Automated Structure Elucidation – the Benefits of a Symbiotic Relationship between the Spectroscopist and the Expert System

Mikhail E. Elyashberg, Kirill A. Blinov, Eduard R. Martirosian

Advanced Chemistry Development
6 Akademik Bakulev St, Moscow 117513, Russian Federation

Sergey G. Molodtsov

Novosibirsk Institute of Organic Chemistry
Siberian Branch of Russian Academy of Science
Lavrentiev Avenue 9, Novosibirsk 630090, Russian Federation

Antony J. Williams

Advanced Chemistry Development Inc.,
90 Adelaide Street West, Suite 600, Toronto, ON, M5H 3V9 Canada

Gary E. Martin

Michigan Pharmaceutical Sciences Structure Elucidation Group
Pfzier Global Research & Development
Pfizer Corp., Kalamazoo, Michigan 49001-0199, USA
Received May 16, 2003

Characteristic features of a new expert system *StrucEluc* are described. The system is intended for the structure elucidation of complex organic molecules using a variety of spectroscopic data including 2D NMR. We review here the results of challenging this system with over 100 structure elucidation problems where the 2D NMR peak tables presented in original journal publications provided the input data. This contribution is focused on methods to overcome difficult situations that can arise when contradictions are present in the input data and/or when the structure is underdetermined as a result of insufficient 2D NMR correlations. Methods by which to address these situations are examined. It has been shown that synergy between the spectroscopist and the expert system allows the solution of problems that seemed to be hopeless at the outset of the structure elucidation process.

J. Heterocyclic Chem., **40**, 1017 (2003).

Introduction.

Advances in both hardware for data acquisition and software for data analysis have enabled structure elucidation. Nevertheless the extraction of a chemical structure from a collection of analytical data still remains a challenge for analytical laboratories. Laboratories in the chemical and pharmaceutical industry commonly isolate a large number of compounds in any given year and many of these can be regarded to be complex. To both simplify and speed up the analysis process and hence the determination of the structure of interest, expert systems have been created that mainly use NMR spectral data as their foundation. A series of reports have been published in which expert systems developed to aid the elucidation process are described (for instance, [1-10]). In these systems 2D NMR data, presented in the form of connectivities between skeletal atoms of the molecule, serve as restrictions for the structure generation process which proceeds from a given molecular formula. A typical input data set generally is comprised of both homonuclear ^{1}H-^{1}H COSY and heteronuclear direct (HMQC or HSQC), and long-range (HMBC or any of a number of more recently developed experiments [11,12]) connectivities. Recently ^{15}N-^{1}H HMBC correlations (see reviews [13,14]) and a series of new 2D NMR techniques are becoming more widely used. After the structure generation process produces a series of hypothetical structures consistent with the atom-to-atom connectivity (homo- and heteronuclear; direct, long-range, and through space) information fed to the program, the most likely structures can be identified on the basis of a comparison of the predicted ^{13}C chemical shifts of the candidate structures *vs*. the observed ^{13}C chemical shifts for the molecule.

Our work in regards to the development of expert systems has shown that the best result is achieved when software allows synergistic interaction between the skills and insights of a spectroscopist and the unbiased nature of a computer program. This means that qualified spectroscopists should be given the freedom to apply their experience and knowledge regarding the elucidation of a compound under study in order to introduce additional constraints for the structures generated by the expert system. The implementation of such possibilities allows a symbiotic relationship between the scientist and a computer, a synergistic effect that is highly beneficial.

According to previous publications, expert systems [1-10] have only allowed restricted application of *a priori*

information. In particular this can prevent the introduction of key fragments that can directly contribute to the structural hypothesis. Similarly, little attention has been paid to the detection of contradictions in 2D NMR data and methods by which these can be resolved. It is known [15,17] that the source of contradictions is, for example, the presence of cross peaks in COSY or HMBC spectra that correspond to couplings over four or more bonds. Commonly computer software applications are defaulted to correlations over fewer bonds.

A number of different ways of eliminating the contradictions have been proposed. In particular, searching for contradictions by repeating the process of structure generation and adding at each iterative step one correlation related to a *weak* 2D NMR peak has been utilized during this work. The authors of ref. [10] also apply this approach. If a particular correlation turns out to be of greater length than that set as a default then the structure generator will produce no structures after this correlation has been added and this indicates the presence of contradictions. However, as will be shown in the current work, the presence of contradictions may not prevent structure generation, and false structures may be generated. In order to overcome this difficulty the application of a stochastic generation algorithm that requires the application of computer calculations on parallel processors has been proposed [8]. Our experience shows that the number of connectivities characterized by spin-spin couplings over four bonds or more can be over ten couplings in one set of 2D NMR data [10]. This makes the resolution of these contradictions using the methods discussed problematic. Our experience has also shown that frequently important information regarding the structure of an unknown substance can be derived from a sub-structure database and related ^{13}C NMR sub-spectra when used in combination with 2D NMR data.

The drawbacks highlighted in the preceding discussion have mostly been overcome in the *StrucEluc* system [15-17]. During development, the capabilities of the *StrucEluc* program package have been challenged and evaluated by using the published 2D NMR data for more than one hundred natural products whose structures were elucidated. Data were fed to the system directly from the published 2D NMR peak tables. In the process, we have demonstrated that the system is very capable of automated structure elucidation. The work described in this report illustrates strategies for determining the structures of natural products in a number of challenging situations, when the solution of the problem by the "*common*" operation mode of *StrucEluc* fails. For several of examples shown, the interaction of a qualified spectroscopist with the software system, which incorporates a capable knowledge base and diverse means of correct structure identification, leads to the successful determination of structures that initially seemed doomed to failure.

Materials and Methods.

The *StrucEluc* system has been described in detail previously [15-17]. Here we shall provide only a brief overview of its unique capabilities. Relative to other systems designed to elucidate structures based on 2D NMR spectral inputs, StrucEluc is equipped with a knowledge base (KB) and three structure generators based on different mathematical algorithms.

The KB consists of three components: 1) a library of about 200,000 molecular structures and their assigned ^{13}C NMR spectra; 2) a *fragment library* (FL) containing about 1,000,000 fragments with corresponding ^{13}C NMR sub-spectral assignments created using proprietary algorithms from the full structures stored in the KB; 3) a *Library of Spectrum-to-Structure Correlations* (LSC) comprising the most common functional groups and their characteristics in both NMR and IR spectra.

The *StrucEluc* system is able to use both the 2D NMR data and the fragment library during the elucidation process. In those cases where the number of available 2D NMR correlations is insufficient to impose effective restrictions during the structure generation process (in this case the number of possible structures can be extremely large and the generation time will be unacceptable), the system searches for appropriate fragments in the library in accordance with their associated sub-spectra. Found fragments (FF) meeting the restrictions arising from the 2D NMR spectra, are retained. Acceptable combinations of good fragments are "projected" onto the set of all atoms within the molecular formula. As a result, the program builds and displays one or more molecular connectivity diagrams (MCD), on which fragments, atoms, and connectivities of different lengths are graphically represented. Using a correlation table, the program automatically establishes the carbon atom properties, these being the atom hybridizations and proximity to heteroatoms (the number and type of heteroatoms are specified). At this stage a qualified specialist is given an opportunity to analyze the MCD and make appropriate revisions including specifying particular atom properties, change the lengths of certain connectivities, draw specific chemical bonds, for example explicitly designating C=O, -C≡N, *etc*.). Subsequently process chemists could also choose to introduce fragments that in their opinion should be present in a molecule (so-called user fragments, UF). In this case the program is adjusted so that both the users and found fragments are used for creation of the next generation of MCDs.

Chemists frequently try to use the assigned spectra of structures related to the molecule being studied to aid in the structural determination and assignment of the NMR spectra of new compounds. In many cases this approach can be very successful. In order to implement this method within the *StrucEluc* system, algorithms enabling auto-

mated generation of a user fragment library have been incorporated. To create the user database the chemist populates a database with structures having assigned ^{13}C NMR spectra. Using proprietary algorithms, fragments are then extracted [16] from these structures. In the situation when an extensive user database exists comprised of molecules similar to that being studied, the probability that the database will contain meaningful fragments to a new analogue in a series is rather high. If necessary, the program is also able to sequentially "assemble" the molecule of the unknown compound from the fragments found in the KB by overlapping common atoms (the so-called *standard* [16] generator is used in this case). Another possible approach is in forming sets from the found fragments. The sets are then used for structure generation using an input molecular formula (the so-called *classic* generator [16] is used). All structures produced by any of the available generators can be verified by the LSC, which forms the NMR and IR filters (if IR spectra are available).

An analysis to check the consistency between the 2D NMR data and the generated structures is performed using a heuristic algorithm described elsewhere [17]. In most cases, approximately 90 % of the time, this algorithm can detect the presence of contradictions and suggest the reasons for their presence. If contradictions are found, the program tries to automatically eliminate the contradictions by lengthening those particular connectivities that according to the analysis results may have a length that is contradictory to that set as the default. The preferable structure is selected on the basis of both ^{13}C and ^{1}H NMR chemical shift prediction. ^{13}C NMR spectra are calculated using two methods. The first is the so-called *fast*-method based on additive rules, and the second version referred to as the "*accurate* method" using a HOSE-code fragment-based approach. In each case the average values of deviation of the calculated *vs.* experimental spectrum (d_F and d_A. respectively) are estimated and the structures are ordered by increasing d_F or d_A values. The values of d_F and d_A are calculated as a simple average of the chemical shift deviation per carbon atom. To select the most probable structure initially the fast spectrum prediction is carried out and the structures are ranked in ascending order of d_F values. The accurate spectrum calculation is then applied to the first 10-20 members of the ranked file, and the structures are again ranked in ascending order of d_A values. The first structure of the finally ranked file is considered as the most probable one.

The system is also integrated with a suite of software programs that can predict mass spectrometric fragmentation and many physicochemical parameters (for example logP, solubility, etc.) that can be used as supporting verification parameters if experimental values have been determined. A structure similarity coefficient [18] can also be calculated enabling structures to be ranked by similarity. Such ranking is useful when analyzing elucidated structures as well as for producing a user database and a user fragment library. To calculate a similarity coefficient, the user can choose one of five different methods incorporated into the system.

Previously, in order to characterize the performance of *StrucEluc* for the automated structure determination of organic molecules and natural products, we have used 2D NMR data published in the literature. The ongoing enhancement of the StrucEluc system have been driven by continued application of the system to real problems and development and tweaking of necessary algorithms to resolve problems. To date *StrucEluc* has solved the structures of about 150 natural products based on 2D NMR data published by different researchers primarily in the *Journal of Natural Products* during 2000-2003, as well as using raw data obtained in different laboratories. The structures of 60 compounds elucidated previously using *StrucEluc* are discussed in ref [17]. The molecules analyzed in this investigation were reasonably large and fairly complex, ranging in mass from 200 to 900 a.m.u. and between 15 and 65 skeletal atoms.

The results are reviewed here as a demonstration of system capability. The robust nature of the *StrucEluc* structure evaluation strategy was supported by the fact that for 90% of the elucidation challenges the d_A-ranking placed the correct structure at the top of the list. The high selectivity of the system was proved by the fact that in 75 % of the cases the output lists contained ten or fewer structures. Our research has shown that even preliminary ranking of the output structures by the d_F statistic placed the correct structure at the top of the list in 80 % of the test cases. Considering the present speed and cost of modern personal computers, the issue of computer processing time required for structure determination of a new natural product is in many cases inconsequential. While the extraction and purification of a compound can take many months, a matter of minutes, hours, or even a few days to obtain a structural solution using a personal computer should not be considered excessive. Nevertheless, for 75% of the test cases, the solution was found in under one minute and for 95 % of the cases the solution time did not exceed 30 min (*Celeron* operating at 500 MHz, Windows 98, RAM 128 Mb). An analysis of the computation time with respect to the available data showed that almost all tasks requiring more than 10 minutes for solution were solved on the basis of HMBC correlations alone.

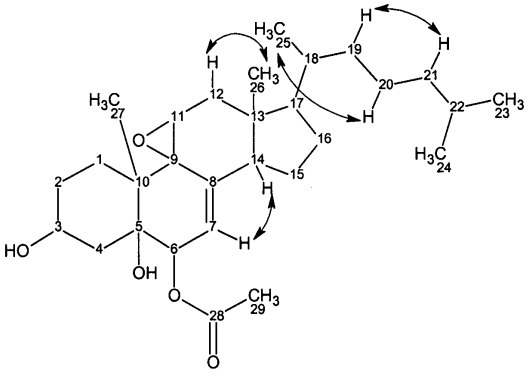
Recently we repeated the same data processing originally applied to a set of 130 problems. On this occasion the number of skeletal atoms in a molecule was between 20 and 90, while the molecular weight reached 1260. The investigation indicated the same characteristics in terms of the performance of the software, which is evidence of the statistical significance of the calculated values.

The strategy of the *StrucEluc* software application has been described in detail elsewhere [15,17] and illustrated by a series of examples highlighting the advantages of allowing interaction between the user and the software program. A common feature in the structural problems used as examples in the previous reports was that all contained non-contradictory 2D NMR data. Further, the molecules used as case studies in the previous reports, obviously, cannot encompass even a small fraction of the tremendously diverse range of problems encountered during the structure elucidation process. Providing even more stringent challenges to evaluate and further refine the capabilities of the *StrucEluc* program package requires the consideration of problem that both necessitate user interaction with the program as well as problems that contain incomplete 2D NMR data sets and/or those with hidden contradictions, *e.g.* substituents whose chemical shifts are outside of the normal interval ranges contained in filter libraries, *etc*. Specific examples of each of the more challenging types of problems were encountered while solving a large number problems using *StrucEluc*. Experimental NMR data reported by the authors of publications describing the identification and confirmation of the structure of new natural products were used as input for *StrucEluc* in these efforts.

Results and Discussion.

In the work of Leone *et al*. [19] five compounds related to a class of *polyoxygenated sterols* were identified and examined. For three of these compounds the authors presented 1D ^{1}H (600 MHz) and ^{13}C (150 MHz) NMR data as well as HMBC and COSY correlations in the form of peak tables. Automated structure elucidation for two of the three natural products was performed without difficulty, since related data were rather complete and coherent. The third compound (**1**) was more challenging and the analysis is presented here in detail.

The given molecular formula was $C_{29}H_{46}O_5$ and Mw=474. The molecular formula and peak lists for the HMQC, COSY and HMBC spectra were fed as inputs into the program. These formed tables containing 43 COSY cross-peaks and 57 HMBC correlations. The default settings of the program are such that COSY and HMBC crosspeaks are defined to result from coupling interactions within 2-3 bonds. However, the comparison of crosspeaks with the molecular structure, a process which can in reality only be performed after the structure of the unknown compound is determined, showed that there are four long-range correlations in the COSY spectrum corresponding to the following couplings: $^{5}J_{HH}$ (H20-H25), $^{4}J_{HH}$ (H19-H21), $^{4}J_{HH}$ (H12-H26), and $^{4}J_{HH}$ (H7-H14). It is evident that these correlations contradict the default options and consequently the correct structure will not be generated. 2D NMR cross-peak intensity information can be useful, though not definitive since the correlation response intensity is a function of both the size of the coupling, and the experimental optimization, in helping to define which correlations are "non-standard". For convenience we shall use this term to denote spin couplings which are present over distances longer than those set as defaults in the system options. Intensity information was however omitted in the peak table presented in the reference [19].



(**1**)

The MCD automatically formed by the program on the basis of the peak tables is shown in Figure 1. HMBC and COSY connectivities are presented here in the form of fragments connecting carbon atoms. The connectivities corresponding to COSY correlations that do not coincide with HMBC correlations are marked with bold lines. For convenience, connectivities having a length of one C-C bond, typical for a COSY correlation, will be called α-connectivities and those corresponding to a length of one to two C-C bonds, typical for an HMBC correlation, will be referred to as β-connectivities. Correlations spreading over one to three C-C bonds will be referred to as γ-connectivities. For the example shown in Figure 1 the methylene group associated with C19 has three α-connectivities thereby contradicting the valence for the carbon atom.

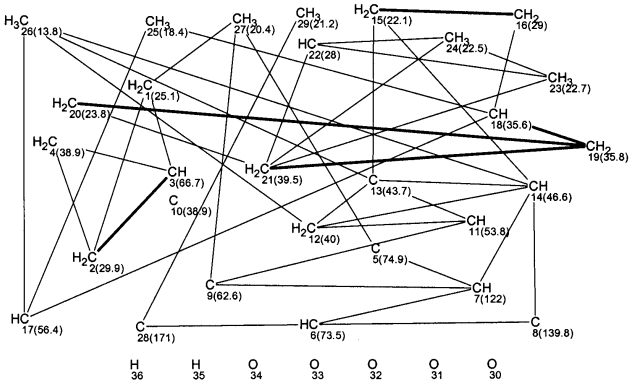


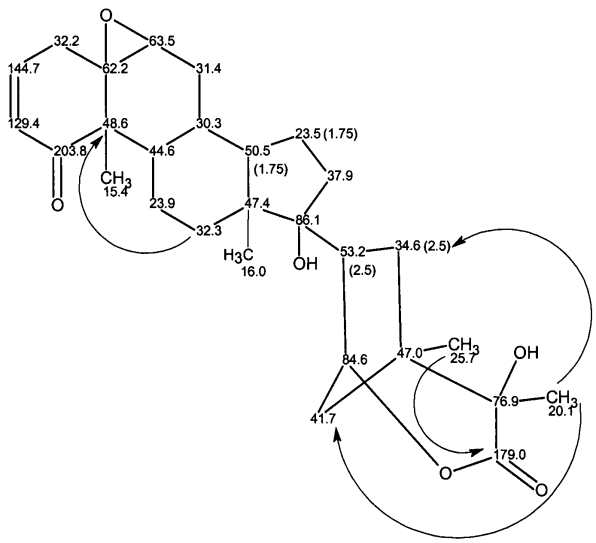
Figure 1. The molecular connectivity diagram created from the HMBC and COSY spectra of compound **1**. COSY-connectivities that do not coincide with HMBC correlations are marked by the **bold** lines.

Visually it is obvious that the COSY data have at least one contradiction.

As mentioned above, the system provides for the automated searching and resolution of such contradictions. When the MCD was checked for contradictions some α-correlations were suggested to contradict the standard values. The algorithms within *StrucEluc* will attempt to resolve contradictions in an iterative mode. The procedure required four iterations and fourteen seconds to resolve contradictions. A "contradiction resolution protocol" has been included in the *StrucEluc* system to track the identity of contradictions. In this case the protocol indicated that the program detected and lengthened connectivities in the following order: (C20-C25), (C19-C21), (C12-C26), and (C7-C14). The protocol also summarized the steps producing a complete list of the new lengths of connectivities introduced to resolve the contradictions.

Based on the MCD with the corrected connectivity lengths, the generation and filtration of structures using spectral libraries and a structural BADLIST containing fragments that are highly unlikely in organic chemistry, and enhanced by the implementation of Bredt's rule was performed. No changes were made to the default atom properties (hybridization, possible attachment to heteroatoms, *etc*). The program generated 7 molecules (k=7) in 20 seconds (t_g=20 s). To determine the most preferred chemical structure the ^{13}C NMR chemical shifts of the structures in the output file were calculated using both the fast and accurate methods described earlier. In addition, the ^{1}H chemical shifts were calculated. The structures arranged in order of increasing average deviation of the calculated spectra, $\boldsymbol{d_A}$ (^{13}C) are shown in Figure 2. Structure **1** was listed as the first entry of the prioritized output file. This is denoted as $\mathbf{r_A}$=1. The value of the deviation $\boldsymbol{d_A}$ (1) =1.8 ppm and was significantly different from the value of that for the second structure ($\boldsymbol{d_A}$ (2) - $\boldsymbol{d_A}$ (1) =1.1 ppm). In our experience [17] a significant difference in $\boldsymbol{d_A}$ between the first and second structure is generally a good indication of the validity of the solution. It is interesting to note that only the first structure contains an epoxy group. In this case the program successfully dealt with identifying and resolving all contradictions and elucidating the correct structure. A chemist was able to detect visually only one contradiction in the MCD. Unfortunately, the program is not always able to resolve such issues because the algorithm for searching and resolving contradictions is based on heuristic principles. This algorithm is a critical component of the *StrucEluc* system and iterative improvement continues. The following example demonstrates a situation in which the program failed to identify all contradictions in the 2D NMR data.

The work of Habtemariam *et al* [20] reported the structure of a novel steroid (**2**) isolated and identified as *17-epi-acnistin-A* on the basis of both spectroscopic and X-ray diffraction data. In order to aid the discussion here the figure includes all chemical shifts assigned to the carbon atoms.



(**2**)

The high-resolution electron impact mass spectrum of **2** indicated a molecular ion at *m/z* 470, giving the molecular formula $C_{28}H_{38}O_6$. ^{1}H, ^{13}C, ^{1}H-^{1}H COSY, ^{1}H-^{1}H NOESY, HMQC and HMBC (optimized for 7 Hz long-range coupling) NMR spectra were recorded using a Bruker AMX-400 spectrometer. However, only the ^{1}H, ^{13}C and HMBC data were listed in the article without any indication of the observation of HMBC correlations of non-standard lengths ($> {}^{3}J_{CH}$). Only this information was used to determine the structure using *StrucEluc*. It was noted that in addition to non-standard correlations in the HMBC data, two proton shifts, 1.75 and 2.5 ppm, were both associated with two ^{13}C resonances each: 1.75/50.5, 1.75/23.5; and 2.5/53.2, 2.5/34.6 ppm. (see structure **2**). These correlations increased ambiguity in the 2D NMR data. Comparison of the HMBC correlations with the chemical structure (**2**) allowed identification of the presence of four γ-correlations of non-standard length as indicated with arrows in the structure (**2**).

The 2D NMR data were converted to connectivities to create the molecular connectivity diagram. The MCD showed that CH_2 (23.5) and CH (53.2) groups had no correlations with other carbon atoms. According to the original article, the carbon atom with δ 23.5 ppm in fact does not correlate with any other atom, while the CH δ 53.2 ppm was deprived of connectivities by the program due to accidental degeneracy of the signals in the ^{1}H NMR spectrum as mentioned above. The presence of two non-correlating carbon atoms in the MCD usually results in a long structure generation time and the number of potential structures will correspondingly be large even in the case when the 2D NMR data are consistent.

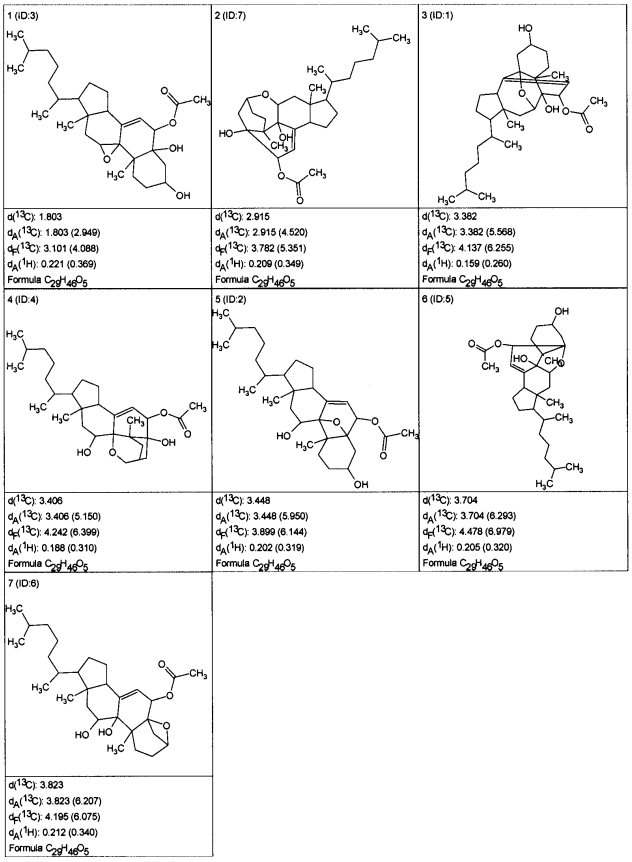


Figure 2. The results of the structure elucidation of compound **1**.

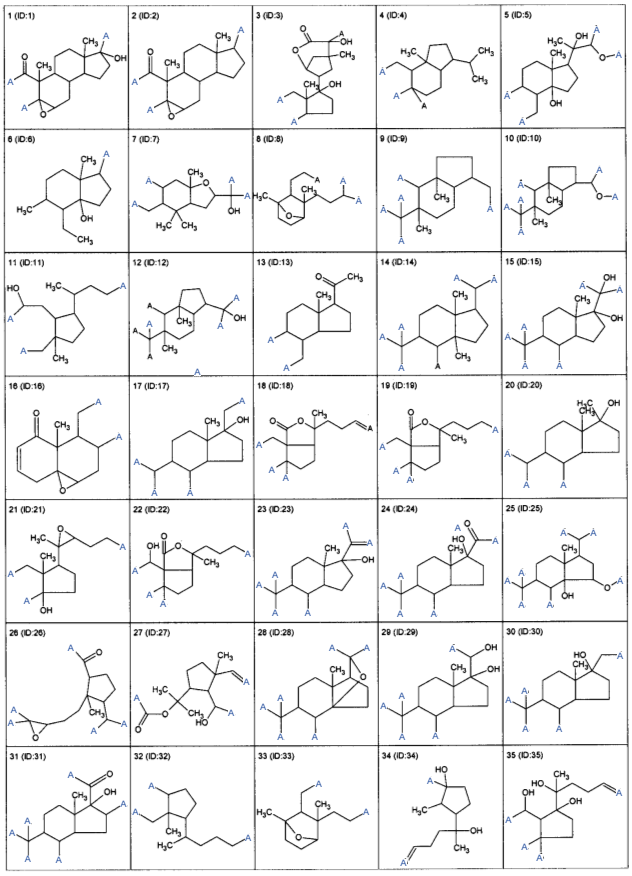


Figure 3. The initial structures in the found fragment list obtained as a result of a fragment search using the ^{13}C NMR spectrum of compound **2** as input.

The search for contradictions in the 2D NMR data produced no results. This indicated that the program found one or more arrangements of atoms under which the corresponding topological distances between carbon atoms, represented by their chemical shifts, are the same as the default values. In other words, the program found that it was possible to build at least one structure meeting all of the constraints and the atom properties assigned by the program with the help of the correlation table. With the absence of contradictions the structure generation procedure was initiated with no additional restrictions. As expected, the structure generation process was rather long. The program had produced no resulting chemical structures within 45 minutes and the generation process was aborted. Following the general strategy of problem solving using the *StrucEluc* system that we described previously [17], we attempted to use fragments stored in the system database.

The program selected L=1846 library fragments from the database that were consistent with the molecular formula of the analyzed compound taking into account both the chemical shifts and the multiplicities in the 1D ^{13}C NMR spectrum. The first 35 records from the resulting file, with the selected fragments ranked in decreasing order of their molecular formula are shown in Figure 3. Among the selected fragments there are a number that are present in the "unknown" structure. The high probability that there will be fragments consistent with a truly unknown structure is critically important if the structure is to be successfully solved by *StrucEluc*. Obviously, however, when dealing with a true unknown, there is also no way to know if in fact any of the fragments are really consistent with the structure of the molecule. The MCD creation from all of the found fragments was started using an initial value of E=0.5 ppm where E is the allowed deviation of the chemical shifts of carbon atoms present in the fragments from the values of the experimental ^{13}C NMR spectra. With E = 0.5 ppm, no MCDs were created. When this situation is encountered, the value of E used in the MCD creation process is usually increased incrementally. With a value of E = 2 ppm the program created one molecular connectivity diagram (see Figure 4). It can be seen that the fragment of the MCD denoted by bold lines corresponds to a portion of the published structure of the molecule **2**. Two oxygen atoms from the original molecular formula were incorporated but the carbons resonating at δ 23.5 and 53.2 were not present in the fragment structure. The check for contra-

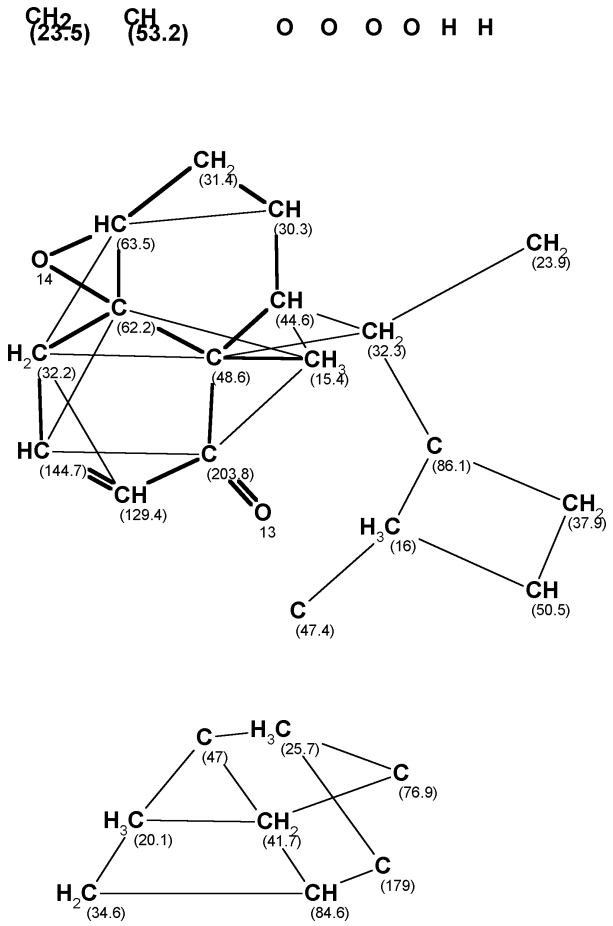


Figure 4. The molecular connectivity diagram created from the HMBC spectrum of compound **2** using found fragments.

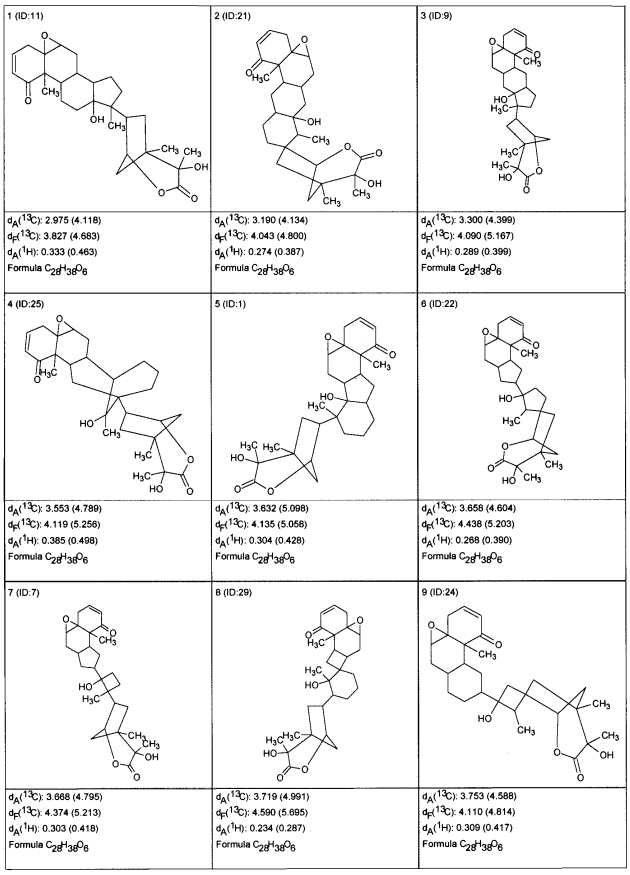


Figure 5. The initial structures in the output structural file produced from the MCD shown in Figure 4.

dictions was rerun with the result that this MCD passed all algorithmic tests and the program again was unable to recognize any contradictions.

In reality the user would not be aware of the presence of contradictions when working with a true unknown structure so the common process of starting structure generation from the MCD was followed to mimic the actions of an investigator with no prior knowledge to work from. Due to the presence of two free carbon atoms, and even with the presence of a large fragment which had consumed 13 skeletal atoms, the generation time was relatively long: t_g = 5 min 40 sec with a resulting structure set of k=36. After removing duplicates, 15 structures were stored (k=36→15). The first 9 structures from the file were arranged by increasing deviation of the calculated ^{13}C NMR spectra ($\boldsymbol{d_A}$) from the experimental spectra as shown in Figure 5.

A value of $\boldsymbol{d_A}$ = 2.97 ppm was determined for the preferred structure within the limits of allowed values [17] and, consequently the identified solution was not rejected. Comparison of the favored structure with structure **2** shows that they differ from each other only by the arrangement of both the OH and CH_3 groups attached to the five-membered ring, confirming that the structural formula of the unknown compound obtained in the presence of contradictions, was very similar to the actual structure. The calculation of the Tanimoto similarity match factor [18] for the suggested structure with the real structure gave a value of 0.975. In the presence of contradictions in the data *StrucEluc* was able to provide a reasonable solution to the problem. Understanding why the program failed to identify the non-standard connectivities can be explained by comparing the assignments of the chemical shifts in the real and preferred chemical structures (see Figure 6). As shown in Figure 6, all connectivity lengths in the preferred structure match the default values. The values of the carbon chemical shifts and their properties assigned by the software are within the range of intervals contained within the correlation table.

To ascertain the correctness of a particular solution, a check of its convergence may be of value. In order to check convergence for this particular example the fragments found in the database were used as a basis for

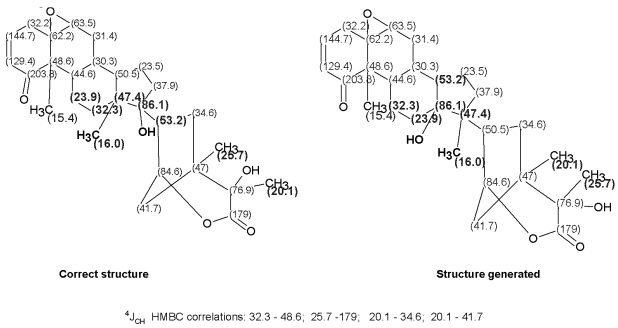
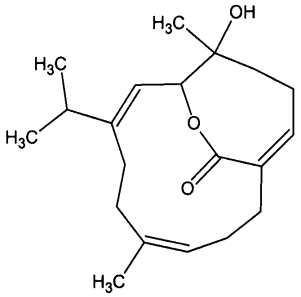


Figure 6. The comparison of assignments based on chemical shifts in the published (**2**) and *StrucEluc* elucidated structures.

standard structure generation by assembling structures from fragments having common atoms. It should be noted that the ^{13}C NMR spectrum and the structure of the analyzed molecule were not present in the system's knowledgebase.

Previous experience indicates that if the number of fragments is large, in the experience of the authors about 2000-5000 fragments or more, it is advisable to use only the large fragments for the first attempt at structure generation. This allows a quick evaluation of the possibility of obtaining a solution within a reasonable time. The first 100 fragments from the list were selected and *standard* structure generation was initiated on the basis of the 1D ^{13}C NMR spectrum. Within 1 min 20 sec one structure was obtained consistent with structure **2**. Visual comparison of the experimental and calculated spectra showed a very high degree of similarity with $\boldsymbol{d_A}$ =0.93 ppm, a value which supports the correctness of the found solution. A high degree of flexibility in the software design, the presence of structure generators based on different algorithms and ability for the user to take active part in the process, allows the correct structure to be identified even in the presence of hidden contradictions in HMBC data.

Shigemory and co-workers [21] recently isolated two new diterpenoids containing an eight-membered lactone ring. The structures were elucidated utilizing a combination of spectral data. In their work the authors present a table containing ^{1}H, ^{13}C, HMQC and HMBC spectra of one of the diterpenoids, compound **3,** with a molecular formula of $C_{20}H_{30}O_3$:



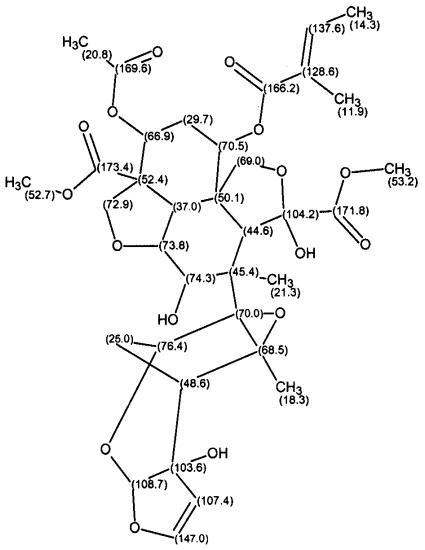
3

These data were used to further challenge the *StrucEluc* system. Having formed the MCD, the program checked connectivities between the skeletal atoms for the presence of contradictions. The data passed all tests. When structure generation was initiated the process stopped immediately and no structures were generated. This may have been caused by the presence of contradictions that cannot be recognized by the program or the presence of chemical shifts which are out of range of the characteristic intervals of the spectral filter libraries (LSC) in the experimental ^{13}C NMR spectrum. The latter can easily be checked simply by repeating the generation process with the spectral filter switched off. When the process was repeated with the spectral filter switched off, one structure was generated that was identical to **3** and characterized by the following deviations: $\boldsymbol{d_A}$= 3.395, $\boldsymbol{d_F}$= 3.122, $\boldsymbol{d_H}$=0.304. Since the *StrucEluc* system can be trained, it is valuable to determine which spectral features cause outliers to the corresponding ranges. This may be examined by using the filter options that provide three degrees of severity for the characteristic intervals: *tight, medium, and loose*. The *tight* filter corresponds to an interval width commonly used for the present fragment. The two other options automatically widen the ranges according to specific rules. The system uses the *tight* option by default and ensures the highest selectivity of the spectral filter. To understand the reason for rejection of the molecule the resulting structure was filtered initially with the *medium* degree of severity and then the *loose* one. The program reported that the C=C-CO-O fragment was found in the structure and that the conjugated carbonyl group is expected to have a characteristic interval of 160.0-170.0 ppm. In reality, the carbonyl in the experimental spectrum was observed at 171.3 ppm. In this case a slight shift of the C=O signal outside the interval range results for the diterpenoid under study. This is probably due to the fact that the carbonyl group belongs to a system containing fused eight- and twelve-membered rings. To prevent the rejection of compounds of this class in future, the corresponding interval was increased to 172 ppm.

Experimental data for the next example were obtained from a report devoted to the structure elucidation of the insect antifeed *azadirachtin* (**4**) [22]. In this publication the elucidation was performed with the assistance of the LSD program [6]. The molecular formula $C_{35}H_{44}O_{16}$ and the 1D ^{13}C, COSY, HMQC and HMBC spectra were used as inputs for *StrucEluc*.

Initially an attempt was made to elucidate the structure using the common mode and without any editing of the molecular connectivity diagram. The generation process was very time consuming and the program was halted after several hours. Since the molecule is rather large and complicated, containing 51 skeletal atoms, the authors [22] artificially set the multiplicity of all carbon atoms and assumed the presence of four ester and three hydroxyl

groups to elucidate the structure with the assistance of the LSD program. When we made the same assumptions and drew the corresponding functionalities in the molecular connectivity diagram (the R_c=3-7 constraint on the cycle sizes was also fixed during the structure generation) we obtained the results: k=9990→333, t_g=51 min and the correct structure was distinguished by its $\boldsymbol{d_A}$ value ($\boldsymbol{r_A}$ =1).



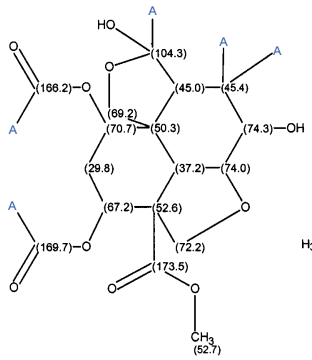
4

The main disadvantage of the solution of this problem using the approach just cited is that it required that a great many assumptions be made and numerous constraints were imposed by the user. This shortcoming can, however, be overcome by using a fragment database search to provide structural core components for *StrucEluc* to work from.

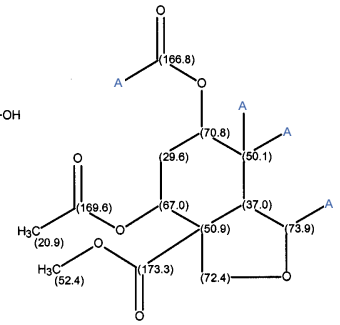
An initial check indicated that the *azadirachtin* molecule was absent from the full structure database. A fragment search resulted in a list of fragments with L=2885. The larger substructures are of the most interest for this process and the first 500 fragments were chosen for creation of the MCD diagrams. The minimum value of the chemical shifts error was set to E = 0.5 ppm. No further assumptions or constraints were imposed.

With these starting conditions the program created 8 MCDs. Examination of the MCDs showed that only two contained the large fragments **5** and **6** that could influence the process of problem solving.

All MCDs, except the two selected above, were deleted. These two MCDs passed the confirmatory connectivity check and structure generation was performed without any constraints (R_c=3-20). The results produced were: k=79→76, t_g =29 s and the priority of the correct structure was obvious since $\boldsymbol{r_A}$=$\boldsymbol{r_F}$=$\boldsymbol{r_H}$=1. It is notable that when the generation was performed from the MCD containing the largest fragment, **5,** only three structures including the correct one were generated in 3 s.

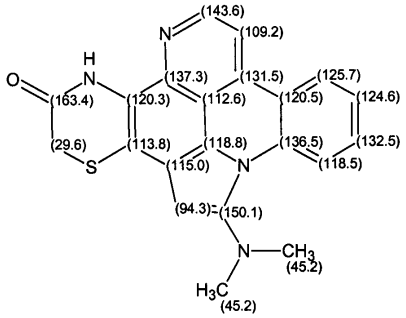


5



6

With this approach a *large and complex* natural product molecule was identified both without any atom property constraints and without the input of either carbonyl or hydroxyl groups in contrast to the numerous constraints imposed in the initial report [22]. There were also no assumptions regarding the ring sizes contained within the molecule. In addition, the time required for structure generation decreased from 51 min to 30 s. The structure output file decreased in size by a factor of four.



7

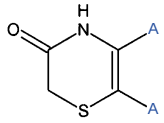
The experimental data for the next example reported here was taken from reference [23]. The authors performed the structure elucidation of *cycloshermilamine D* (**7**), a recently isolated alkaloid. The molecular formula is $C_{21}H_{16}N_4O_1S_1$ (Mw=372). In their article, the authors presented the COSY, HMQC and HMBC correlations obtained from the 2D NMR experiments.

Cycloshermilamine-D is a proton-deficient molecule; the problem is further complicated by the presence of three different heteroatoms including S (an element with a valence that can vary as 2, 4, or 6). The number of correlations detected by the authors proved to be very small (8 correlations were revealed by the COSY experiments and 20 from the HMBC experiment). When structure generation was initiated, an enormous number of structures were produced within the first several minutes. The generation

process was aborted since it was evident that the introduction of additional structural information would be necessary to help solve the problem.

The fragment search gave a list with L=1371. Many attempts were made to create MCDs from the found fragments using different E values, but the number of MCDs turned out to increase very quickly (for instance, at E=4 ppm $n_{(MCD)} \approx$ 12,000). As none of these molecular connectivity diagrams resulted in a structure, the time consuming process of creating and checking the connectivities from the found fragments was aborted. It became clear that only the introduction of a user fragment (UF) could help us to solve the problem.

The main difference between a found fragment and a user fragment is that the FF (found fragment) already contains assigned carbon atoms, while the carbon atoms of a UF (user fragment) have no associated assignments. In our example, the presence of the following fragment was assumed:



8

This hypothesis can be justified by the fact that substructure **8** is present in congenerous compounds – *shermilamines D* and *E* [24] as well as in *segolins* [25] that were isolated and identified earlier. Such assumptions can be made when families of structures are under analysis.

The accurate ^{13}C NMR spectrum prediction of substructure **8** was performed and 5 MCDs were created from this UF with E =7 ppm. The result of checking the MCDs for contradictions gave $n_{(MCD)}$ =5→2. At this stage structure generation was performed with only one constraint of R_c=5-6 imposed, taking into account that the compound under investigation, as authors of ref. [23] supposed, belongs to the *pyridoacridines*. The structure generation process delivered: k=1987→373, t_g= 8 min 25 s. After structure ranking by the $\boldsymbol{d_F}$ values accurate ^{13}C NMR spectra were calculated for the first 20 ranked structures. As a result the correct structure delivered the best match (minimal $\boldsymbol{d_A}$ value).

The application of identifying *user fragments* in the list of *found fragments* by searching structure **8** through the list of found fragments was examined. The found fragments shown in Figure 7 were identified. As indicated in the figure the fragments produced from different molecules during the process of creating the fragment DB have different chemical shift assignments.

An attempt to create MCDs from this set of 6 fragments was pursued. At E=0.5 ppm, $n_{(MCD)}$ =18→4 resulted.

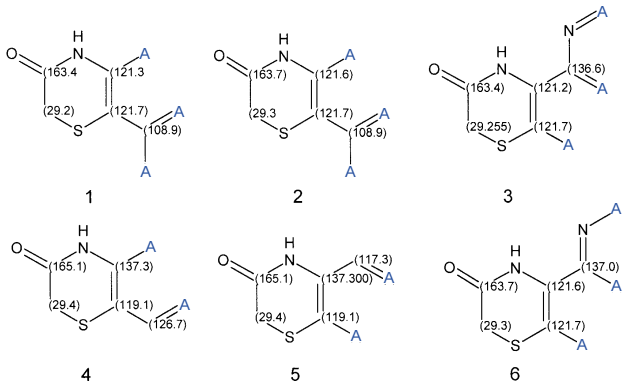


Figure 7. Library fragments containing the user-defined fragment (**8**).

Structure generation was started with only one constraint imposed on the ring sizes: R_c=5-6. The result was: k=291→52; t_g= 2 min 7 sec. Accurate ^{13}C NMR spectral prediction was performed for the first 20 structures ranked previously by the $\boldsymbol{d_F}$ values. The true structure was again correctly distinguished as shown in Figure 8. It is worth noting that structures 4-5 are very similar to the determined structure, but they are significantly distinguished from *cycloshermilamine D* by a deviation difference of

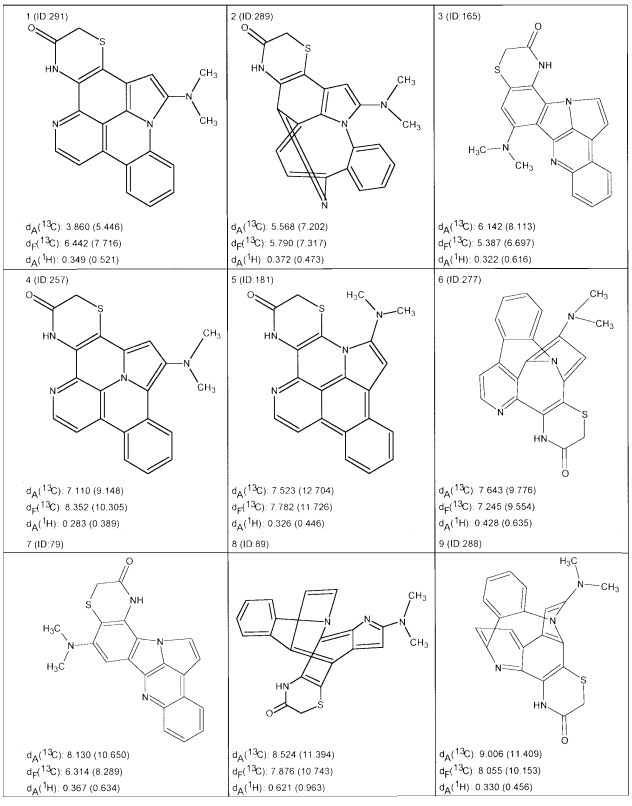
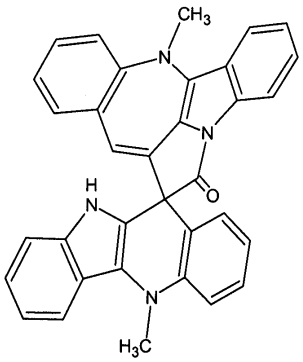


Figure 8. The initial structures in the ranked answer file generated from the HMBC and COSY connectivities and using the library fragments shown in Figure 7.

about 3.5 ppm. In the case of structures 4 and 5 shown in Figure 8, the disparity in the calculated values of d_A can be readily attributed to the electronic effects associated with changing the location of the pyrrole nitrogen.

The results obtained in this example are indeed impressive since the problem initially seemed to be hopeless due to the lack of correlations. Experience indicates that the accurate ^{13}C NMR calculation is capable of assigning carbon atoms for a user fragment with a degree of precision that allows the fragment to be accepted by the program as a useful starting point. It has also been shown that found fragments can be used to supply a user fragment with *realistic* carbon chemical shifts. Therefore, despite the fact that the list of found fragments was not used for direct MCD creation, the structural and spectral information stored in the ACD/Labs database, nevertheless, proved to be valuable for user fragment assignment and, eventually, for obtaining the solution to the problem.

Further investigations confirmed the high efficiency of the combined use of all three processes in the structure generation system described in an earlier section of this article. In particular, for compound **9** investigated by



9

Martin and co-workers [26-28] proved especially challenging. This structure could not be elucidated using 2D H-H, C-H and N-H correlations due to a large number of missing and/or overlapped cross-peaks. Our attempts to generate structures using only these available data proved to be unsuccessful and no result structure was generated after several tens of hours. To augment the basic *StrucEluc* program, a user fragments library was created from eight related *Cryptolepis* alkaloids examined in earlier work. These are presented in Figure 9. The 1D and 2D NMR data for these compounds, published in references [29-35], were entered into *StrucEluc*, and the elucidation process was performed on each one in turn. Structures 1.9-3.9 were found in the system database, while structures 4.9-8.9 were determined by 2D NMR correlations. The ^{13}C NMR assignments for 1.9-8.9 were entered into a file from

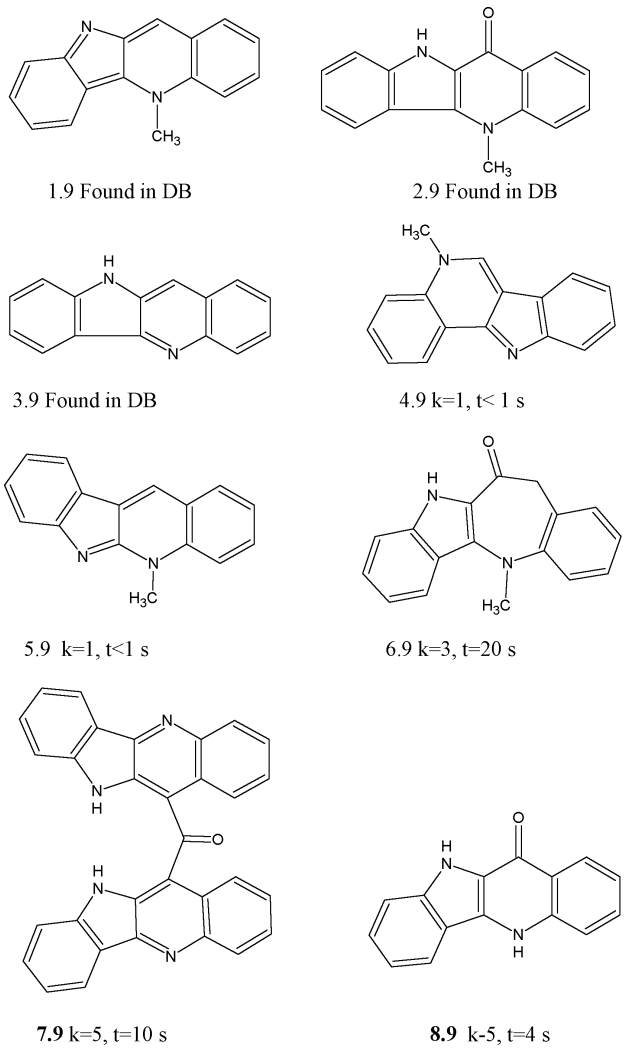
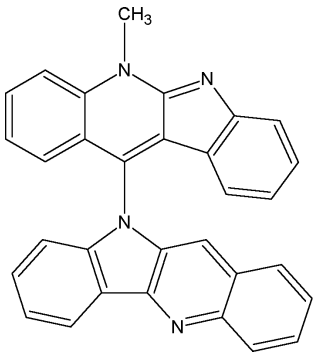


Figure 9. Elucidation of the structure of *Cryptolepis* alkaloids. Designations: *k* – number of structures in the output file, *t* – elapsed time for structure generation.

which *StrucEluc* automatically created a user library containing 342 fragments.

A search through this user database using the 1D ^{13}C NMR spectrum of cryptospirolepine detected fragments whose addition to the MCD provided much of the information missing from the experimental 2D NMR data. As a result, the structure generation process took 18 minutes and produced only one structure—the correct one, *cryptospirolepine*.

The user's fragment database created in the process of determining the structure of *cryptospirolepine* found practical application recently for the identification of degradants of *cryptospirolepine* [36-37], and also in the elucidation of the structure of *quindolinocryptotackieine* (**10**) a molecule whose structure had remained intractable to human solution for twelve years [38].



10

Conclusions.

In this work the methodology for the structure elucidation of complex organic compounds from 2D NMR data using the *StrucEluc* expert system has been investigated. Challenging situations have been revealed in the process of solving about 150 problems based on the identification of newly isolated natural products. For these studies 2D NMR data published in the original articles were used. Difficult situations arise under particular circumstances:

- The 2D NMR data have long-range correlations with spin couplings of $^{4}J_{HH}$ and/or $^{4}J_{CH}$ or $^{n}J_{HH}$ /$^{n}J_{CH}$, where n>4. The program defaults are set for correlations more typical for COSY and HMBC data.
- The number of experimentally available 2D NMR correlations is markedly less than the number of theoretically possible correlations for a given structure.
- The number of 2D correlations is small due to the proton-deficient nature of the molecule and/or specific stereochemical factors. A molecule under investigation is so large that even a rich 2D NMR data can lead to an enormous time overhead for structure generation.

It has been shown that the *StrucEluc* system incorporates a wide range of facilities that can overcome the difficulties just cited. These capabilities include:

- the search and removal of contradictions in 2D NMR data;
- a powerful knowledge base containing 200,000 structures and 1,000,000 fragments with assigned ^{13}C NMR spectra (sub-spectra);
- a flexible algorithm to allow overlapping of the knowledge base with the 2D NMR data;
- a structure generator offering three different generation algorithms dependent on situation;
- facilities to allow collaborative interaction between the researcher and the software;
- intuitive aids allowing explanation of the system operations at all stages of structure identification.

It has been shown using a number of examples that the use of fragments found in the knowledge database, as well as fragments introduced by the user, allows the solution of problems that cannot be solved except by manual interpretation. The possible methodology of searching for the correct solution in the presence of non-obvious contradictions is considered. The authors conclude that in the hands of a qualified spectroscopist the synergistic interaction between the *StrucEluc* system and the scientist can be a powerful approach to simplifying the structure elucidation of newly isolated natural products and other complex chemical structures. To date, the *StrucEluc* program has already been successfully applied to the automated structure elucidation of a series of *cryptospirolepine* derivatives and for the identification of *cryptospirolepine* degradants [36] and the solution of the structure of *quindolinocryptotackieine* (**10**) that was intractable to human solution for twelve years [38].

In spite of the promising results obtained to date the *StrucEluc* system obviously has need for further development. The development of new pulse sequences which provide ever more information regarding the nature of long-range heteronuclear couplings specifically will provide added information which will further enable the elucidation of a structure. In future work the intention is to enable the program to determine the stereochemistry of the identified molecule. This remains a significant but not necessarily insurmountable challenge

REFERENCES AND NOTES

* Correspondence to: Dr. Antony J. Williams, Advanced Chemistry Development, 90, Adelaide Street, Suite 702, Toronto, Ontario, M5H 3V9, CANADA; E-mail address: tony@acdlabs.com; Tel: 919-928-9035; Fax: 425-790-3749.

[1] M. Munk, *J. Chem. Inf. Comput. Sci*., **38**, 997 (1998).

[2] K. Funatsu and S. Sasaki, *J. Chem. Inf. Comput. Sci*., **36**, 190 (1996).

[3] C. Peng, S. Yuan, C. Zheng and Y. Hui, *J.Chem. Inf. Comput. Sci*., **34**, 805 (1994).

[4] C. Peng, S. Yuan, C. Zheng, Y. Hui, H. Wu, K. Ma and X. Han, *J. Chem. Inf. Comput. Sci*., **34**, 814 (1994).

[5] C. Peng, S. Yuan, C. Zheng, Z. Shi and H. Wu, *J. Chem. Inf. Comput. Sci*., **35**, 539 (1995).

[6] J.-M. Nuzillard and G. Massiot, *Tetrahedron*, **47**, 3655 (1991).

[7] C. Steinbeck, *Angew. Chem. Int. Ed. Engl*., **35**, 1984 (1996).

[8] C. Steinbeck, *J.Chem. Inf. Comp. Sci*., **41**, 1500 (2001).

[9] J. Meiler and M. Will, *J. Chem. Inf. Comp. Sci*., **41**, 1535 (2001).

[10] T. Lindel, J. Junker and M. Koeck, *Eur. J. Org. Chem*., **3**, 573 (1999).

[11] G. E. Martin, "Exploitation of Long-range Heteronuclear Coupling Constants – New Techniques for Long-Range Heteronuclear Shift Correlation Spectrocopy," *Ann. Rep. NMR Spectrsosc*., G. A. Webb, Ed., Academic Press, New York, 2002, vol **46**, pp 37-100.

[12] T. Sprang and P. Bigler, *Magn. Reson. Chem.*, **41**, 177 (2003).

[13] G. E. Martin and C. E. Hadden, *J. Nat Prod.*, **63**, 543 (2000).

[14] R. Marek and A. Lyčka, *Curr. Org. Chem.*, **6**, 35 (2002).

[15] K. Blinov, M. Elyashberg, S. G. Molodtsov, A. J. Williams and E. Martirosian, *Fresenius J. Anal. Chem*., **369**, 709 (2001).

[16] M. Elyashberg, K. Blinov and E. Martirosian, Autom. Inf. Manage., **34,** 15 (1999).

[17] M. Elyashberg, K. Blinov, A. J. Williams, S. Molodtsov and E. Martirosian, *J. Nat. Prod*., **65**, 693 (2002).

[18] P. Willett, J. Barnard and G. Downs, *J. Chem. Inf. Comput. Sci*., **38**, 983 (1998).

[19] P. Leone, J. Redburn, J. Hooper and R. Quinn, *J. Nat. Prod.*, **63**, 694 (2000).

[20] S. Habtemariam, B. Skeleton, P. Waterman and A. White, *J. Nat. Prod*., **63**, 512 (2000).

[21] H. Shigemori, S. Shimamoto, M. Sekiguchi, A.Ohsaki and J.Kobayashi, *J. Nat. Prod*., **65**, 82 (2002).

[22] S. Ley, K. Doherty, G. Massiot and J.-M. Nuzillard, *Tetrahedron*, **50**, 12267 (1994).

[23] C. Koren-Goldshlager, M. Aknin and Y. Kashman, *J. Nat. Prod*., **63**, 830 (2000).

[24] C. Koren-Goldshlager, M. Aknin, E. Gaydou and Y. Kashman, *J. Org. Chem*. **63**, 4601 (1998).

[25] A. Rudi and Y. Kashman, *J. Org. Chem*., **54**, 5331 (1989).

[26] A. N. Tackie, G. L. Boye, M. H. M. Sharaf, P. L. Schiff Jr., R. C. Crouch, T. D. Spitzer, R. L. Johnson, J. Dunn, D. Minick and G. E. Martin, *J. Nat. Prod*. **54**, 653 (1993).

[27] C. E. Hadden, G. E. Martin, A. N. Tackie and P. L. Schiff Jr., *J. Heterocyclic Chem*., **36**, 1115 (1999).

[28] G. E. Martin, C. E. Hadden, A. N. Tackie, M. H. M. Sharaf and P. L. Schiff Jr., *Magn. Reson. Chem.*, **37**, 529 (1999).

[29] S. Ablordeppey, C. Hufford, R. Borne and D. Dwumu-Badu, *Planta*, **56**, 416 (1990).

[30] R. C. Crouch, A. O. Davies, T. D. Spitzer, G. E. Martin, C. H. Phoebe Jr., M. H. M. Sharaf, P. L. Schiff Jr. and A. N. Tackie, *J. Hetercyclic Chem*., ***32***, 1077 (1995).

[31] M. H. M. Sharaf, P. L. Schiff Jr., R. C. Crouch, A. O. Davies, C. W. Andrews, G. E. Martin, C. H. Phoebe Jr., and A. N. Tackie, *J. Hetercyclic Chem***.**, **32**, 1631 (1995).

[32] M. H. M. Sharaf, P. L. Schiff Jr., G. E. Martin, Jr., C. H. Phoebe and A. N. Tackie, J. *Heterocyclic Chem*., **33**, 239 (1996).

[33] M. H. M. Sharaf, P. L. Schiff Jr., D.Minick, R. L. Johnson, R. C. Crouch, C. W. Andrews, G. E. Martin, C. H. Phoebe Jr. and A. N. Tackie, *J. Heterocyclic Chem*., **33**, 789 (1996).

[34] C. E. Hadden, W. K. Duholkc, J. E. Guido, R. H. Robins, G. E. Martin, M. H. M. Sharaf and P. L. Schiff Jr., *J. Heterocyclic Chem*., **36**, 525 (1999).

[35] T. D. Spitzer, R. C. Crouch, G. E. Martin, M. H. M. Sharaf, P. L. Schiff Jr., A. N. Tackie, and G. L. Boye, *J. Heterocyclic Chem*., **28**, 2065 (1991).

[36] G. E. Martin, C. E. Hadden, D. J. Russell, B. J. Kaluzny, J. E. Guido, W.K. Duholke, B. A. Stiemsma, T. J. Thamann, R. C. Crouch, K. Blinov, M. Elyashberg, E. Martirosian, S. G. Molodtsov, A. J.Williams, and P. L. Schiff Jr., *J. Heterocyclic Chem.*, **39**, 1241 (2002).

[37] K. Blinov, D. Carlson M. Elyashberg, G. E. Martin, S. G. Molodtsov, E. R. Martirosian, and A. Williams. *Magn. Reson. Chem.*, **41**, 359 (2003).

[38] K. Blinov, M. Elyashberg, E. R. Martirosian, S. G. Molodtsov, A. J. Williams, A. N. Tackie, M. H. M. Sharaf, P. L. Schiff, Jr., R. C. Crouch, G. E. Martin, C. E. Hadden, J. E. Guido, and K. A. Mills, *Magn. Reson. Chem.*, **41**, 577 (2003).