the structure elucidation in the common mode, structure generation was not accomplished after more than 48 h. Attempts to solve these problems using fragments found in the system knowledge base (FF) also failed. The failure to utilize library fragments was due to the following issues:

- Fragments appropriate for a given problem are missing in the knowledge base.
- Appropriate fragments are found but the number of possible variants of carbon atom assignments in these fragments is so huge (more than 100 million) that the computer runs out of resources attempting to sort out all possible permutations.
- The number of MCDs built up by the program is so huge that the completion of structure generation is simply too long and human intervention halts the process.

Among the structures that we failed to determine in the common mode and using library fragments were the following three alkaloids from the cryptolepine series: cryptolepicarboline (**I**), cryptospirolepine (**II**) and 5,5′-dimethyl-5′H-10,11′-biindolo[3,2-*b*]quinolin-11(5H)-one (**III**). The structures of these natural products are shown in Scheme 1 and the related data are given in Table 1.

These molecules are relatively large, highly unsaturated and have four condensed benzene rings and double bonds contained in other cycles. All the molecules have large fragments (displayed in bold) containing no hydrogen atoms. These fragments account for half of the skeleton of the molecule, thus limiting the utility of COSY data. For structures **I**–**III** COSY correlations are observed only for the contiguous protons in the 1,2-disubstituted benzene rings. Cryptospirolepine (**II**) has an especially complex structure consisting of two planar fragments linked through a carbon spiro-center and therefore lying in perpendicular planes. Each fragment makes up a system of conjugated bonds. One can expect long-range correlations (^nJ(C,H), ^nJ(H,H), $n \geq 4$) inside the planar fragments and little or no spin–spin coupling information transfer between the fragments. An unexpected chemical shift for the carbonyl amide group in the ^{13}C NMR spectrum is observed at 188.4 ppm. It is more likely to correspond to a ketone carbon resonance since it is more representative of this type of structure. The examination of 2250 structures found



Scheme 1

Table 1. Properties of structures **I**–**III** and the 2D NMR data used in their identification[a]

No.	Molecular formula	n	DBE	n_s	^{13}C NMR	$^1H-^1H$ COSY	$^{13}C-^1H$ HMBC	$^{15}N-^1H$ HMBC	Ref.
I	$C_{27}H_{18}N_4$	31	21	14	+	–	39	–	37
II	$C_{34}H_{24}N_4O$	39	25	19	+	19	46	8	38
III	$C_{32}H_{22}N_4O$	37	24	19	–	12	32/45[b]	3	39

[a] Designations: n = the number of heavy atoms; DBE = the total number of double bonds and cycles (the double bond equivalent); n_S = the number of skeletal atoms in 'silent' fragments. The subsequent columns indicate the presence or absence of various types of NMR spectra and the number of peaks discovered in the corresponding spectra.

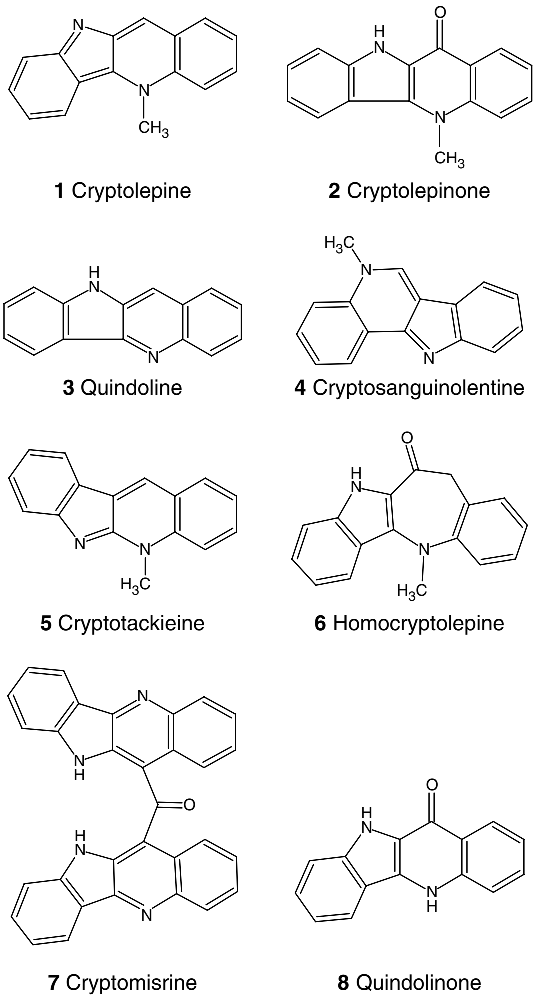
[b] 32 peaks were registered in the standard mode with the optimization for the coupling constant value set to 6Hz. 45 peaks were discovered under phase-sensitive conditions with a cryoprobe and optimized for 8 Hz.

in the system knowledge base and containing an amide group in a five-membered cycle confirmed that no such structures had a similar carbonyl chemical shift. In the IR spectrum of cryptospirolepine the frequency of the carbonyl absorption is also unusually low at 1611 cm^{-1}. These factors all contribute to a very challenging elucidation process for cryptospirolepine. The structure determination of **III**, a degradation product of cryptospirolepine, is also a complex analytical problem. In addition to fused benzene rings, the molecule contains a large 'silent' fragment containing 19 carbon atoms and two *N*-CH_3 groups. The ketone carbonyl resonance also has an uncommonly low chemical shift value of 167.1 ppm characteristic of esters and amides, though representative of pyridones and quinolones.

The StrucEluc system contains algorithms and programmatic features that allow users to create their own knowledge base. This knowledge base generally contains structures referring to chemical classes that are of particular interest to a given researcher as well as fragments excised from these structures using specific algorithms. As in the main system knowledge base, the carbon atoms of fragments and molecules are attributed the corresponding chemical shifts in the ^{13}C NMR subspectra. The system can therefore be adjusted to expedite the structural characterization of new members related to existing classes of compounds for which some data are available upon which specialized knowledge bases can be built. A search for fragments consistent with the ^{13}C NMR spectrum can be performed in each database separately or in any combination.

For this reason, we considered this approach for determining alkaloid structures **I**–**III** using a user database adjusted for cryptolepine structure analysis. Assuming that the unknown compounds were members of the cryptolepine series, we introduced the information for earlier published members of the series. To create a user database, we selected the compounds contained in the cryptolepine series shown in Fig. 2.

The spectral data referring to these compounds were obtained from the literature.[43–49] 2D NMR spectra for compounds **4**–**8** were found in the literature while others only provided 1D NMR spectra.



1 Cryptolepine **2** Cryptolepinone

3 Quindoline **4** Cryptosanguinolentine

5 Cryptotackieine **6** Homocryptolepine

7 Cryptomisrine **8** Quindolinone

Figure 2. Structures of the cryptolepine family employed for forming a user database.

Creating a user database

As shown in Fig. 2, all cryptolepines are highly unsaturated and can present difficulties regarding their structure elucidation. Initially ^{13}C NMR chemical shift data were input into the program and the StrucEluc library was searched. The search results indicated that molecules **1**–**5** are present in the system knowledge base. Since 2D NMR data for compounds **4**–**8** were available, these structures were used to challenge the system. First, we attempted to determine the structure in the common mode. These efforts were successful for all of the compounds to provide the results shown in Table 2. The structures for compounds **4**–**7** were elucidated in the automated mode without any operator interaction. To

Table 2. Results of structure elucidation of compounds **4**–**8**[a]

		Available data					
No.	Molecular formula	COSY	HBMC	*k*	*t*	r_A	Ref.
4	$C_{16}H_{12}N_2$	+	+	1	<1 s	1	46
5	$C_{16}H_{12}N_2$	+	+	1	<1 s	1	46
6	$C_{17}H_{14}N_2O_2$	−	+	24 → 3	20 s	1	47
7	$C_{31}H_{18}N_4O$	+	+	5 → 5	10 s	1	48
8	$C_{15}H_{10}N_2O$	−	+	26 → 5	4 s	1	49

[a] Designations: *k*—number of structures in the output file, the figures after the arrows show the number of non-identical structures; *t*—structure generation time elapsed; r_A—position of a correct structure in the ranked output file

elucidate the quindolinone structure **8**, the fragments found in the StrucEluc knowledge base were used. The generation time for all structures was in all cases less than 20 s.

Alkaloids **1**–**8** of the cryptolepine series along with the assigned ^{13}C NMR spectra were included in the user library of full structures. The algorithm used to create a user fragment library is similar to the algorithm used to create the StrucEluc system knowledge base and has been described elsewhere.[33] The two basic steps are as follows:

- The program excises as complete as possible a set of fragments from all the structures which are incorporated into the structural file. In so doing, the program is guided by a collection of rules providing generation of 'chemically reasonable' fragments.
- Atoms in the fragments are assigned chemical shift values that they have in the corresponding structure.

The procedure described above produced a user library containing 342 fragments from the cryptolepine series. This user database was then used to elucidate the remaining structures **I**–**III**.

The solution of challenging problems

Cryptolepicarboline

Searching the ^{13}C NMR spectrum of cryptolepicarboline through the user fragment library yielded 68 fragments. To create the MCDs the following options were specified: (a) minimal number of fragments to be included into each MCD is 3 ($q_{\mathrm{fr}} = 3$); (b) minimal percentage of the total number of skeletal atoms absorbed by the fragments of each MCD is set to 50 ($p_{\mathrm{at}} = 50\%$). These options are specified to minimize the number of free skeletal atoms in each MCD. Fifty MCDs were created for $E = 1.5$ ppm, and 25 of these survived the test for contradictions. One of the MCDs shown (Fig. 3) displays three fragments accounting for about 70% of the skeletal atoms.

The low E value indicates that among the 68 selected fragments there were many carbon atoms with assignments very close to the experimental values. Structure generation using a constraint for the ring size of $R_c = 5-7$ resulted in $k = 32 \rightarrow 13, t = 10$ s. NMR spectra prediction positioned the cryptolepicarboline molecule first in the ranked file, where $d_A(2) - d_A(1) = 3.3$ ppm.

Cryptospirolepine

Searching the ^{13}C NMR spectrum of cryptospirolepine in the user fragment library yielded 60 fragments. These produced 180 MCDs using a value of $E = 4$ ppm. Attempts to generate a structure failed since the calculations were too time consuming. Additional structural information was introduced using the 'generalized portrait' approach discussed earlier.[33] The full list of fragments was 'refined' and fragments corresponding to the nature of the structure analyzed were selected using program facilities. For a general representation, our so-called 'generalized portrait' of the cryptospirolepine structure, most functional groups in the selected fragments were identified.

The system fragment library was searched using the ^{13}C NMR spectrum of cryptospirolepine as input. A list of fragments was produced ($L = 1437$). All ion-containing structures were removed to provide 1287 fragments. The first 12 functional groups along with the number of parent fragments are shown in Scheme 2.

703 fragments (55%) contain a benzene ring and almost half of these (338) show a 1,2-Ar substitution. Owing to the atypically downfield chemical shift value of the carbonyl group (188.4 ppm), none of the found fragments contains a tertiary amide. The hypothesis was that the molecule should contain at least one 1,2-Ar fragment. The 1,2-Ar-containing fragments shown in Scheme 3 were automatically sorted out of the 60 fragments extracted from the *user database* in the first step.

A number of these fragments are identical. However, they have different chemical shift assignments relating to the original parent molecules that are contained within the user library. MCDs were generated from these fragments using the following parameter values: $q_{\mathrm{fr}} = 3$, $p_{\mathrm{at}} = 50\%$, $E = 6$ ppm. As a result, it took 3 min for the program to build $n_{(\mathrm{MCD})} = 490$ where the first 216 MCDs contained four fragments each. One of these is shown in Fig. 4 (only HMBC connectivities are visible).

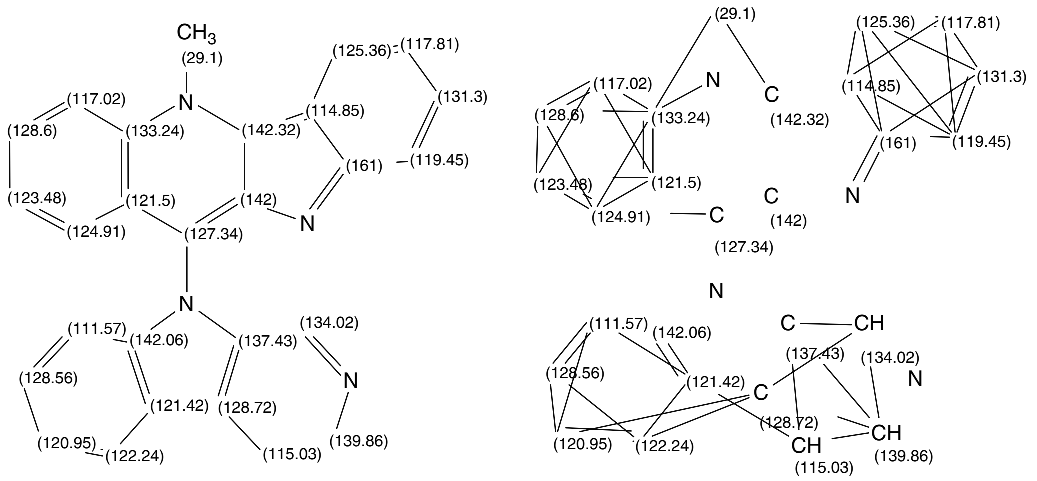
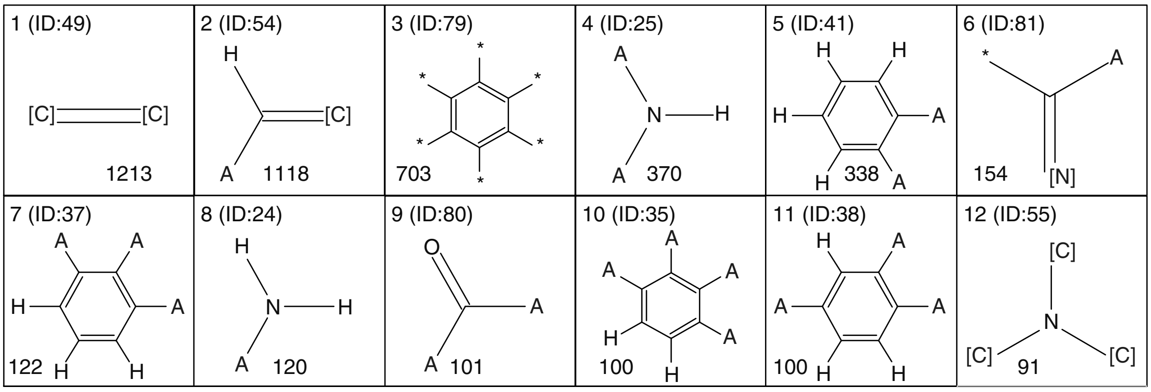
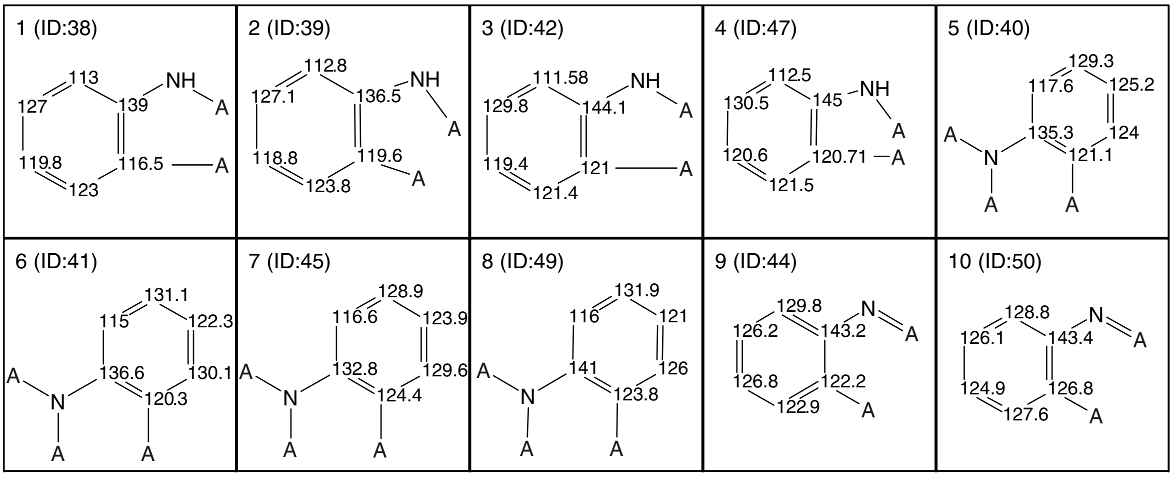


Figure 3. One of 50 molecular connectivity diagrams of cryptolepicarboline (right) in comparison with the molecular structure having assigned atoms (left). Unambiguous connectivities only are shown.



Scheme 2



Scheme 3

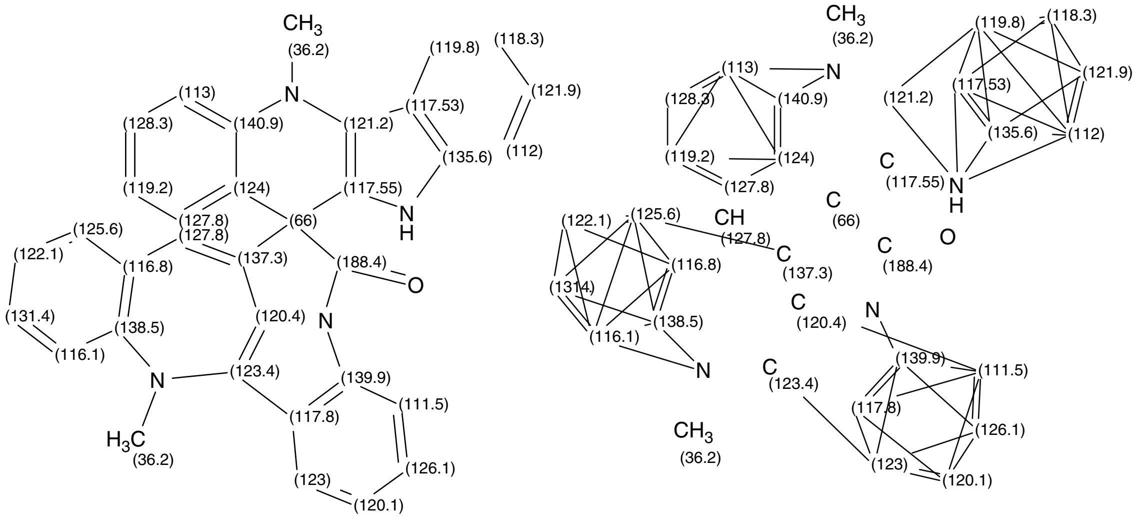


Figure 4. One of the first 216 molecular connectivity diagrams of cryptospirolepine (right) in comparison with the molecular structure having assigned atoms (left). Unambiguous connectivities only are shown.

In these MCDs, the fragments account for more than 70% of the skeletal atoms, suggesting that they are probably good fragments. It was proven by IR spectral experiments[38] that the analyzed molecule contains an amide group so the ketone functional group was added to the BADLIST to produce $k = 192 \rightarrow 1$, $t = 18$ min 30 s under the constraint of $R_c = 5$–7. The only structure consistent with the data was the structure of cryptospirolepine.

Compound III

For the elucidation of this structure, raw spectral data were processed and input into the program using the processing tools available in the ACD/Labs SpecManager software. A 1D ^{13}C NMR spectrum was not available, as is common in natural product structure elucidation when very small samples are involved. The ^{13}C shift inputs were thus created from the HSQC and HMBC spectra. Eighteen peaks were identified in the HSQC (2 CH_3 and 16 CH) data and 13 peaks were extracted from the HMBC to give a total of 31 peaks. According to the molecular formula, the molecule contains 32 carbon atoms. It was concluded that one quaternary carbon atom did not show an HMBC peak and one was added to the spectrum with a chemical shift of 130 ppm, in the middle of the aromatic interval. The number of peaks in the HMBC spectra acquired in standard and phase-sensitive mode were different, 32 and 45 peaks, respectively. These additional responses are probably due to improved resolution in the congested regions of the spectrum, although possibly higher order couplings were being detected. To avoid contradictions, the extra peaks observed in the second HMBC experiment were attributed to a range of potential couplings and allowed to be ^{2-4}J(C,H). Searching the ^{13}C NMR spectrum in the user fragment library resulted in 101 fragments. 3144 MCDs were created with $E = 2$ ppm and $q_{fr} = 4$ in about 25 min. Checking for contradictions reduced the number of MCDs to 1376. Structure generation was performed with all spectral filters off and a cycle size constraint of $R_c = 5$–7. A carbonyl group was added to the GOODLIST. The result of structure generation was $k = 785 \rightarrow 75$, $t = 30$ min. As the structures were ranked by the d_A values, the target structure was moved to first position. The first three structures of the ranked file are shown in Scheme 4.

The difference of deviations $d_A(2) - d_A(1) = 0.5$ ppm is small but the increase of both d_H and d_F structure to structure allowed us to conclude that the structure of compound **III** is correctly identified.

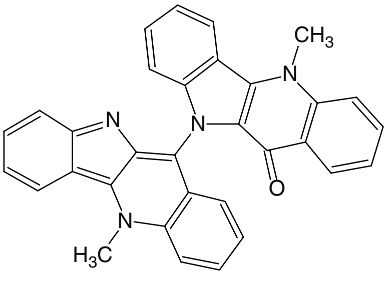
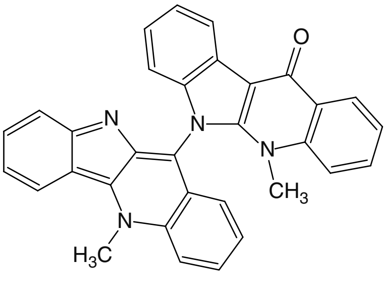
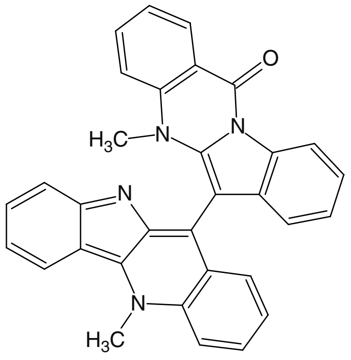
It is common for an experienced spectroscopist to detect molecular fragments simply by visual analysis of 1D and 2D NMR data. The approach is based on experience, knowledge and insight of a highly qualified researcher, and the structural information extracted can therefore be invaluable. Providing spectroscopists with software tools that can facilitate the assembly of the molecular structure in an interactive mode while allowing them to modify their hypotheses is of obvious value. This approach is expected to have a synergistic effect.

The ability of the StrucEluc system to act as an assistant to the elucidation process was tested for this example. Visual analysis of the molecular connectivity diagram of **III** allowed the expert to clearly see three 1,2-Ar fragments and suggest the presence of the fourth one. HMBC connectivities identify the connection of the two aromatic fragments via the *N*-CH_3 group and binding of another *N*-CH_3 group to the third benzene fragment. Several other bonds were drawn manually, in particular, atom C(167.1) was connected by a double bond to the oxygen atom. The resulting diagram is shown in Fig. 5.

The results of structure generation from this manually created MCD with the aforementioned options were $k = 550 \rightarrow 85$, $t = 1$ min. The correct structure was identified from the other structures since $d_A(2) - d_A(1) = 2$ ppm. The application of the spectroscopist's insight has a beneficial effect and allowed progress without the user fragment library, thus saving a considerable amount of time. The above example indicates that a highly qualified expert capable of determining very complex structures relying on his or her knowledge can result in dramatic increases in the speed of the elucidation process and improve the reliability of a system such as StrucEluc.

EXPERIMENTAL

The software used for the results reported here was ACD/Structure Elucidator, StrucEluc, version 6.0, a Windows-based software program composed of a number of separate modules. The entire software suite was composed of 1D and 2D NMR data processing, MS data processing, ^{1}H and ^{13}C NMR chemical shift prediction and assigned structure databasing tools, and an integrated chemical structure-drawing program. During the elucidation process the program displayed a number of possible structures, with comparison of on-screen experimental and fragment spectra or, in the case of failure, a set of structural fragments corresponding to portions of the spectrum that were used to assemble the structure of the unknown compound. StrucEluc included filters for ^{13}C and ^{1}H NMR, IR peaks, mass spectral (M_r) data, elemental composition and a self-training system. If the accuracy of spectral calculations for a new class of compounds was poor, user databases with experimental

1 (ID:15404)	2 (ID:15405)	3 (ID:15406)
		
$d_A(^{13}C)$: 3.371 (5.569) $d_F(^{13}C)$: 5.354 (8.260) $d_A(^{1}H)$: 0.271 (0.460)	$d_A(^{13}C)$: 3.879 (7.126) $d_F(^{13}C)$: 5.704 (9.375) $d_A(^{1}H)$: 0.352 (0.532)	$d_A(^{13}C)$: 4.765 (6.510) $d_F(^{13}C)$: 5.779 (8.384) $d_A(^{1}H)$: 0.474 (0.684)

Scheme 4

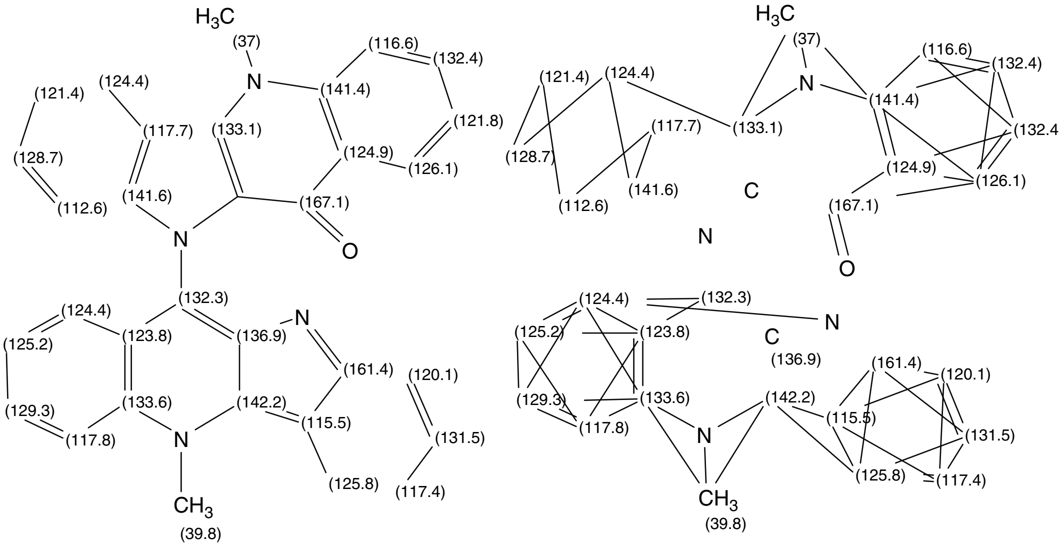


Figure 5. Molecular connectivity diagram of compound III (right) displaying fragments deduced by the expert in comparison with the molecular structure having assigned atoms (left). Unambiguous connectivities only are shown.

chemical shifts could be constructed and utilized during the elucidation process. Both ^{1}H and ^{13}C databases of assigned structures with NMR chemical shifts and coupling constants were available for searching by chemical structure and substructure. The number of entries in each database was >110 000 for ^{1}H NMR and >111 000 for ^{13}C NMR. Calculations were performed on a PC (Celeron operating at 500 MHz, Windows 98, 128 Mb RAM).

CONCLUSIONS

The work reported here shows that the StrucEluc expert system is a powerful tool to allow automated and assisted natural product structure elucidation using 1D and 2D NMR spectra. Its performance has been proven by the elucidation of over 100 new natural products. Depending on the quality and quantity of 2D NMR data, two main modes of structure determination are provided in the system: *common* and *fragment* modes. The common mode allows structure generation on the basis of a molecular connectivity diagram built up via the transformation of 2D NMR correlations into the connectivities between molecular skeletal atoms. In those cases where the structural information extracted directly from 2D NMR spectra is insufficient for the identification of the molecule within a reasonable processing time, the system can perform a search of the ^{13}C NMR spectra in a library containing about 1 000 000 fragments and their corresponding ^{13}C NMR subspectra. Frequently it is necessary to apply this technique for the analysis of natural products containing few hydrogen atoms. In most cases the selected fragments make up for the lack of structural information.

If this technique fails, the user can optimize the performance of the software around a particular class of compounds. For this purpose, the user can generate a fragment library using a set of assigned molecules related to that under analysis. The efficiency of the user fragment library has been shown by application to three natural products of the cryptolepine series that the program failed to determine both in the common mode and using the fragments from the system library.

In addition to library fragments, the StrucEluc system allows the introduction of fragments whose presence is proposed by the user. To mitigate the human factor in creating structural hypotheses, a 'generalized portrait' of the unknown molecule can be created by examining the distribution of the most common functional groups in fragments found in the system library. This helps the user to identify potentially beneficial fragments for the elucidation process. The flexibility in tailoring the structure elucidation process to the compound or family of compounds being investigated allows the expert to fully integrate his accumulated knowledge and insights in the problem solving process ultimately facilitating the elucidation of unknown structures that are extremely challenging to performing manually.

The StrucEluc algorithms and programs are constantly evolving and improving. In particular, we plan to introduce methods for analyzing stereochemical properties of a molecule and enhance the ability to extract high-quality inputs from the experimental data. Development is under way to allow the analysis and inclusion of data from the recently developed *J*-based experiments[50] and extraction of additional connectivity data by the manipulation and analysis of HSQC-TOCSY data.

REFERENCES

1. Martin GE, Crouch RC. *J. Nat. Prod.* 1991; **54**: 1.
2. Martin GE, Crouch RC. In *Modern Methods of Plant Analysis*, vol. 15, Linskens HF, Jackson JF (eds). Springer: Berlin, 1994; 25.
3. Martin GE, Zektzer AS. *Magn. Reson. Chem.* 1988; **26**: 631.
4. Müller L. *J. Am. Chem. Soc.* 1979; **101**: 4481.
5. Bodenhausen G, Ruben DJ. *Chem. Phys. Lett.* 1980; **69**: 185.
6. Bax A, Subramanian S. *J. Magn. Reson.* 1986; **67**: 565.
7. Bax A, Summers MF. *J. Am. Chem. Soc.* 1986; **108**: 2093.
8. Crouch RC, Martin GE. *J. Nat. Prod.* 1992; **55**: 1343.
9. Crouch RC, Martin GE. *Magn. Reson. Chem.* 1992; **30**: 66.

10. Martin GE, Guido JE, Robins RH, Sharaf MHM, Schiff P Jr, Tackie AN. *J. Nat. Prod.* 1998; **61**: 555.
11. Martin GE, Crouch RC, Zens AP. *Magn. Reson. Chem.* 1998; **36**: 551.
12. Martin GE, Hadden CE, Tackie AN, Sharaf MHM, Schiff PL Jr. *Magn. Reson. Chem.* 1999; **37**: 529.
13. Schlotterbeck G, Ross A, Hochstrasser R, Senn H, Kühn T, Marek D, Schett O. *Anal. Chem.* 2002; **74**: 4464.
14. Martin GE. In *Encyclopedia of Nuclear Magnetic Resonance*, vol. 9. *Advances in NMR*, Grant DM, Harris RK (eds). Wiley: Chichester, 2002; 98–112.
15. Russell DJ, Hadden CE, Martin GE, Gibson AA, Zens AP, Carolan JL. *J. Nat. Prod.* 2000; **63**: 1047.
16. Crouch RC, Llanos W, Mehr KG, Hadden CE, Russell DJ, Martin GE. *Magn. Reson. Chem.* 2001; **39**: 555.
17. Martin GE. In *Encyclopedia of Nuclear Magnetic Resonance*, vol. 9. *Advances in NMR*, Grant DM, Harris RK (eds). Wiley: Chichester, 2002; 33–35.
18. Elyashberg ME, Serov VV, Martirosian ER, Zlatina LA, Karasev YuZ, Koldashov VN, Yampolskiy YY. *J. Mol. Struct.* 1991; **230**: 191.
19. Elyashberg ME, Martirosian ER, Karasev YZ, Thiele H, Somberg H. *Anal. Chim. Acta* 1997; **337**: 265.
20. Munk ME, Velu VK, Madison MS, Robb EW, Badertscher M, Christie BD, Razinger M. In *Recent Advances in Chemical Information II*, Collier H (ed). Royal Society of Chemistry: Cambridge, 1992; 247.
21. Christie BD, Munk ME. *J. Am. Chem. Soc.* 1991; **113**: 3750.
22. Funatsu K, Sasaki S. *J. Chem. Inf. Comput. Sci.* 1996; **36**: 190.
23. Will M, Fachinger W, Richert JR. *J. Chem. Inf. Comput. Sci.* 1996; **36**: 221.
24. Peng C, Yuan S, Zheng C, Hui Y. *J. Chem. Inf. Comput. Sci.* 1994; **34**: 805.
25. Yuan S, Peng C, Zheng C. *NMR-SAMS* (originally known as *CISOC-SES*). Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences: Shanghai, 1988–94; http://guru.specres.com/nmrsams.html.
26. Lindel T, Junker J, Koeck M. *Eur. J. Org. Chem.* 1999; **3**: 579.
27. Köck M, Junker J, Maier W, Will M, Lindel T. *Eur. J. Org. Chem.* 1999; **3**: 573.
28. Nuzillard J-M, Massiot G. *Tetrahedron* 1991; **47**: 3655.
29. Steinbeck C. *Angew. Chem., Int. Ed. Engl.* 1996; **35**: 1984.
30. Steinbeck C. *J. Chem. Inf. Comput. Sci.* 2001; **41**: 1500.
31. Elyashberg ME, Blinov KA, Martirosian ER, Molodtsov SG, Williams AJ, Martin GE. *J.Pharm. Biomed. Anal.* submitted.
32. Elyashberg ME, Blinov KA, Williams AJ, Molodtsov SG, Martirosian ER. *J. Nat. Prod.* 2002; **65**: 693.
33. Elyashberg ME, Blinov KA, Martirosian ER. *Autom. Inf. Manage.* 1999; **34**: 15.
34. Blinov KA, Elyashberg ME, Molodtsov SG, Williams AJ, Martirosian ER. *Fresenius' J. Anal. Chem.* 2001; **369**: 709.
35. Martin GE, Hadden C, Crouch RC, Krishnamurthy VV. *Magn. Reson. Chem.* 1999; **37**: 517.
36. Martin GE. *Annu. Rep. NMR Spectrosc.* 2002; **46**: 37.
37. Sharaf MHM, Shiff PL Jr, Tackie AN, Phoebe CH Jr, Howard L, Meyers C, Hadden CE, Wrenn SK, Davis AO, Andrews CW, Minick D, Johnson RL, Shockcor JP, Crouch RC, Martin GE. *Magn. Reson. Chem.* 1995; **33**: 767.
38. Tackie AN, Boye GL, Sharaf MHM, Schiff PL, Crouch RC, Spitzer TD, Johnson RL, Dunn J, Minick D, Martin GE. *J. Nat. Prod.* 1993; **54**: 653.
39. Martin GE, Hadden C, Russell D, Kaluzny B, Guido J, Duholke W, Stiemsma B, Thamann T, Crouch R, Blinov K, Elyashberg M, Martirosian E, Molodtsov S, Williams AJ, Schiff PL Jr. *J. Heterocycl. Chem.* 2002; **39**: 1249.
40. Bremser W. *Anal. Chim. Acta* 1978; **103**: 355.
41. Capon RJ, Ford J, Lacey E, Gill JH, Heiland K, Friedel T. *J. Nat. Prod.* 2002; **65**: 358.
42. Jayasuriya H, Lingham RB, Graham P, Quamina D, Herranz L, Genilloud O, Gagliardi M, Danzeisen R, Tomassini JE, Zink DL, Guan Z, Singh SB. *J. Nat. Prod.* 2002; **63**: 1091.
43. Ablordeppey SY, Hufford CD, Borne RF, Dwumu-Badu D. *Planta* 1990; **56**: 416.
44. Hadden CE, Duholkc WK, Guido JE, Robins RH, Martin GE, Sharaf MHM, Schiff Jr PL. *J. Heterocycl. Chem.* 1999; **36**: 525.
45. Spitzer TD, Crouch RC, Martin GE, Sharaf MHM, Schiff PL Jr, Tackie AN, Boye GL. *J. Heterocycl. Chem.* 1991; **28**: 2065.
46. Sharaf MHM, Schiff PJ Jr, Martin GE, Phoebe CH Jr, Tackie AN. *J. Heterocycl. Chem.* 1996; **33**: 239.
47. Sharaf MHM, Schiff PJ Jr, Crouch RC, Davies A, Andrews CW, Martin GE, Phoebe CH Jr, Tackie AN. *J. Heterocycl. Chem.* 1995; **32**: 1631.
48. Sharaf MHM, Schiff PJ Jr, Minick D, Johnson RL, Crouch RC, Andrews CW, Martin GE, Phoebe CH Jr, Tackie AN. *J. Heterocycl. Chem.* 1996; **33**: 789.
49. Crouch RC, Davies A, Spitzer TD, Martin GE, Phoebe CH Jr, Sharaf MHM, Schiff PJ Jr, Tackie AN. *J. Heterocycl. Chem.* 1995; **32**: 1077.
50. Marquez BL, Gerwick WH, Williamson RT. *Magn. Reson. Chem.* 2001; **39**: 499.