

## ARTICLES

**Structure Elucidator: A Versatile Expert System for Molecular Structure Elucidation from 1D and 2D NMR Data and Molecular Fragments**

Mikhail E. Elyashberg,<sup>§</sup> Kirill A. Blinov,<sup>§</sup> Antony J. Williams,<sup>\*,†</sup> Sergey G. Molodtsov,<sup>‡</sup>  
Gary E. Martin,<sup>||</sup> and Eduard R. Martirosian<sup>§</sup>

Advanced Chemistry Development, Moscow Department, 6 Akademik Bakulev Street, Moscow 117513,  
Russian Federation, Advanced Chemistry Development, Inc., 90 Adelaide Street West, Suite 600,  
Toronto, Ontario, M5H 3V9 Canada, Novosibirsk Institute of Organic Chemistry, Siberian Branch of Russian  
Academy of Science, Lavrentiev Avenue 9, Novosibirsk 630090, Russia, and Rapid Structure Characterization  
Group, Global Research and Development, Pfizer, Kalamazoo, Michigan 49001-0199

Received May 28, 2003

*StrucEluc* is an expert system that allows the computer-assisted elucidation of chemical structures based on the inputs of a series of spectral data including 1D and 2D NMR and mass spectra. The system has been enabled to allow a chemist to utilize fragments stored in a fragment database as well as user-defined fragments submitted by the chemist in the structure elucidation process. The association of fragments in this way has been shown to dramatically speed up the process of structure generation from 2D NMR data and has helped to minimize or eliminate the need for user intervention thereby further enabling the vision of automated elucidation. The use of fragments has frequently transformed very difficult 2D NMR elucidation challenges into easily solvable tasks. A strategy to utilize molecular fragments has been developed and optimized based on specific challenging examples. This strategy will be described here using real world examples. Experience gained by solving more than 150 structure elucidation problems from a variety of literature sources is also reviewed in this work.

## 1. INTRODUCTION

Two generations of the *StrucEluc* expert system (ES) have been described previously.<sup>1,2</sup> The first generation system, *StrucEluc-1*,<sup>1</sup> was developed for the structure elucidation of organic molecules from 1D <sup>13</sup>C NMR spectra and was found to be successful for molecules in the mass range of 150–300 and containing up to 20–25 skeletal atoms. The second system (*StrucEluc-2*)<sup>2</sup> is capable of elucidating the chemical structure of large molecules, up to a mass of 1285 amu to date and 90 skeletal atoms; typically in the area of natural product structure elucidation, this task is mainly accomplished from 2D NMR spectral data. In general, the system has been designed to elucidate structures containing up to 250 skeletal atoms.

The capabilities of the *StrucEluc-2* system in terms of general utility as a tool for the structure elucidation of natural products have been demonstrated previously.<sup>3</sup> In this study, *StrucEluc-2* was applied to the structure determination of 60 recently isolated natural products. The experimental data were obtained from a series of original articles published mainly in 2000. The strategy of structure elucidation was discussed in detail.

It should be emphasized that problems described in ref 3 were solved using only heteronuclear (HMQC, HMBC, or COLOC) and homonuclear (H–H COSY) 2D NMR correlations and without using any additional structural information. In this mode of program operating, referred to as the *common* mode, the system creates connectivities from the spectral data and generates all possible structures in accordance with default settings for the number of intervening bonds between corresponding skeletal atoms and with atom properties including the state of hybridization and the possibility of neighboring heteroatoms being taken into account.

Further investigations challenged *StrucEluc-2* with problems that could not be solved or proved to be very time-consuming due to a lack of information in the 2D NMR data. In these cases it proved necessary to introduce additional structural information, if available, to facilitate the elucidation process. In the real world, it is common for a chemist or spectroscopist faced with the need to elucidate a structure to have prior knowledge of reaction components in a synthesis, knowledge of the class of compounds that may have been isolated, or even hypothetical structures for validation rather than full elucidation.

The hypothesis that the utilization of molecular fragments found from the system knowledge base<sup>1</sup> or potential substructures proposed by the chemist would be helpful to circumvent the difficulties was tested. The *fragment approach* has been used in a number of first generation expert

\* Corresponding author phone: (919)570-0217; fax: (425)790-3749; e-mail: tony@acd labs.com.

<sup>†</sup> Advanced Chemistry Development, Inc.

<sup>‡</sup> Siberian Branch of Russian Academy of Science.

<sup>§</sup> Advanced Chemistry Development.

<sup>||</sup> Pfizer.

systems. It is based on correlation tables containing substructures and their associated characteristic intervals of spectral feature, e.g., chemical shift variation (for example, see refs 4–6). In contrast, systems including both SpecSolv<sup>7</sup> and *StrucEluc-1*<sup>1</sup> employ databases containing substructures and their associated <sup>13</sup>C NMR subspectra. At present, the *StrucEluc* database contains over 215 000 chemical structures and more than one million substructures. The database continues to grow as further literature data is added. The value of including substructures directly into the elucidation process is that a fragment, considered as a macroatom, can absorb a significant number of the skeletal atoms and leads to a reduction in the complexity of the problem. This results in the acceleration of the structure generation process, which is typically the most time-consuming stage of the structure elucidation process.

Nevertheless, even in the case when 2D NMR data are employed, the utilization of molecular fragments is hampered by the fact that *all* carbon atoms existing in a fragment utilized in solving the problem *must* be supplied with chemical shifts. Moreover, the values of these chemical shifts must be as close as possible to the observed values for the atoms of the corresponding fragments in the experimental <sup>13</sup>C NMR spectrum of the unknown under study. In addition, the accommodation of one or more fragments within a set of connectivities derived from the 2D NMR data is a problem that requires the development of new algorithms. To our knowledge, no attempts have been made to investigate the possibility of using molecular fragments for structure generation from 2D NMR connectivities.

In this work, attention has been focused on methods that allow the utilization of fragments stored in both the <sup>13</sup>C NMR database (*found fragments*, FF) and those introduced by the user (*user fragments*, UF) in combination with 2D NMR data. To develop an appropriate strategy and refine the methodology for the utilization of fragments, a series of challenging tasks related to the chemistry of natural products has been examined.

During this work, more than 150 structural problems were solved in order to develop and test fragment utilization-based methods to elucidate the chemical structures. Advantages and disadvantages inherent to this approach were identified. The utilization of fragments provides the following main advantages: (1) some problems that cannot be solved in the *common* mode are solvable when fragments are utilized in the process; (2) commonly user intervention or the necessity of structural assumptions is minimal; and (3) generally the time for structure generation is drastically reduced when compared with the *common* mode.

## 2. FUNCTIONAL SCHEME OF THE SYSTEM

Many abbreviations will be introduced in this section and will be used for the sake of brevity and ease of communication. For reference purposes they have been included at the end of this manuscript and are listed prior to the references section. The flowchart represented in Scheme 1 provides an overview of the system. As shown, the system consists of three interrelated branches, each having access to the system database. The *knowledge base* (**KB**) of *StrucEluc* consists of three components: (1) a library of 215 000 molecular structures and their associated <sup>13</sup>C NMR spectra; (2) a

*fragment library* (**FL**) containing more than 1 million fragments with their corresponding <sup>13</sup>C NMR subspectra; and (3) a *Library of Spectrum-to-Structure Correlations* (**LSC**) comprising the most widespread functional groups and their accompanying characteristics in both their NMR and IR spectra. The **FL** is created from the full molecular structures stored in the **KB** using proprietary fragmentation algorithms.

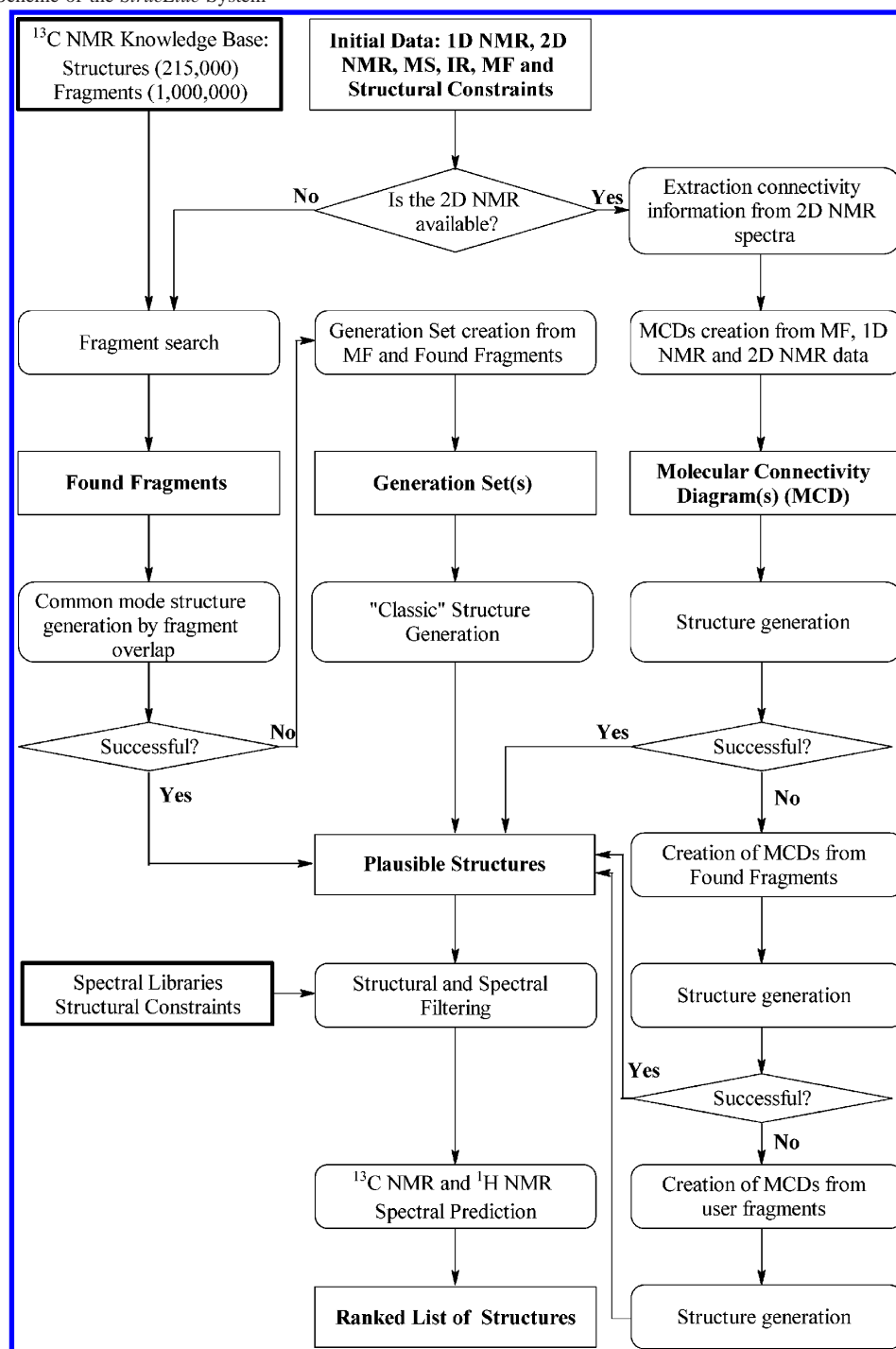
Depending on the initial data available and the complexity of the molecule being analyzed, the system offers a wide range of methods for solving a problem. These methods are outlined briefly below.

**2.1. System Operation with 1D NMR Spectra. 2.1.1. "Standard" Mode.** In this mode (see Scheme 2), similar to that first described in ref 7, a <sup>13</sup>C NMR spectrum is used as the main experimental data input for the elucidation process. The chemical shifts, multiplicity, and intensity (number of carbon atoms) of signals are specified. Quite frequently, an experimental <sup>13</sup>C NMR spectrum is sufficient to elucidate the structure of a molecule of up to 20–25 skeletal atoms. However, IR and <sup>1</sup>H NMR spectra, if available as supplementary data, can certainly assist with the solution of the problem as explained later. In principle, the program is capable of working *without* either a molecular mass (**M**) or molecular formula (**MF**).

The program starts with a search of the entire <sup>13</sup>C NMR spectrum in the **KB**. If the spectrum is found in the database, the corresponding structure is displayed; otherwise, the search of the **FL** is carried out. As a result of the search in the **FL**, the program displays *L* fragments that have corresponding subspectra that do not contradict the experimental spectrum. The generated fragments are then ranked in order of decreasing number of carbon atoms. If <sup>1</sup>H NMR and IR spectra are available, the program can filter the selected substructures with the help of the **LSC** library, which can noticeably reduce the number of selected fragments and move the "good" fragments to the beginning of the list. The possibility of merging the fragments by overlapping common atoms is checked with the intent of assembling these fragments into a final structure. After the merging of fragments, the <sup>13</sup>C NMR subspectra of the integrated fragment are predicted to help identify whether the fragments have been merged correctly.

If the result of checking all possible fragment combinations is the detection of a structure having no free bonds, then <sup>13</sup>C NMR spectrum prediction and signal assignment are performed. Each structure is then verified in accordance with all available spectral data, general rules of structural chemistry, and any constraints imposed by the chemist. If several structures are generated, then they are ranked in order of increasing *d* value which is calculated as the sum of deviations found for each <sup>13</sup>C NMR signal divided by the total number of signals in the spectrum. If an attempt to assemble a structure fails, then the program automatically switches to the "classical" mode of the **ES** that requires the availability of the **MF**. The X-PERT<sup>4</sup> system serves as an example of this "classical" approach.

**2.1.2. The "Classical" Approach.** The "classical" approach for the *StrucEluc* system (see Scheme 2) starts with a molecular formula calculation and the formation of sets of fragments. The methods have been described previously in ref 8. The first *l* basis fragments, (*l* = 10–30), *l* ≤ *L*, are

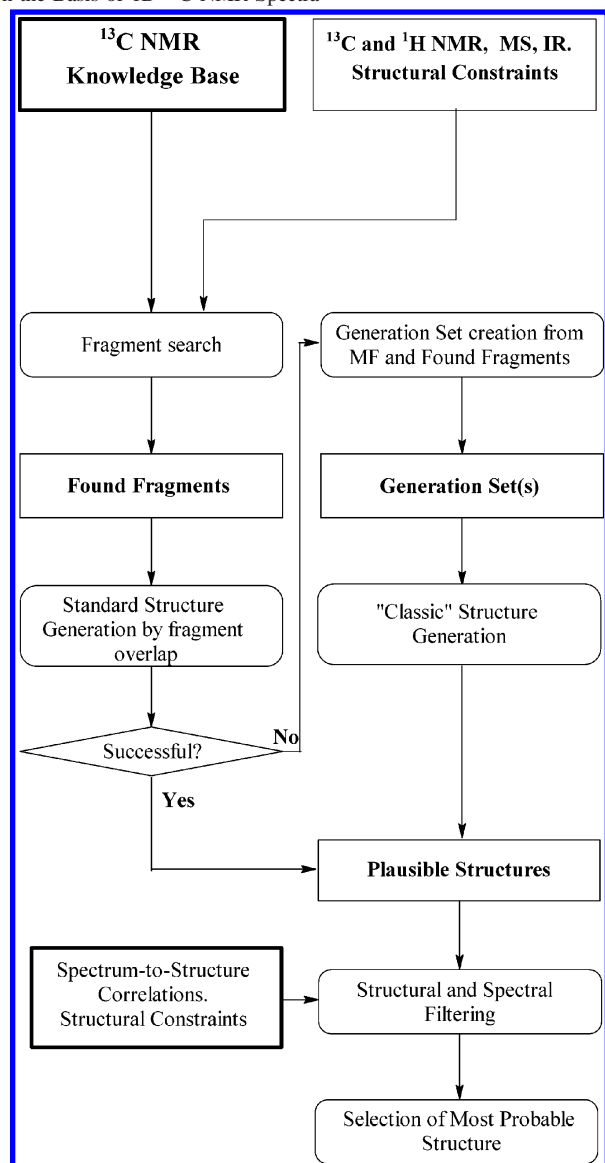
Scheme 1. General Scheme of the *StrucEluc* System

chosen from the list of selected fragments. They will be the largest fragments.

The subspectrum of the first basis fragment is compared with the experimental spectrum. The chemical shifts that were not originally identified in the experimental spectrum are associated. The first *basis* fragment is then combined with combinations of smaller fragments to provide complete spectral interpretation. The sets from the second, third, etc. basis fragments are similarly formed. Each set is then checked for correspondence with the suggested MF. A chemist may add sets of fragments and apply appropriate restrictions common to most "classical" expert systems: GOODLIST, BADLIST, minimum and maximum ring cycle

sizes, multiplicities of bonds, etc., see refs 4–6. The program is capable of displaying a "generalized portrait" of the analyzed molecule. This overview is a distribution of functional groups included in the Typical Functional Group Library (TFGL) by the frequency of their occurrence in the fragments selected from the FL.<sup>1</sup> Groups that are completely absent and those that occur most frequently are correspondingly used to automatically form the BADLIST and GOODLIST.

All of the generated structures are checked for their correspondence with the available spectra as well as with the rules of organic chemistry and stereochemistry. Finally, to choose the most probable structure, <sup>13</sup>C NMR spectra of

**Scheme 2.** A Flowchart of “Standard” and “Classic” Modes Operating on the Basis of 1D  $^{13}\text{C}$  NMR Spectra

candidate structures are predicted, and the structures are ranked in order of increasing average deviations of the predicted spectra from the experimental data. The method that finds the most probable structure is described below in section 2.2.4. It has been shown that the system operating in this 1D NMR mode can generally elucidate structures of medium size, up to 20–25 skeletal atoms.

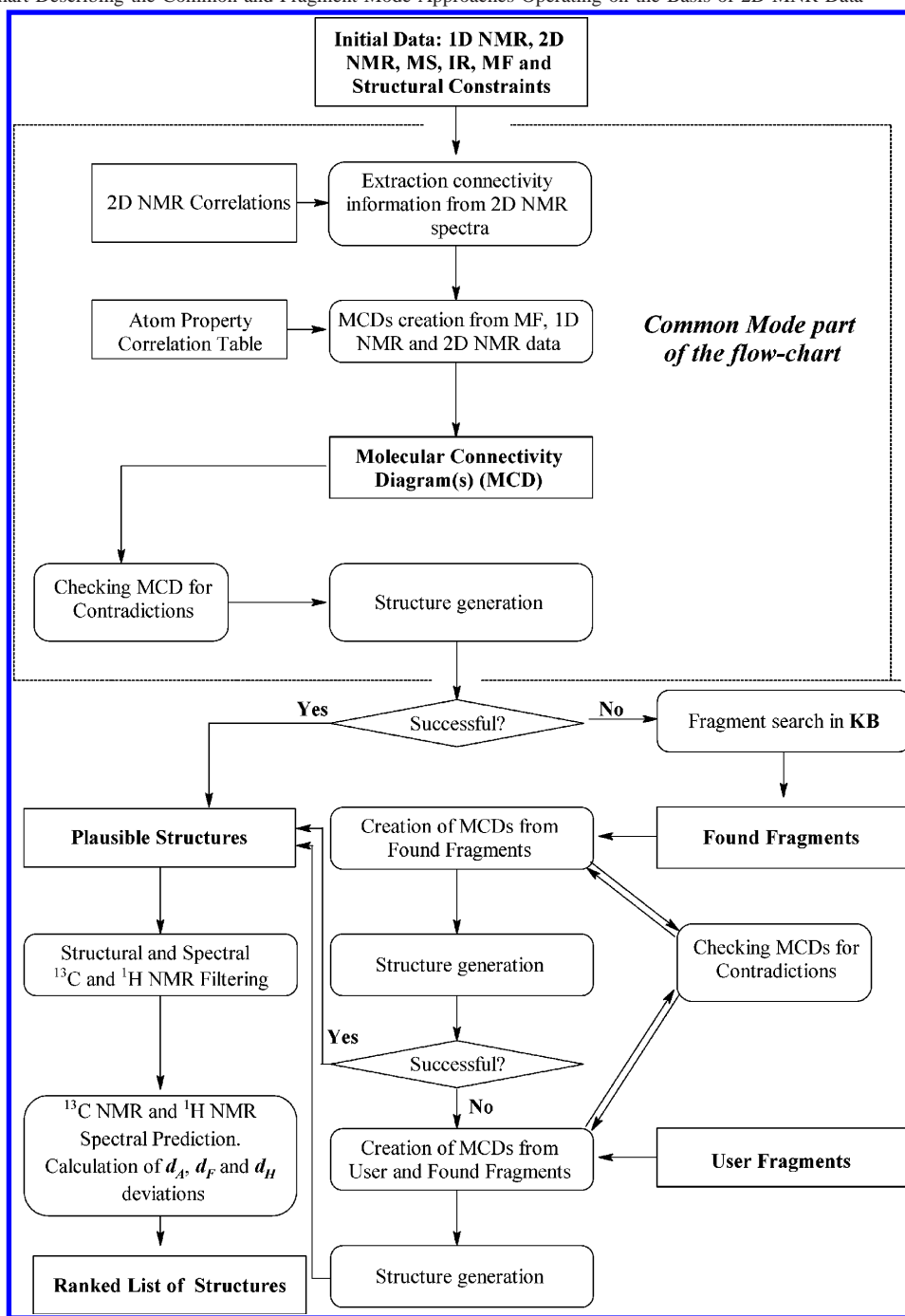
**2.2. Molecular Structure Elucidation from 2D NMR Spectra.** **2.2.1. The “Common” 2D NMR Mode.** If an analyzed molecule is large and is related to newly isolated complex natural products, it is most probable that both the “standard” and “classical” approaches will fail to elucidate a final structure. Experience has shown reliable elucidation of the structure of large molecules (containing > 20–25 skeletal atoms) is generally impossible without employing 2D NMR data.

A third module of the *StrucEluc* system (see Scheme 3) is based on a number of programs developed for elucidating a molecular structure from a combination of 2D NMR spectra. The most typical combination providing a basis for structure determination includes H–H COSY, HSQC/

HMQC, and HMBC. The *StrucEluc* system presently operates with the following 2D NMR methods: H–H COSY, HSQC/HMOC, HMBC, ROESY, NOESY, and INAD-EQUATE. Other methods can be also used by the system through a flexible procedure that allows input and processing of experimental 2D NMR data. In addition to spectral data the program also needs the molecular formula to proceed in this mode.

First, the coordinates of the 2D cross-peaks should be determined and entered into the program. This procedure can be carried out either automatically or manually. The program is presently capable of reading data represented in the primary vendor formats of Varian, Bruker, and JEOL. The program forms peak tables where signal intensities are also displayed. These tables are converted into tables indicating carbon atom connectivities. To provide analysis and editing of the initial data, all  $\text{CH}_3$ ,  $\text{CH}_2$ ,  $\text{CH}$ , and  $\text{C}$  groups as well as heteroatoms and hydrogens attached to heteroatoms are displayed on the screen with their respective  $^1\text{H}$  and  $^{13}\text{C}$  chemical shifts. The resulting presentation corresponds to a molecular connectivity diagram (MCD) and an example is shown in Figure 2. HMBC and COSY connectivities are shown as subgraphs (fragments) connecting carbon atoms and/or carbon and nitrogen atoms when the corresponding  $^1\text{H}$ – $^1\text{H}$  COSY and  $^{15}\text{N}$  HMBC data are available. Different possible distances between the connected atoms are marked on the MCDs with specific colors. The program analyzes the  $^{13}\text{C}$  and  $^1\text{H}$  chemical shifts of the  $\text{CH}_n$  groups ( $n = 0$ –3) and automatically sets parameters showing the allowed hybridization and possible heteroatom neighborhood for each carbon atom. To carry out this procedure, special Atom Property Correlation Tables (APCT) are derived from fragments stored in the FL. In the latest version of the *StrucEluc* system, the relationship to neighboring heteroatoms is marked as “forbidden” (fb), “at least one”, “at least two”, “at least three”, “four” and “not defined” (nd). The possible states of atom hybridization are designated as  $sp^3$ ,  $sp^2$ ,  $sp$ , *not sp*, and “not defined”. Those descriptors for the carbon atoms allow the system to analyze 2D NMR data and to efficiently apply restrictions during the process of structure generation. The chemist can then edit these parameters using other available information. For example, if the molecule belongs to the  $\text{CHNO}$  class and the carbon atom C(175.5) is marked as  $sp^2$  at least two, the system will only generate  $\text{O}=\text{C}=\text{O}$  and  $\text{N}=\text{C}=\text{O}$  fragments on the basis of that atom. The chemist is also offered the opportunity to draw bonds of any multiplicity between the atoms and to set some functional groups (for instance  $\text{C}=\text{O}$ ,  $\text{O}=\text{C}=\text{O}$ ,  $\text{O}=\text{H}$ , etc.). In so doing, the structural information evident from the  $^1\text{H}$  NMR and IR spectra can be successfully used. The lengths of the COSY and HMBC connectivities are taken as *default* to be 1 and 1–2 bonds correspondingly, thus indicating the number of bonds between skeletal atoms. We call these “standard length” connectivities. Since the connectivity tables contain information that initially will be used during the structure generation process, they should be examined for consistency. For these purposes, specific criteria allowing the detection of the presence of contradictions between some atom pairs have been employed. As a rule, the main cause of contradictions is in setting the distances between intervening nuclei that are shorter than those in the molecule under investigation. The method of detecting the



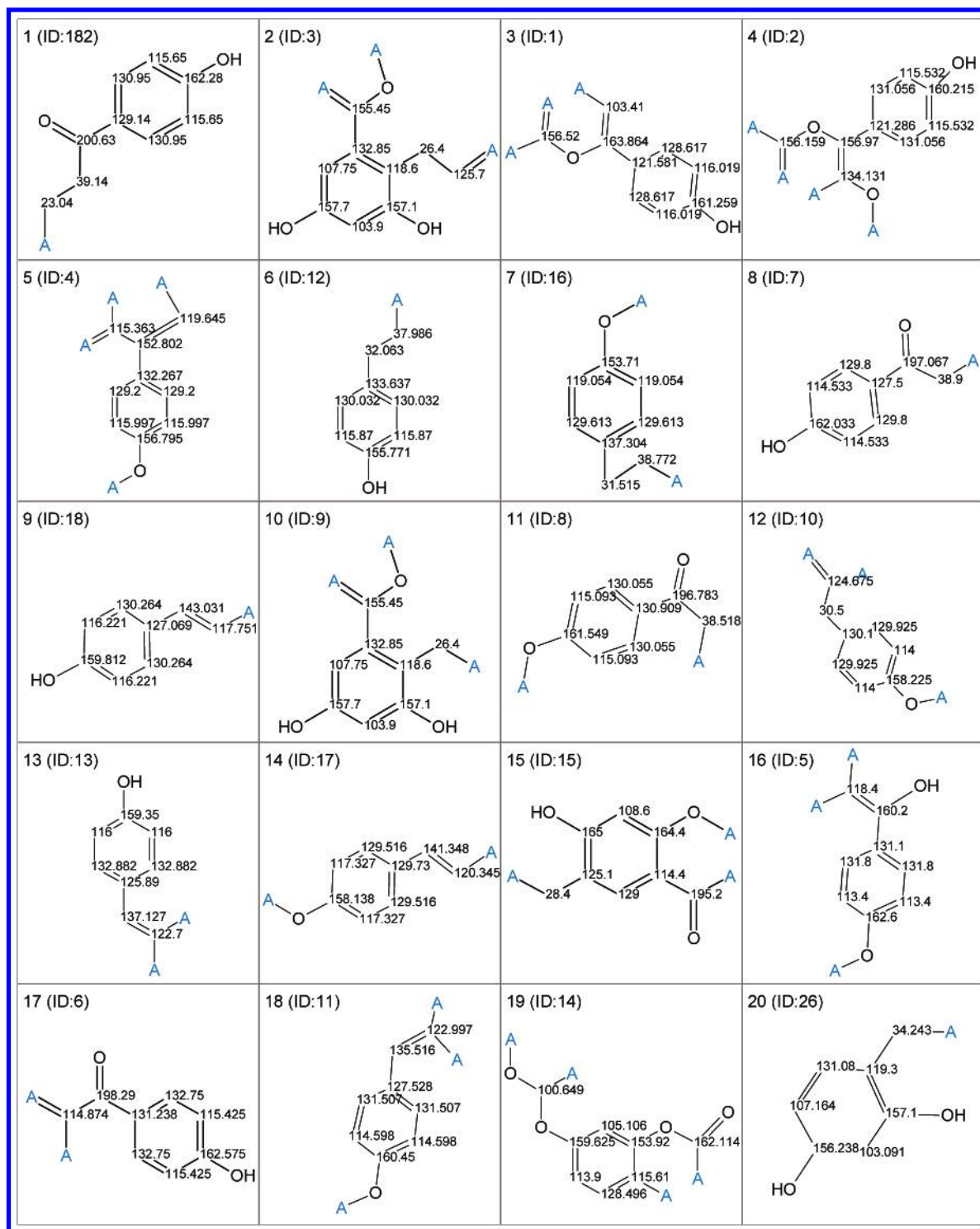
**Scheme 3.** A Flowchart Describing the Common and Fragment Mode Approaches Operating on the Basis of 2D NMR Data

presence of contradictions and their subsequent removal is described in ref 9.

The data formed at this stage are used as the input data for the 2D structure generator developed in this work. Structures are generated under the constraints determined from the MCD diagrams and any additional constraints that may be introduced by the chemist. The structure generator is based on mathematical algorithms described in refs 10–12. Generated structures are verified using the approaches discussed previously: filtering with LSC, GOODLIST, BADLIST, etc. <sup>13</sup>C NMR spectrum prediction is performed for all structures included in the output file, and the structures are ranked in order of increasing *d* value using the method described in detail in section 2.2.4.

The system can also be used for verification of the proposed structure and the associated <sup>13</sup>C and <sup>1</sup>H NMR signal assignments by validation of all available two-dimensional NMR spectra. If any contradictions are found, for instance if the distance between a pair of carbon atoms in a proposed structure is greater than that postulated from a particular connectivity, the program displays a textual message detailing the probable cause(s) of the conflict(s) and showing the connectivities in graphical form.

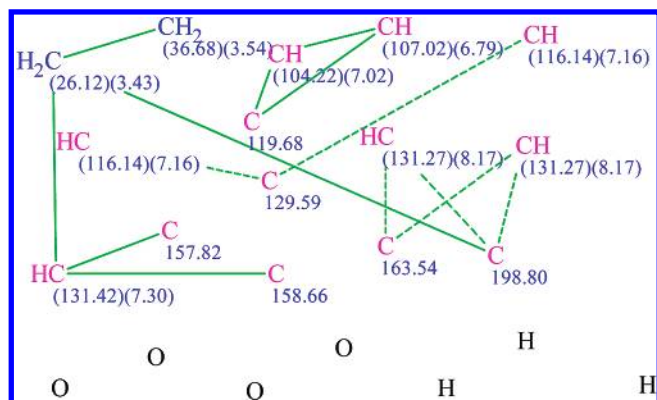
**2.2.2. Application of Fragments in Combination with 2D NMR Data.** As previously reported,<sup>3</sup> computer-assisted structure elucidation using 2D NMR data is quite efficient for the structures of complex natural compounds. However, if the structural restrictions imposed by the MCD are not



**Figure 1.** The first 20 fragments obtained in the found fragment list during the elucidation of compound 1.

sufficient for the generation of a reasonable number of possible structures within an appropriate time, it is to be expected that the utilization of molecular fragments can greatly facilitate solution of the problem solution. Commonly appropriate fragments to aid in the solution of a problem can be found in the knowledge base. The main advantage of these fragments is that all fragment carbon atoms are supplied with the <sup>13</sup>C NMR assignments obtained from the full structures that were used for creation of the fragment DB.

The first step of the process is a fragment search of the KB (see Scheme 3). As a result, a set of  $L$  found fragments is selected. The next step is the creation of the MCDs using the found fragments (FF). For this purpose, either all FFs or any selected number of them can be incorporated by the operator. The main idea of the algorithm that implements this procedure is as follows. Prior to creation of the MCD, the chemist defines the number of fragments,  $l$  ( $l \leq L$ ), that will be used in this procedure and sets an error,  $E$ , that defines the maximum difference allowed between the chemi-



**Figure 2.** Molecular connectivity diagram created for compound 1.  $^{13}\text{C}$  and  $^1\text{H}$  NMR chemical shifts of carbon and hydrogen atoms are shown in brackets.

cal shifts of the fragment carbons and the corresponding values observed in the experimental spectrum under study. It is important to note that both parameters,  $l$  and  $E$ , are closely interrelated and choosing the most efficient values may be a matter of trial and error.

The  $^{13}\text{C}$  NMR subspectrum of each fragment is compared with all experimental chemical shifts. The number of hydrogen atoms attached to a carbon atom is taken into account during this process. Consider a fragment contains  $f$  carbon atoms and an arbitrary atom  $\text{C}_i$  of the fragment has a chemical shift  $\delta_i$  ( $i = 1 \div f$ ) and multiplicity  $m_i$ . Suppose that the experimental chemical shifts  $\delta_{i1}, \delta_{i2}, \dots, \delta_{iq}, \dots, \delta_{ip}$  meet conditions  $|\delta_i - \delta_{iq}| \leq E$  and  $m_i = m_{iq}$ . Then all possibilities of substituting the  $d_i$  for the experimental values  $\delta_{i1}, \delta_{i2}, \dots, \delta_{iq}, \dots, \delta_{ip}$  must be verified.

If the conditions  $|\delta_i - \delta_{iq}| \leq E$  and  $m_i = m_{iq}$  hold for all  $f$  carbon atoms, then the given fragment is recognized as a candidate for inclusion in the process of creating the MCD. If this condition does not hold, then the fragment is excluded from consideration. It is possible that one experimental chemical shift  $\delta_{iq}$  can also substitute chemical shifts assigned to several carbon atoms within the fragment. All rearrangements of the experimental chemical shifts within the corresponding carbon atoms of the fragment should be considered. The chemical shift distribution of carbon atoms that produce a conceivable assignment of a given fragment carbon atom has to be verified. During the verification process, the program checks whether the carbon atom assignments correspond to the experimental chemical shift correlations comprising the skeletal atoms making up the fragment. The fragments that survive the test are then included in the set of *prospective* fragments.

The more skeletal atoms that are “absorbed” by the fragments the shorter the process of structure elucidation. With this in mind, the algorithm that combines the prospective fragments within one molecular connectivity diagram was developed. To realize this procedure, all possible combinations of prospective fragments are searched, and only combinations that are in agreement with the experimental 2D NMR correlations are chosen. The fragment combinations that pass this examination form a set of *prospective fragment combinations*. These fragments are then “projected” onto the MCDs together with any remaining free atoms. The user can then visually analyze these MCD diagrams.

The total number of MCDs,  $n_{\text{MCD}}$ , depends on the following parameters which are defined by the user:  $l$ , number of found fragments which will be used for the creation of MCDs ( $l \leq L$ );  $n_f$ , the minimal number of fragments that must be present in each MCD;  $q$ , the minimum percentage of all skeletal atoms that must be absorbed by the fragments present in each MCD.

As noted above, in the general case, the more atoms from the fragments consumed by the MCD, the greater the likelihood that the process of structure generation from a given MCD will be more time efficient.

The speed of structure generation depends on the size of the molecular fragments. If the number of small fragments composing the MCD is large enough, then this will speed up structure generation. Generation is also much faster even when the MCD is comprised of only a small number of large fragments. Depending on the size of the molecule being analyzed and the size of fragments placed at the beginning of the ranked list of found fragments, the  $n_f$  value is usually defined as a number from 1 to 4. The most efficient results are obtained if  $q$  is significant, generally 40–60%.

The conclusion of all further verification procedures is a check of all produced MCDs for contradictions. The program offers an option that deletes all MCDs that are recognized as contradictory. The diagrams remaining after checking can be used in the structure generation process. The user has the opportunity to omit the connectivity verification because contradictory MCDs will in any case be detected and rejected in the process of structure generation. Moreover, for the process of structure generation, the user can select one or more MCDs that are attractive to the user who may have prior knowledge of a particular structure class or target structure. To alleviate having to choose a preferable MCD, they are automatically ranked in order of the increasing number of free carbons. In this way, it is possible to select a series of appropriate MCDs, starting from that ranked first.

The number of MCDs produced from a given set of fragments can be rather large with the  $n_{\text{MCD}}$  value sometimes being  $> 1000$ . To allow editing of a large set of MCDs, we developed a procedure which allows the transfer of all changes made in one MCD to all the MCDs contained within the set. In particular, it is possible to specify options which transfer atom coordinates, the display zoom factor, atom properties, manually drawn connectivities, and connectivities automatically modified during the process of checking the MCD for the presence of contradictions.<sup>9</sup> This procedure essentially alleviates using prior information in the fragment mode of structure elucidation.

In the process of analyzing a novel compound, it is entirely possible that there will be no fragments in the database that will reduce the magnitude of the challenge. It is natural in such cases to expect that the introduction of user-defined fragments may help to form the MCDs. The main qualitative difference between a found fragment (FF) and a user fragment (UF) is that the FF already contains carbon atoms with assigned chemical shifts, while the carbon atoms of the UF have no carbon chemical shift assignments. Two ways have been suggested to introduce user fragments into the program:

Calculate the carbon chemical shifts of the fragment using the “accurate” method (see section 2.2.4.);

Search the KB for fragments that *comprise* the user fragment.

It is likely that fragments from at least one of the two sources would be available for use by the program.

**2.2.3. Choice of “E” Value.** In the process of MCD creation from fragments, the *E* value is of great importance since it markedly influences the result of applying the fragments. There are a number of principles governing the selection of the *E* value. As a rule, the smaller the value of *E* then the smaller the number of molecular connectivity diagrams,  $n_{\text{MCD}}$ , created from found fragments. The advantage of a small number of MCDs is of course that it can reduce the time for structure generation,  $t_g$ . At the same time,  $t_g$  is also a function of the *fragment dimensions*. Larger fragments generally shorten the structure generation process. However, if a fragment is large and correspondingly contains many assigned carbon atoms, then as a consequence it is not as likely that *all* carbon atoms, especially the terminal ones, of a large fragment will fit the experimental shifts, thereby satisfying a narrow interval for  $\pm E$ . The program automatically sets the *E* value for terminal atoms equal to 12 ppm to account for this issue.

Large fragments are the most useful but to utilize them in the structure elucidation process a large *E* value is necessary. A large *E* value can correspondingly, in turn, increase the  $n_{\text{MCD}}$  value. The optimal approach would be to set a large enough *E* value and select only MCDs containing large fragments for the structure generation. This principle therefore justifies manual (user) or automatic rejection of MCDs containing small fragments. With testing, it has been shown that the optimal program parameter controlling the minimum number of carbon atoms in the fragment used for MCD creating should be set to a value of 5. Unfortunately, it is impossible to determine an optimal *E* that is valid for a diverse range of problems. The value of *E* should be optimized for each task by gradually increasing the *E* value starting from 0.5 ppm.

There is one more situation that requires relatively large *E* values. Experience has shown (see section 4.1.2) that the program sometimes accepts incorrect fragments as legitimate when the *E* value is small. Obviously if the magnitude of the error is sufficiently large, the program will select fragments of increased structural diversity.

**2.2.4. Selection of the Preferable Structure.** The *StrucEluc* system provides a three-step procedure for identifying the most probable structure in the output file.

*First Step.*  $^{13}\text{C}$  NMR spectra are predicted for all generated structures using an incremental method, our so-called “fast” method, and  $d_F$  values, the average deviation of an experimental  $^{13}\text{C}$  NMR spectrum (or the chemical shifts derived from corresponding projections of 2D NMR spectra), versus predicted chemical shifts, are calculated.

*Second Step.* Duplicate structures in the output file are deleted. Among the generated structures there are usually duplicates that differ from each other only in terms of the assignment of the chemical shifts to different carbon atoms. If this possibility is not appropriately considered when deleting isomorphous structures, then the structure with the correct assignment of the chemical shifts could conceivably be the deleted isomorphous structure. To avoid this eventuality, the system executes a special procedure for duplicate removal. For each duplicate family only the structure that

has the minimum  $d_F$  value is retained in the file as “the best representative” of the family. After duplicates are removed, the structures are then ranked by the  $d_F$  value and sorted in ascending order. The smallest  $d_F$  value indicates the best match between the experimental and calculated spectra, and this structure will therefore be the first in the list. Experience shows that for the fast calculation of  $^{13}\text{C}$  NMR spectra and their subsequent ranking places usually the correct structure is the first or second one in the list. Only in rare instances will the correct structure be listed below fifth place. Such a preliminary ranking of the big resulting files can help to reject hundreds and thousands of structures that are known to be unsuitable.

*Third Step.* During the third stage, more accurate  $^{13}\text{C}$  NMR spectra are calculated for the first 10–25 structures of the ranked file. Accurate predictions are performed using a database of fragments with their corresponding assigned subspectra. The description of each nuclear environment is defined using the HOSE code approach<sup>13</sup> (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 accurate proton chemical shifts can be predicted and displayed together with the corresponding deviation value  $d_H$ . If the experimental  $^1\text{H}$  NMR spectrum is not recorded, the chemical shifts are automatically found from 2D NMR projections. For proton NMR prediction, the predicted proton–proton couplings are enhanced by three-dimensional optimization of the structure. The position of the correct structure in the file determines its rank depending on the type of ranking parameter, *i.e.*,  $d_A$ ,  $d_F$ , or  $d_H$  correspondingly. The “rates” of the correct structure in the ranked file are denoted as  $r_A$ ,  $r_F$ , and  $r_H$ . If the correct structure is the first in the list ranked by  $d_A$  values, then  $r_A = 1$ . As a rule, the final structural ranking is carried out according to increasing  $d_A$  values, while magnitudes of the  $d_F$  and  $d_H$  parameters serve as secondary aids for estimating reliability of the correct structure selection. When the structures at the beginning of the ranked file possess different structural elements, prediction of the MS match factor ( $m_i$ , where *i* is the position of a structure in the ranked file) may also be useful for the confirmation of the preferable structure. The system utilizes a routine that is capable of calculating the percentage of peaks in the experimental MS spectrum that can be interpreted on the basis of a given structure. The calculation of the MS match factor is relatively time-consuming, so it is used only in cases when the difference  $\Delta_{(2-1)} = d_A(2) - d_A(1)$  is small. Here  $d_A(1)$  and  $d_A(2)$  represent the deviations corresponding to the first and second structures in the ranked file.

In ambiguous cases, it may be useful to display the calculated  $^1\text{H}$  NMR spectra because of the complexity of some multiplet patterns. Also, to facilitate structure analysis in the output file, the *StrucEluc* system is supplied with a feature that calculates structural similarity coefficients.<sup>14</sup> In this way, if the investigator has an idea of the class of structure under investigation this structure can be used as an input to allow rank ordering relative to the structural similarity of the results file.

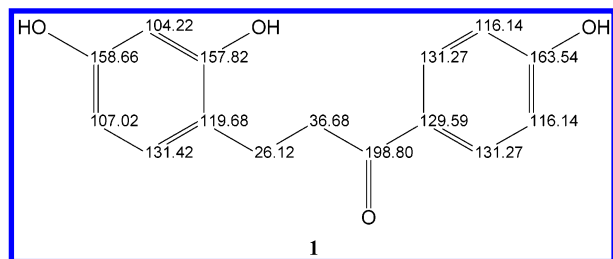


**Table 1.** NMR Data for Compound 1 in Pyridine-d<sup>5</sup>

| atom number | $\delta_C$ , multiplicity | $\delta_H$ | HMBC       |
|-------------|---------------------------|------------|------------|
| 1           | 119.68, s                 |            | 7.02, 6.79 |
| 2           | 157.82, s                 |            | 7.30       |
| 3           | 104.22, d                 | 7.02       | 6.79       |
| 4           | 158.66, s                 |            | 7.30       |
| 5           | 107.02, d                 | 6.79       | 7.02       |
| 6           | 131.42, d                 | 7.30       |            |
| 7           | 129.59, s                 |            | 7.16       |
| 8           | 131.27, d                 | 8.17       |            |
| 9           | 116.14, d                 | 7.16       |            |
| 10          | 163.54, s                 |            | 8.17       |
| 11          | 116.14, d                 | 7.16       |            |
| 12          | 131.27, d                 | 8.17       |            |
| 13          | 36.68, t                  | 3.54       | 3.43       |
| 14          | 26.12, t                  | 3.43       | 7.30, 3.54 |
| 15          | 198.80, s                 |            | 8.17, 3.43 |

### 3. EXAMPLES OF APPLYING DIFFERENT SYSTEM MODES

To demonstrate the main features of the approaches suggested above, a number of examples will be cited. Consider the elucidation of the chemical structure from the 1D and 2D NMR spectra of a rather simple molecule, 2,4,4'-trihydroxydihydrochalcone (**1**), recently isolated and characterized by Gonzales and coauthors.<sup>15</sup> This example allows demonstration of all methods of structure elucidation currently available in *StrucEluc*. A molecular formula of C<sub>15</sub>H<sub>14</sub>O<sub>4</sub> was deduced from the high-resolution MS spectrum. The NMR spectral data (<sup>1</sup>H and <sup>13</sup>C NMR, DEPT, HMBC) obtained in ref 15 and used here are represented in Table 1. For the explanations of the analysis to be clear, the target structures of the examples described in this article will be displayed, if necessary, with already assigned carbon chemical shifts.



**3.1. Examination of 1D NMR Spectra. 3.1.1. "Standard" Mode.** When the *Search Fragments by CNMR Spectrum* command was performed (see Scheme 2), the program picked out  $L = 180$  fragments and ranked them according to decreasing molecular formula. The first 20 fragments from the list are shown in Figure 1. Note that all carbon atoms of the fragments stored in the database are supplied with <sup>13</sup>C NMR assignments taken from the full structures that were used to create the fragment DB.

The comparison of fragments with structure **1** shows that fragments #1 and #20 exist in the molecule under investigation and have one overlapping group which is a CH<sub>2</sub>. *Standard* structure generation accompanied by structure filtering with spectral libraries was performed. This is the application of the first system branch. The program combined the found fragments using at least one *overlapping* atom. As a result, the program produced only one structure coinciding with compound **1** with a structure generation time of  $t_g = 45$  s (in this article, all calculations are performed

on an PC Intel Celeron operating at 500 MHz, Windows 98, RAM 128 Mb). The single structure was actually generated in 4 s, while the remaining time was consumed by attempts to assemble other conceivable structures from the first 30 fragments, this number being the default for the assembly process. A single correct structure was obtained from the 1D <sup>13</sup>C NMR spectrum in fully automatic mode.

**3.1.2. "Classical" Mode.** An attempt was made to apply the second system branch to the data utilizing the algorithms in the so-called "classical" generation mode (see Scheme 2) based on combining *nonoverlapping* fragments and free atoms. According to the methodology described previously, the program automatically selected the first 20 found fragments ( $l = 20$  is the default setting in this case) from the list and then created 223 fragment sets. Classical structure generation was performed, and the spectral libraries and internal BADLIST were used for structure filtering during the generation process. As a result, the number of structures generated,  $k$ , gave 112 resultant molecules, while the generation time,  $t_g$ , was 1 min 30 s. After the removal of identical structures, the number of structures was reduced to 73. This is denoted by the representation of  $k = 112 \rightarrow 73$  and will be used throughout this article. The <sup>13</sup>C NMR prediction and structure ranking allowed the program to reliably distinguish the correct structure as the first structure. The difference between the first and second structure is given by the difference between  $d_A(1) = 1.64$  ppm and  $d_A(2) = 3.98$  ppm which gives  $\Delta_{(2-1)} = d_A(2) - d_A(1) = 2.34$  ppm. Further details regarding the process of evaluation of the structure reliability will be discussed in section 7.

**3.2. Application of 2D NMR Data Analysis. 3.2.1. The "Common" 2D NMR Mode.** The possibility of elucidating an unknown structure from the 2D NMR data in an automated or semiautomated manner is certainly attractive. In this case, the user would not need to make any assumptions that would serve as constraints for the system. However, to demonstrate the role of different structural constraints the two common mode methods (see Scheme 3) of solving a given problem will be considered: without user intervention and with the influence of the user by inclusion of additional information.

**3.2.1.1. Solution Without Any User Assumptions.** As described earlier, the program automatically created and displayed a molecular connectivity diagram (MCD). The MCD reflecting the HMBC data from Table 1 is shown in Figure 2.

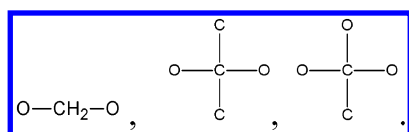
The atom properties were automatically determined for all carbon atoms by the program using the APCT table (see section 2.2.1). In this example, for the carbon atoms of the benzene rings the parameters of hybridization and atom neighborhood with attached heteroatoms were automatically set to *sp<sup>2</sup>/not defined* and for the carbon with a chemical shift of 198.8 ppm atom the values were set to *sp<sup>2</sup>/at least one*. As a result, the number of structures generated and filtered was  $k = 32 \rightarrow 4$ , with  $t_g = 26$  min. The relatively large  $t_g$  value is explained by utilization of HMBC data only since the COSY spectrum was not cited by the authors of ref 15.

The structures were ranked using the methodology described earlier. The first structure was characterized by  $d_A(1) = 1.51$  ppm,  $d_F(1) = 1.74$  ppm, and  $d_H(1) = 0.42$  ppm and was identical to compound **1**. The correctness of the

structure determination was corroborated by both fast and accurate spectrum predictions, *i.e.*,  $r_A = r_F = 1$ . The reliability of the structure selection was also confirmed by the large difference  $\Delta_{(2-1)} = 4.9$  ppm. It is worth noting that the chemical shift assignments suggested by the program coincided with the assignments performed by the authors in ref 15.

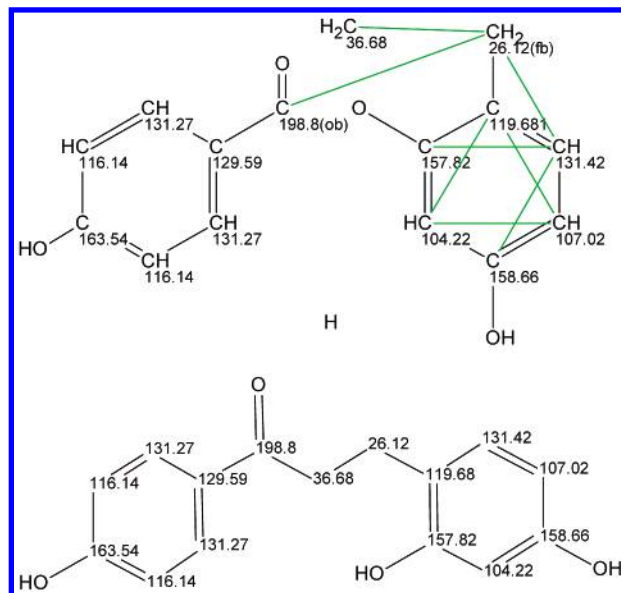
**3.2.1.2. Imposing Additional Constraints.** One manner by which to formulate additional structural restrictions is to search for functional groups in the fragments selected from the database. This may enable the formation of a “generalized portrait” of the molecule. When the command *Search Functional Groups* is applied, the program sorts the Typical Functional Group Library described earlier through the structures of the *L* found fragments. As a result, both the GOODLIST and the BADLIST can be created in an automated fashion, and the functional groups identified are available for viewing on the screen.

In this example, the following oxygen-containing groups were created in the BADLIST:



This indicates that it is highly unlikely that carbon atoms having the *sp<sup>3</sup>/at least two* property are present in the molecule which is under consideration. With this in mind, for the second run the following additional constraints were imposed: the properties of carbon atoms having chemical shifts between 104.22 and 131.42 ppm were replaced with (*sp<sup>2</sup>/forbidden*). For the three carbon atoms with chemical shifts observed between 157.82 and 163.54 ppm, the properties were set to *sp<sup>2</sup>/at least one*. A C=O bond was drawn manually for the carbon at 198.8 ppm. Following structure generation,  $k = 16 \rightarrow 4$  and  $t_g = 8$  s and the same output file was obtained. When these additional assumptions were employed,  $t_g$  decreased from 26 min to 8 s, a change of a factor of 200. The alleviation of the difficulty of solving the problem by increasing user intervention suffers from a shortcoming. If at least one of the user-defined constraints is erroneous, the solution obtained will be incorrect and the structural output file will not contain the right molecule. Even with this shortcoming, the investigator's knowledge, experience, and intuition should be used as much as possible to facilitate the elucidation process.

The structure elucidation process belongs to the class of “inverse problems”<sup>16</sup> and the chance of fully replacing human intellect is unlikely at best. Moreover, in accordance with the Bohr principle of complementarity, the methodology of computer assisted structure elucidation includes two major elements that complement each other. They are the deterministic logic of the computer and the knowledge and intuition of the investigator. The interaction of these elements in the process of solving the problem is what gives rise to the synergistic effect that allows the elucidation of complex molecules. It is therefore necessary to find a rational way of combining connectivities deduced from the experimental 2D NMR data with additional information from a scientist to obtain a solution to the problem in a reasonable time. Certainly employing the fragments from the database can

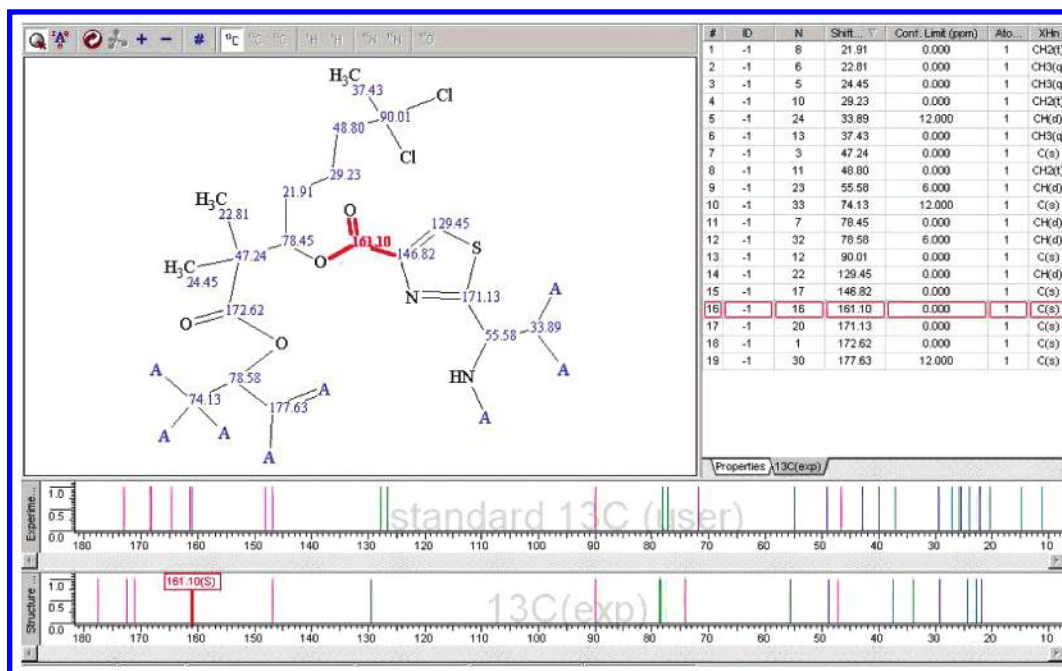


**Figure 3.** Molecular connectivity diagram created from the fragments obtained for compound **1**. Compare chemical shifts assignments of the found fragments with those of the “unknown” structure displayed below.

be used to minimize user intervention and, consequently, the risk of obtaining an incorrect solution.

**3.2.2. Utilization of Fragments in the KB.** Consider again the process of structure elucidation of the 2,4,4'-trihydroxy-dihydrochalcone molecule using fragments stored in the database (see Scheme 3). As shown above,  $L = 180$  and the first 20 fragments in the list are shown in Figure 1. In this example, the following options for creation of the MCD were specified:  $l = L = 182$ ;  $E = 0.5$  ppm;  $n_f = 2$ ;  $q = 50\%$ . The  $n_{\text{MCD}}$  value was equal to 6. To illustrate the mechanism of fragment combination, Figure 3 displays one of the created MCDs and the target structure.

Two “good” fragments existing in the fragment list form a combination that absorbs all skeletal atoms of the molecule. This is a fortuitous situation. After creation of the MCDs, the *Check MCDs* command was executed and indicated that all MCDs are consistent. It should be noted that the final resulting structure with the correct chemical shift assignment can be generated from a combination of these fragments only. The result is that the resulting structure has carbon assignments that coincide with those intrinsic to the molecule under investigation. Structure generation was performed for all six MCDs. During this process *no constraints* were entered for the atom properties. The number of structures generated,  $k$ , was 2 and  $k = 2 \rightarrow 1$  with  $t_g = 1$  s. If the generation is performed directly from the initial 6 MCDs, *without* initially checking the MCDs, the  $t_g$  value is the same. This indicates that it is advisable to start structure generation from the initial set of MCDs and ignore preliminary checking since this does consume some time. There have been examples in this work where the checking time is significant. As shown, the application of fragments drastically shortened the generation time. Compared to the first run performed in the common mode, the reduction in time is about 1500 times. Obviously, in this example, the application of found fragments indeed obviated any user intervention in the solution process and provided a dramatic reduction in the processing time.

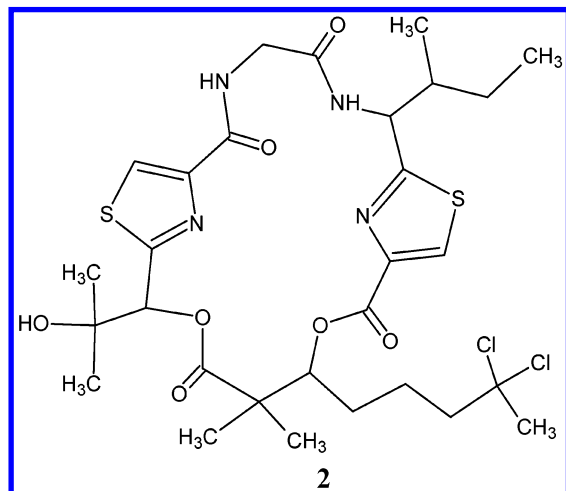


**Figure 4.** Fragment #1 in the found fragments list. The experimental  $^{13}\text{C}$  NMR spectrum of compound 2 is displayed in the upper window and the fragment subspectrum in the lower window.

#### 4. SOLVING 2D NMR PROBLEMS USING FRAGMENTS FROM THE SYSTEM KB

During the process of testing the *StrucEluc-2* system, a number of problems have been encountered where the program failed to solve using the common mode, despite the absence of any contradictions in the 2D NMR data. These cases will be used as examples to illustrate the strategy of utilizing fragments for solving these challenging problems. Experimental data were obtained from literature articles where the structures analyzed were absent from the full structure library of the system.

**4.1. Lyngbyabellin A.** An attempt to automatically determine the structure of *lyngbyabellin A* (**2**), a novel cytotoxic compound isolated and investigated most recently by Luesch et al.,<sup>17</sup> was not successful. Compound **2** has a molecular formula of  $\text{C}_{29}\text{H}_{40}\text{Cl}_2\text{N}_4\text{O}_7\text{S}_2$  and a molecular mass of  $M = 690$ . In this example, 2D NMR data obtained from COSY, HMQC, and HMBC experiments as well as a list of IR vibrations were obtained from the original work.<sup>17</sup>



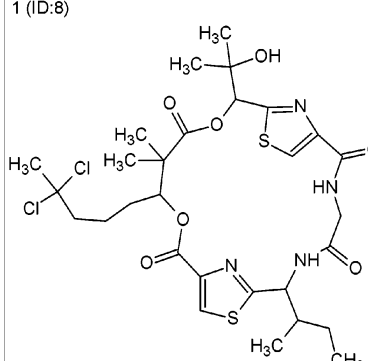
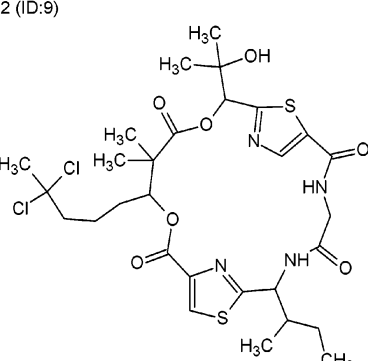
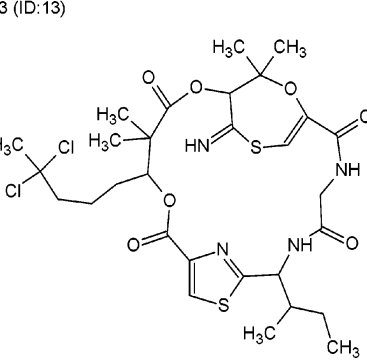
**4.1.1. "Common" Mode.** The molecule under investigation is fairly complex. It contains four types of non-carbon skeletal atoms, N, O, S, and Cl to give a total of 15 heteroatoms and two thiazole rings. The presence of two sulfur atoms implies atom properties of  $sp^3/nd$ ,  $sp^2/nd$ . Sulfur can exist in different valence states and as a result can introduce a large number of possible isomers during structure generation.

Several attempts were made to solve this problem. Many different forms of user intervention were given, including specifying atom properties and drawing  $\text{C}=\text{O}$  bonds. The structure generation process was aborted after generating more than 150 000 structures within a 24-h period. With this in mind, it was concluded that only the application of the fragment approach could be helpful for solving the problem.

**4.1.2. Utilization of Fragments from the KB.** For this example, the number of found fragments was fairly large:  $L = 7427$ . Initial viewing of the found fragments indicated that the fragments listed at the start of the file had  $^{13}\text{C}$  chemical shifts that were very close to those observed experimentally. For the first fragment, the general pattern of both the experimental and predicted spectra appeared similar when they were visually compared (see Figure 4). In addition, this fragment contained 19 carbon atoms and a total of 28 skeletal atoms. It was decided that these coincidences were likely not serendipitous. It was natural to suggest that this and related fragments could be related to the structure under investigation.

With these assumptions, the process of creating connectivities from the first 25 fragments ( $l = 25$ ) was initiated using  $q = 60\%$ ,  $n_f = 1$ , and an initial  $E$  value of 0.5 ppm. The program created no MCD for  $E$  values lying in the interval 0.5–9 ppm. 144 diagrams were created using  $E = 10$  ppm. The absence of any vibrational frequency in the region near  $2600\text{ cm}^{-1}$  as well as the  $^1\text{H}$  NMR spectrum suggested that an SH group be introduced into the user



|  |  |   |
|--|--|---|
| <p>1 (ID:8)</p>  <p> <math>d_A(^{13}\text{C})</math>: 3.077 (4.673)<br/> <math>d_F(^{13}\text{C})</math>: 3.836 (5.854)<br/> <math>d_A(^1\text{H})</math>: 0.277 (0.430)<br/>           Formula <math>\text{C}_{29}\text{H}_{40}\text{Cl}_2\text{N}_4\text{O}_7\text{S}_2</math><br/>           FW: 691.6881<br/>           SimCoeff 1.000         </p> | <p>2 (ID:9)</p>  <p> <math>d_A(^{13}\text{C})</math>: 4.383 (8.004)<br/> <math>d_F(^{13}\text{C})</math>: 4.564 (7.337)<br/> <math>d_A(^1\text{H})</math>: 0.232 (0.297)<br/>           Formula <math>\text{C}_{29}\text{H}_{40}\text{Cl}_2\text{N}_4\text{O}_7\text{S}_2</math><br/>           FW: 691.6881<br/>           SimCoeff 0.957         </p> | <p>3 (ID:13)</p>  <p> <math>d_A(^{13}\text{C})</math>: 4.667 (8.299)<br/> <math>d_F(^{13}\text{C})</math>: 5.093 (8.112)<br/> <math>d_A(^1\text{H})</math>: 0.477 (0.908)<br/>           Formula <math>\text{C}_{29}\text{H}_{40}\text{Cl}_2\text{N}_4\text{O}_7\text{S}_2</math><br/>           FW: 691.6881<br/>           SimCoeff 0.689         </p> |
|--|--|---|

**Figure 5.** The elucidated structures most similar to the correct structure. Those displayed are selected from the beginning of the ranked answer file and ranked in decreasing  $C_{\text{sim}}$  values.

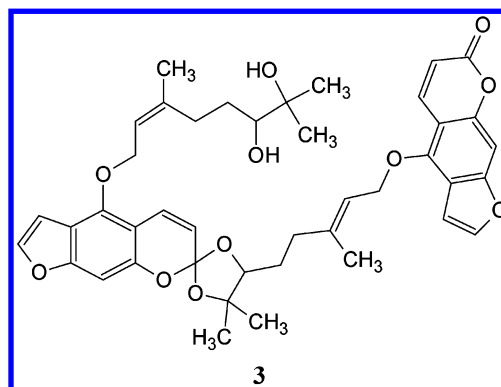
BADLIST. Structure generation was performed, but no structures were produced from these MCDs.

Successively increasing the  $E$  value to  $E = 14$  ppm produced 272 MCDs when 128 new MCDs were produced. This indicates that 272 different distributions of the carbon assignments are possible at this given error value ( $E$ ). Only 32 MCDs survived a check for contradictions. Eight of these contained the first fragment, and the others were composed of two fragments. Structure generation was performed using a single constraint—four-membered rings, which are uncommon in natural products, were forbidden. The results gave  $k = 20 \rightarrow 20$  and  $t_g = 2$  min. The correct structure has minimum deviations for all spectral predictions:  $d_A(1) = 3.08$  ppm,  $d_F(1) = 3.84$  ppm, and  $d_H(1) = 0.28$  ppm, while  $r_A = r_F = r_H = 1$  and the difference  $\Delta_{(2-1)} = d_A(2) - d_A(1) = 1.3$  ppm. The  $\Delta_{(2-1)} = 1.3$  ppm value indicates the reliability of the solution as described in detail in section 7.

When structure generation was repeated without any limitation on ring cycle sizes, 25 structures were generated, and the spectrum prediction again endorsed the priority of the correct structure. To determine the similarity of the generated structures to the real structure, we calculated Tanimoto coefficients for structural similarity.<sup>14</sup> The structures were ranked in decreasing  $C_{\text{sim}}$  values, and the first three are shown in Figure 5. The structure most similar to the correct one gave  $C_{\text{sim}} = 0.96$  and was the second in the ranked output file in accordance with the spectral data. Figure 5 directly demonstrates the high resolving power of the method based on accurate  $^{13}\text{C}$  NMR spectrum prediction.

It was concluded that a visual comparison of the  $^{13}\text{C}$  NMR spectrum of large fragments ranked by the program at the top of the list of FFs with the experimental spectrum can aid in the choice of fragments suitable for creating the MCDs. At present, this approach may be the only possible way to solve a difficult problem without making risky assumptions.

**4.2. Paradisin C.** Reference 18 details the isolation and identification of a new natural compound, *paradisins C* (**3**). The elucidation was based on high-resolution FAB-MS data which gave a molecular formula of  $\text{C}_{42}\text{H}_{46}\text{O}_{11}$  and using a series of 2D NMR spectra (COSY, HMQC, and HMBC).



An attempt to solve the problem in the “common” mode failed. The structure generation process continued for over 24 h and was aborted at that time. Searching through the knowledge base using the  $^{13}\text{C}$  NMR spectrum as the input gave 2946 fragments. As in the previous example it was determined based on visual comparison that the carbon chemical shifts of the first fragments in the list were close to the experimental values. This provided a basis to set the conditions for forming the MCDs which would lead to the consumption of the “free” skeletal atoms. MCDs were formed using  $l = 500$ . With  $E = 5$  ppm,  $q = 40\%$ , and  $n_f = 3$ , the program created  $n_{\text{(MCD)}} = 384$  MCDs. The process of structure generation without any additional restrictions resulted in  $k = 15$ ,  $t_g = 2$  min 30 s,  $d_A(1) = 2.10$  ppm,  $d_F(1) = 2.47$  ppm,  $d_H(1) = 0.14$ ,  $\Delta_{(2-1)} = 0.76$  ppm, and  $r_A = r_F = r_H = 1$ .

These examples indicate that using fragments found in the database allows the user to solve problems that seemed to be impossible for the system to cope within the “common” mode.

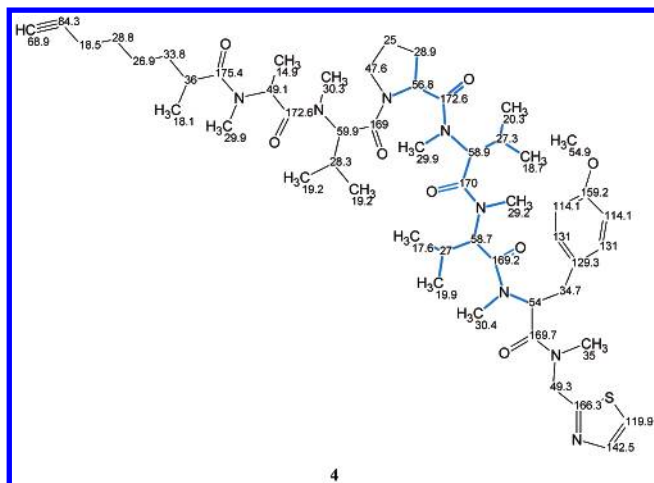
## 5. UTILIZATION OF BOTH USER AND FOUND FRAGMENTS

Using fragments from the KB often leads to solution of the problem, but unfortunately the KB is restricted to published data and may not include suitable fragments



specific to proprietary chemistries. In this section, two examples related to the structure elucidation of peptides will be considered that could not be solved by utilizing fragments found in the KB. In this case, it was demonstrated that these problems could be solved with the help of additional data including user fragments.

**5.1. Apramide G.** Luesch et al.<sup>19</sup> extracted and identified a new metabolite *apramide G* (**4**) whose molecular mass was determined as  $M = 976$  and a molecular formula of  $C_{52}H_{80}N_8O_8S$ . In the publication,<sup>19</sup> tables of data are listed that contain  $^1H$  and  $^{13}C$  NMR spectra and 2D NMR correlations from COSY, HMQC, and HMBC. Various methods of solving this problem were investigated. An initial attempt to solve the problem in the “common” mode indicated that the processing time would be excessive. The structure contains a thiazole ring and therefore a sulfur atom with all of its valence issues. The structure also contains a  $C\equiv C$  group. If the maximum bond multiplicity is set to 3 during the structure generation process, a large number of structures result and the generation time correspondingly increases. For the first attempt at elucidation the structural data obtained by the authors from the spectrum analysis were used.

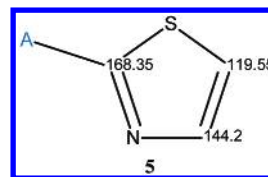


**5.1.1. “Common” Mode.** Checking the MCD revealed that there were no nonstandard connectivities. The authors had found that for the fragment  $CH_2-C\equiv CH$ , one COSY correlation corresponded to  $^4J_{HH}$ . This was taken into account when creating the MCD. Attempts were made to solve the problem in the “common” mode using the data obtained by the authors:<sup>19</sup> the molecule contained six  $NCH_3$  groups (6 singlets in the  $^1H$  NMR spectrum) and one  $CH_2-C\equiv CH$  group (a triplet in the  $^1H$  NMR spectrum at 1.77 ppm with a coupling constant of  $^4J_{HH} = 2.4$  Hz provided evidence for the presence of the terminal acetylene group). A correlation between  $\delta_H(1.77$  ppm) and  $\delta_C(68.9$  ppm) suggested  $sp$  hybridization. Instead of the  $nd/nd$  property, this atom and the quaternary  $\delta_C(84.3$  ppm) were manually attributed with the  $sp/fb$  parameter.

After 16 h of operation in the “common” mode, the program did not generate any structures that complied with the structural and spectral filters. This result was not unexpected. The formal number of double bonds in the structure under investigation is 17, and the molecule is proton deficient. However, the main difficulty here is the size of the molecule, 69 skeletal atoms, and the fact that the molecule

contains a significant number of heteroatoms including a sulfur atom.

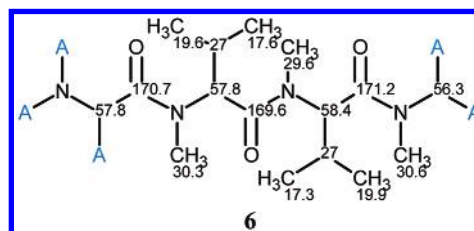
**5.1.2. Utilization of User Fragments.** The authors<sup>19</sup> identified from the experimental data the presence of a thiazole ring. This is supported by 2 doublets in the  $^1H$  NMR spectrum at 6.56 and 7.45 ppm ( $J = 3.2$  Hz) though this could just as easily have been a furan ring based on these data. With this information, a thiazole fragment, **5**, was entered as a User Fragment. The chemical shifts shown in structure **5** were calculated using the accurate method and MCD creation was initiated.

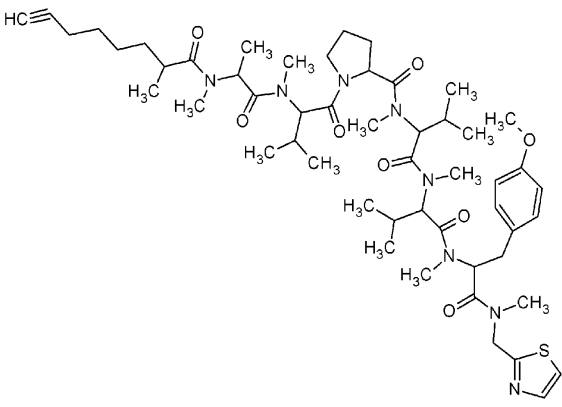
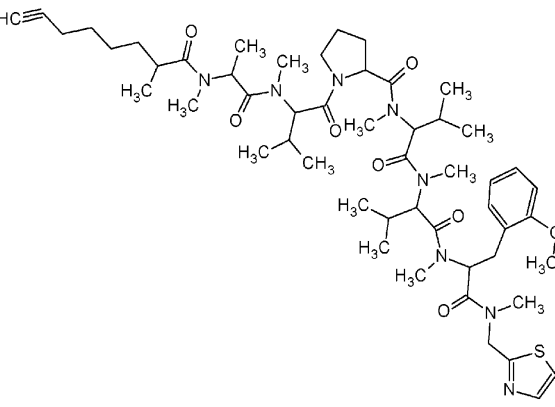


With  $E = 2.5$  ppm, one MCD was created containing this fragment. Bonds were manually included from the six  $CH_3$  groups to the nitrogen atoms according to the information provided by the authors. The properties of these C atoms were adjusted to  $sp$  hybridization. These restrictions were still insufficient to solve the problem, and the generation process was aborted after 2 h. It was clear that the process would be time-consuming since over 7000 structures were generated in that time. An attempt was therefore made to use fragments from the system knowledge base.

**5.1.3. Utilization of Found Fragments.** After a fragment search in the database the number of selected fragments was  $L = 13\,934$ . To prune the fragments that are unlikely and to move prospective fragments to the top of the list, OH and NH groups were placed into the BADLIST since there were no absorptions in the IR spectrum above  $3000\text{ cm}^{-1}$ . The fragments were then filtered with simultaneous exclusion of ionic structures since the system does not yet allow ionic structures. The result gave  $L = 13\,934 \rightarrow 5121$ , and the number of fragments was reduced by a factor of almost 3. Visual investigation of fragments at the start of the list revealed that there were some fragments that showed good correspondence with the experimental spectrum.

The first 10 fragments were selected from the list to form the MCDs. To confirm whether the problem could be solved with only the found fragments, the thiazole ring was deleted from the User Fragments list before starting the MCD generation process. The following options were used for creation of the MCDs:  $l = 10$ ,  $E = 2.5$ ,  $n_f = 1$ , and  $q = 30\%$ . The program created 12 MCDs, and each contained fragment **6** but with different distributions of the chemical shifts. In structure **4** this fragment is highlighted with blue bold lines. Compare the chemical shifts of the structure and the fragment and obvious similarities exist.



|   |   |
|---|---|
| 1 (ID:3030)   | 2 (ID:3364)   |
|    |   |
| $d_A(^{13}\text{C})$ : 1.128 (1.525)<br>$d_F(^{13}\text{C})$ : 2.349 (3.293)<br>$d_A(^1\text{H})$ : 0.272 (0.366)<br>Formula $\text{C}_{52}\text{H}_{80}\text{N}_8\text{O}_8\text{S}$<br>FW: 977.3067 | $d_A(^{13}\text{C})$ : 1.538 (2.088)<br>$d_F(^{13}\text{C})$ : 2.610 (3.585)<br>$d_A(^1\text{H})$ : 0.269 (0.364)<br>Formula $\text{C}_{52}\text{H}_{80}\text{N}_8\text{O}_8\text{S}$<br>FW: 977.3067 |

**Figure 6.** The two top structures from the answer file (*Apramide G*).

Before structure generation, atom properties were adjusted in the first MCD and automatically transferred to other diagrams. After 3 h of structure generation, the process was stopped for the same reasons as described in the previous step. To solve the problem in a reasonable time, user fragments, UF, in combination with found fragments, FF, were considered.

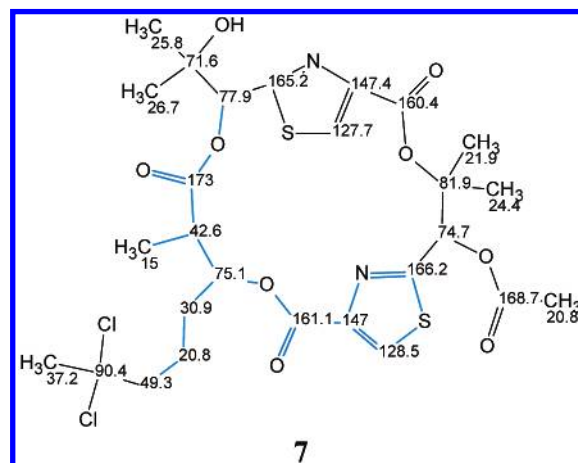
**5.1.4. Combining User and Found Fragments.** With settings of  $E = 2.5$ ,  $n_f = 2$ , and  $q = 30\%$ , the program created 12 MCDs, each containing the user fragment **5** and fragment **6**. Structures were then generated with one restriction only, that is a ring cycle size of  $R_c = 5-6$ . The results are  $k = 4991 \rightarrow 2143$ ,  $t_g = 31$  min,  $d_A(1) = 1.13$  ppm,  $d_F(1) = 2.35$  ppm,  $d_H(1) = 0.27$ ,  $\Delta_{(2-1)} = 0.41$  ppm, and  $r_A = r_F = r_H = 1$ . The first two structures are shown in Figure 6.

The second structure differs from the first only by the position of the substituent on the benzene ring. It can be concluded that the usage of both user fragments and found fragments in the database can be combined to solve the problem. The unresolved question is obviously how much information can or should be extracted by a scientist to feed into the program initially?

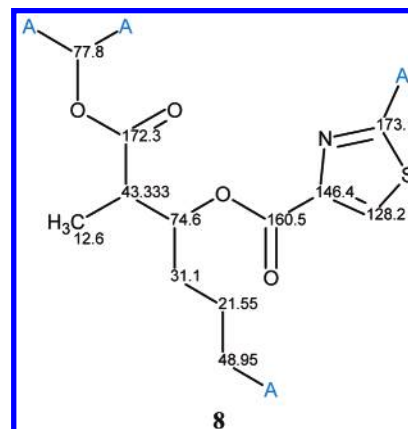
**5.2. Hetcochlorin.** A similar approach allowed the elucidation of *hetcochlorin* (**7**), a novel natural product of molecular formula  $\text{C}_{27}\text{H}_{34}\text{Cl}_2\text{N}_2\text{O}_9\text{S}_2$  isolated and characterized in the report by Marquez and co-workers.<sup>20</sup> The spectral data included HMBC data with 53 connectivities and COSY with 4 connectivities.

The molecule contains 2 thiazole rings and based on earlier experience it was evident that the structure could likely only be elucidated with the help of user fragments and fragments found in the database. The authors<sup>20</sup> detected the presence of two thiazole rings and four carbonyls from the 1D  $^1\text{H}$  and  $^{13}\text{C}$  NMR spectra.

A search through the knowledge base gave  $L = 3401$ . Visual comparison of the chemical shifts of the carbon atoms



of the first fragment in the list, substructure **8**, with the experimental values indicated they were similar. This is evident by comparing structure **7** and fragment **8** as highlighted by blue lines.

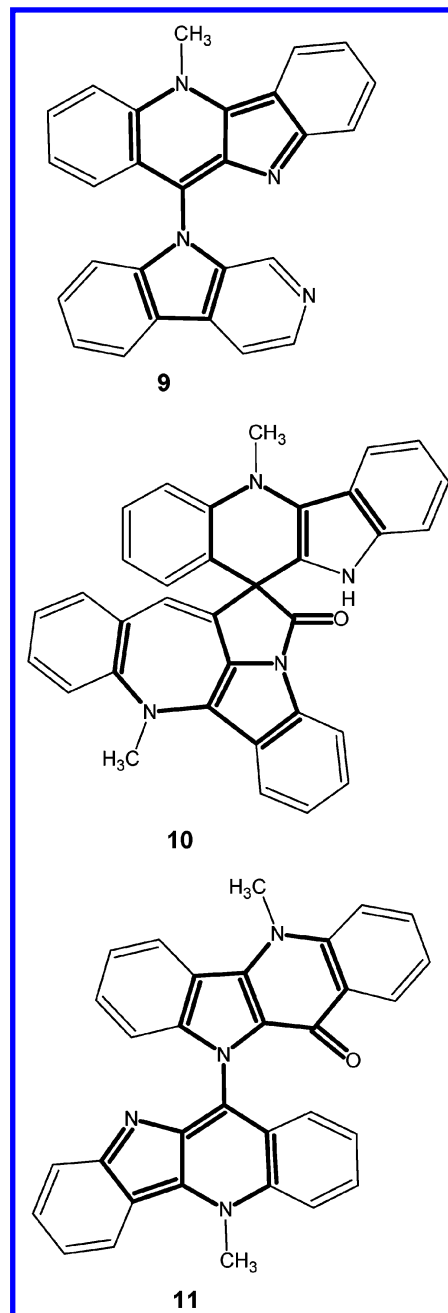


To complete the investigation, two MCDs were obtained from this fragment with  $E = 12$  ppm. Attempts to generate structures from these MCDs failed. When the number of generated structures exceeded 8000, the structure generation process was aborted. It was evident that the usage of user fragments would again be necessary. Since fragment **8** already contained one thiazole ring, an additional thiazole ring was introduced prior to creation of the MCDs. The program created 20 MCDs with settings defined as  $E = 12$  ppm,  $q = 50\%$ , and  $n_f = 2$ . Structure generation without any restrictions resulted in the following:  $k = 8$ ,  $t_g = 15$  s,  $d_A(1) = 2.73$  ppm,  $d_F(1) = 3.54$  ppm,  $d_H(1) = 0.29$ ,  $\Delta_{(2-1)} = 0.60$  ppm, and  $r_A = r_F = 1$ ,  $r_H = 3$ . This approach based on the usage of fragments from various sources allowed the elucidation of a complex molecule when the utilization of both the common mode and fragments from the knowledge base failed.

## 6. UTILIZATION OF THE USER DATABASE

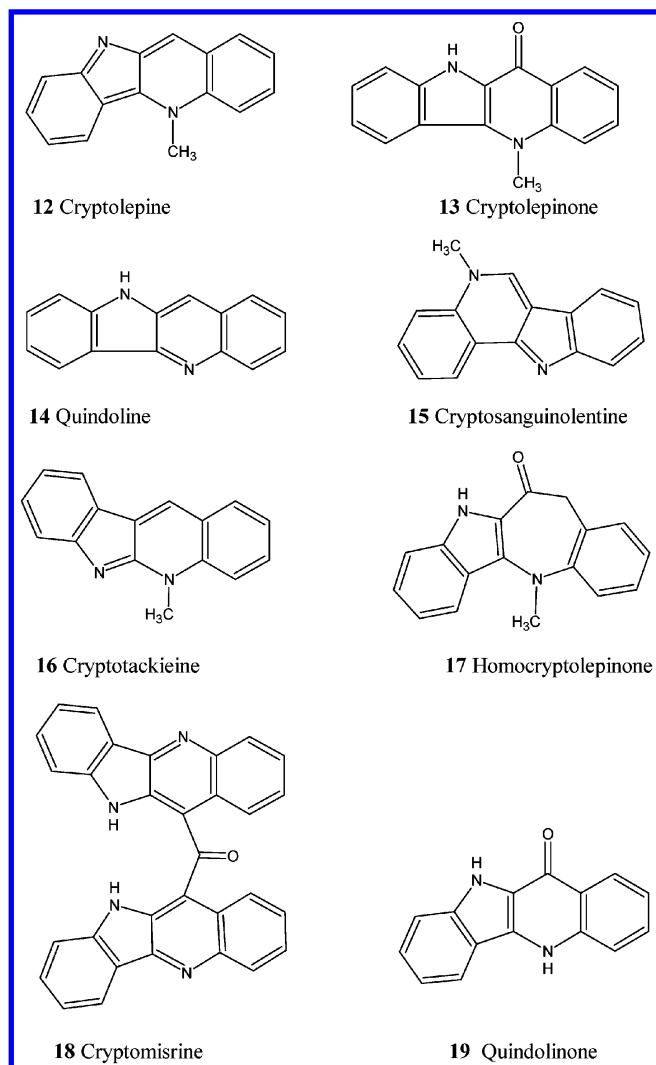
Currently, about 20 million compounds have been identified and well over a quarter million new compounds are synthesized or isolated each year. It is possible that many such compounds will have no analogues in the knowledge bases of expert systems. As a result, it is not always possible to find fragments in the database that will help to elucidate the structure of a compound from a new class. Besides, it is common that an investigator is unable to specify a fragment that may be present in the structure under examination. Our investigation has shown that if the methods described above are ineffective, then the creation of a user database could enable a solution. The *StrucEluc* system provides algorithms and capabilities to create user databases and thereby allow searching for fragments of related compounds. In particular, even if only one compound of a similar structure is known, it can be successfully used for the creation of a user database. With the help of user databases, the system can easily be adjusted for the elucidation of compound classes that are commonly investigated by a given laboratory. Investigations have shown that the failure to utilize library fragments was most frequently due to the following issues: (a) the fragments appropriate for a given problem are missing in the knowledge base; (b) appropriate fragments are found but the number of possible permutations of the carbon atom assignments in these fragments is so huge that the structure generation process is too long. In previous reports,<sup>21,22</sup> the structure elucidation of the three alkaloids from the cryptolepine series, cryptolepicarboline (**9**), cryptospirolepine (**10**), and 5,5'-dimethyl-5'H-10,11'-biindolo[3,2-b]quinolin-11(5H)-one (**11**), was considered. The structures of these natural products which failed to be elucidated both in the common and found fragment modes are shown below.

These molecules are relatively large, highly unsaturated, and have large fragments (displayed in bold) that contain no hydrogen atoms, thereby preventing access to structural information through COSY correlations. In particular, for structures **9–11** COSY correlations are observed for protons contained within the benzene rings only. Cryptospirolepine **10** has an especially complex structure that consists of two planar fragments lying in perpendicular planes joined via a spiro carbon, while each fragment makes up a system of conjugated bonds. Long-range “nonstandard” correlations



( $^nJ_{CH}$ ,  $^nJ_{HH}$ ,  $n \geq 4$ ) inside these planar fragments are likely. These factors as well as some unusual spectral properties all contribute to a very challenging elucidation process for cryptospirolepine. The structure determination of compound **11**, a degradation product of cryptospirolepine that contains a large “silent” fragment encompassing 19 skeletal atoms, is also a complex analytical problem, one whose solution has been comprehensively described previously.<sup>21</sup>

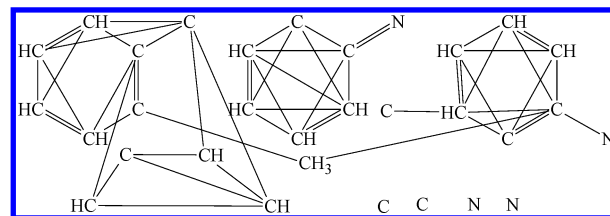
As mentioned earlier, the *StrucEluc* system contains algorithms and program features that allow the user to create their own knowledge base focused on chemical classes that are of particular interest to a given researcher. In this way, fragments are excised from these structures using proprietary algorithms. The carbon atoms of the fragments and molecules are attributed as usual to the corresponding chemical shifts in the  $^{13}C$  NMR subspectra. The system can therefore be adjusted to account for the elucidation of new classes of compounds.



**Figure 7.** Compounds from cryptolepine series that were used for creating a User Fragment Library.

The elucidation of the structures of alkaloids **9–11** was approached using a user database adjusted for cryptolepine structure analysis. The main assumption was that unknown compounds were members of the cryptolepine indoloquinoline alkaloid series since all materials were isolated from the same plant. Taking this into account, information regarding earlier published members of the series shown in Figure 7 was introduced into a user database. Assigned spectral data referring to these compounds were obtained from refs 23–29. This methodology is fairly typical in practice and is frequently used by chemists when manually determining the chemical structures of natural products or reaction products from a synthesis.

**6.1. Structure Determination of Compounds of the Cryptolepine Series.** Initially  $^{13}\text{C}$  NMR spectra were input and the *StrucEluc* library was searched. Molecules **12–16** were found in the system knowledge base. Since 2D NMR data for compounds **15–19** were available, these structures were used to challenge the system. An attempt to determine the structures using the common mode proved successful for all of the compounds. All were elucidated in the automated mode without operator intervention being necessary. The generation time for structures **15–19** was in all cases less than 20 s.



**Figure 8.** One of the MCDs created during cryptolepicarboline elucidation.

Compounds **12–19** of the cryptolepine series were incorporated in the user library as full structures along with their associated  $^{13}\text{C}$  NMR spectra. The algorithm used to create a user fragment library is similar to the algorithm used to create the *StrucEluc* system knowledge base<sup>1</sup> and includes the following two basic steps: (1) the program excises as complete as possible a set of fragments from all structures included in the structural file; and (2) atoms in the fragments are assigned chemical shift values that they have in the corresponding full structure. This procedure produced a user library containing 342 fragments from the cryptolepine series. This user database was then used to elucidate structures **9–11**.

**6.1.1. Cryptolepicarboline.** Searching for the  $^{13}\text{C}$  NMR spectrum of cryptolepicarboline through the user fragment library yielded 68 fragments. For MCD creation, the following options were specified:  $E = 1.5$  ppm,  $n_f = 3$ , and  $q = 50\%$ . 50 MCDs were created, and 25 survived the test for contradictions. One of the MCDs shown below in Figure 8 displays three fragments accounting for about 70% of the skeletal atoms:

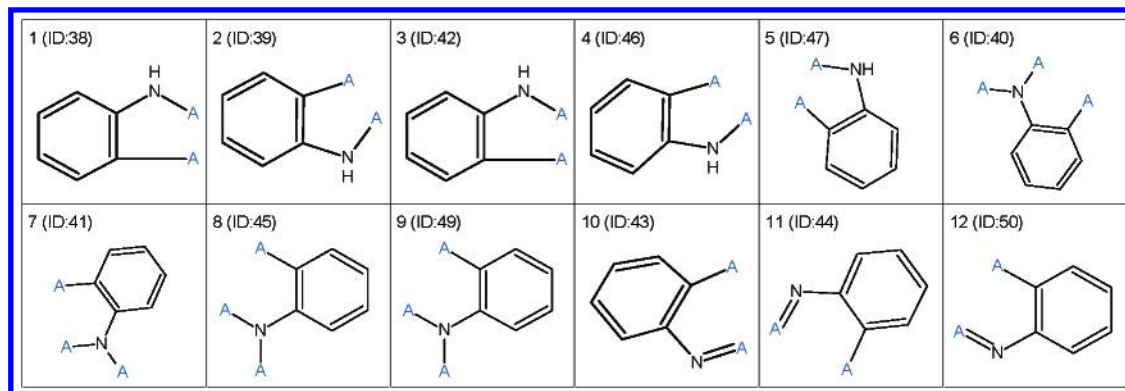
Structure generation using a ring size constraint of  $R_c = 5–7$  resulted in the following:  $k = 32 \rightarrow 13$ ,  $t = 10$  s,  $\Delta_{2-1} = 3.3$  ppm, and  $r_A = r_F = r_H = 1$ .

**6.1.2. Cryptospirolepine.** Searching for the  $^{13}\text{C}$  NMR spectrum of cryptospirolepine in the user fragment library yielded the following results:  $L = 60$ ,  $l = 60$ ,  $E = 4$  ppm, and  $n_{(\text{MCD})} = 180$ . Attempts to generate a structure failed since the calculations were too time-consuming. Additional structural information was obtained from the generalized portrait of the molecule as discussed earlier. A KB fragment library search produced  $L = 1437$ , and a general portrait of the molecule was obtained. It was found that 703 fragments (55%) contained a benzene ring, and almost half of these (338) showed a 1,2-Ar substitution. A hypothesis was tested that the molecule contained at least one 1,2-Ar fragment. Twelve 1,2-Ar-containing fragments were automatically sorted out of the 60 fragments extracted from the user database in the first step (see Figure 9).

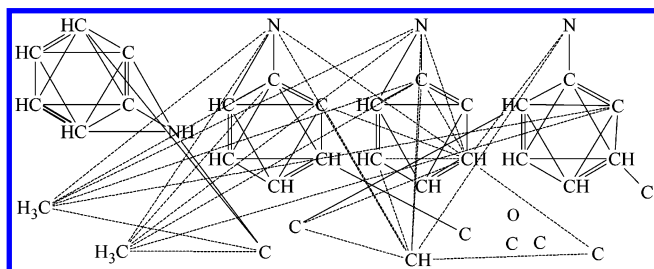
MCDs were generated from these fragments using the following parameter values:  $l = 12$ ,  $E = 6$  ppm,  $n_f = 4$ ,  $q = 50\%$ , and  $n_{(\text{MCD})} = 216$ . One of these MCDs is shown below in Figure 10 with only HMBC connectivities visible.

In these MCDs the fragments account for more than 70% of the skeletal atoms, suggesting they are highly likely to be good fragments. It was proven by IR spectroscopy<sup>30</sup> that the analyzed molecule contained an amide group. As a result a ketone functional group was added to the BADLIST to produce  $k = 192 \rightarrow 1$ ,  $t = 18$  min 30 s. The only structure consistent with the data was the structure of cryptospirolepine.





**Figure 9.** Twelve 1,2-Ar-containing fragments sorted out of the 60 fragments extracted from the User Fragment Library.



**Figure 10.** One of MCDs created during cryptospirolepine elucidation.

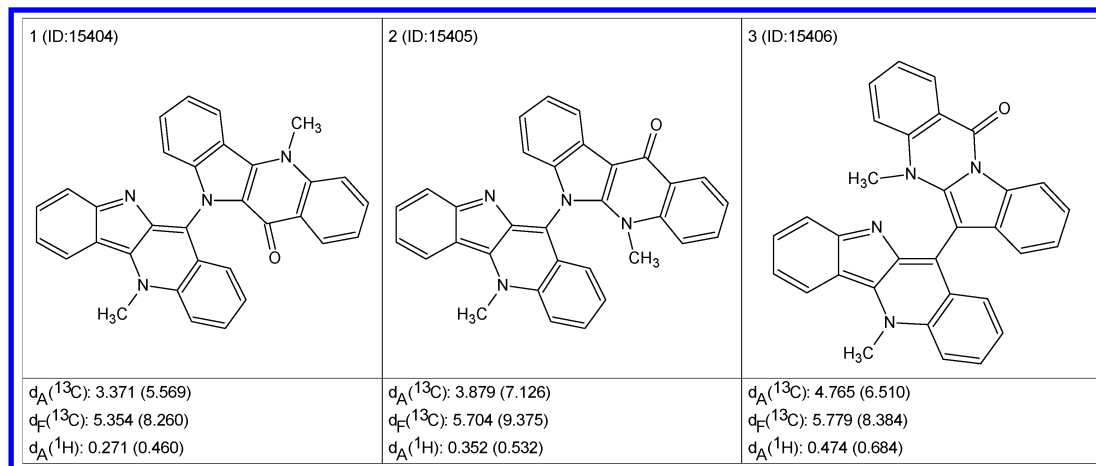
**6.1.3. Cryptoquindolinone.** For the elucidation of this structure, the raw spectral data were processed and input into the program using the processing tools available in the ACD/Labs SpecManager software.<sup>31</sup> A 1D  $^{13}\text{C}$  NMR spectrum was not available as is very common in natural product analysis where quantities of material are limited. The  $^{13}\text{C}$  shift inputs were thus created from the HSQC and HMBC spectra. 18 peaks were identified in the HSQC spectrum (2  $\text{CH}_3$  and 16 CH), and 13 peaks were extracted from the HMBC data to give 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, an estimated value to resonate in the middle of the aromatic interval. The number of peaks in the HMBC spectra acquired in standard and phase-sensitive mode was different, 32 and 45 peaks, respectively. To avoid contradic-

tions, the extra peaks observed in the second HMBC experiment were attributed to a range of potential couplings and allowed to be  $^{2-4}J_{\text{CH}}$ .

Searching for the  $^{13}\text{C}$  NMR spectrum in the user fragment library resulted in 101 fragments. 3144 MCDs were created with  $E = 2$  ppm and  $q_{\text{fr}} = 4$  after about 25 min. Checking for contradictions reduced the number of MCDs to 1376. Structure generation was performed with all spectral filters switched off and a cycle size constraint of  $R_c = 5-7$ . A carbonyl group was added to the GOODLIST. The result of structure generation gave  $k = 7850 \rightarrow 75$  and  $t = 30$  min. As the structures were ranked by  $d_A$  values, the target structure was moved to the first position. The first three structures of the ranked file are shown in Figure 11.

The difference in the deviation  $d_A(2) - d_A(1) = 0.5$  ppm is small. However, the increase in deviation of both  $d_H$  and  $d_F$  from structure to structure suggested that the structure of compound **11** was correctly identified.

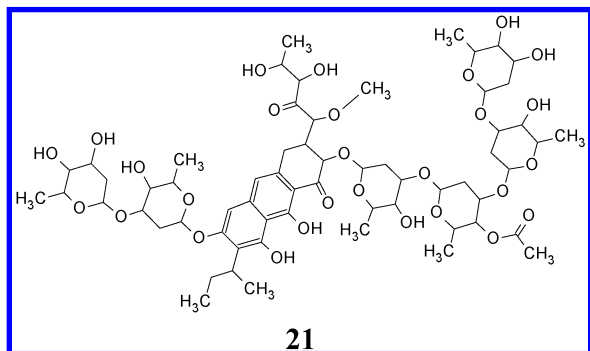
It is common for an experienced spectroscopist to detect molecular fragments simply by visual analysis of 1D and 2D NMR data. This ability is based on experience, knowledge, and insight of a highly qualified researcher, and the structural information "extracted" from a problem by this route is invaluable.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



**Figure 11.** First three structures of the ranked file obtained as the result of identification of compound **11**.



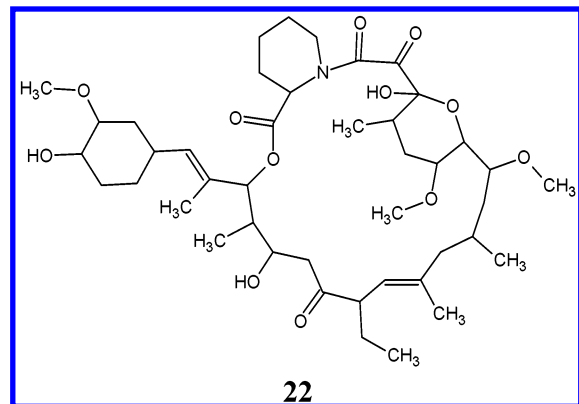
shows that the program can recognize structures of fairly sizable molecules. The largest of these molecules was  $n = 90$ ,  $M = 1284$ ,  $C_{62}H_{92}O_{28}$ . Spectral data were obtained from ref 34. The structure was solved using the *common mode* with COSY and HMBC input data only and a constraint of  $R_c = 6$ . The results gave  $k = 5140 \rightarrow 66$ ,  $t = 4$  h 32 min, and  $r_A = r_F = 1$  and the structure obtained is shown below:



In 60 cases, almost 50% of the problems studied, the size of the molecules exceeded 30 skeletal atoms. With these statistics we challenge the opinion of Steinbeck<sup>35</sup> who insists that a deterministic system is not capable of elucidating the structure of molecules with more than 30 skeletal atoms.

When 2D NMR data is used in combination with found fragments and/or user fragments, two crucial stages can be distinguished: the formation of molecular connectivity diagrams and structure generation. The time consumed for the process of MCD creation is usually measured by seconds and minutes. If MCDs containing appropriate fragments are not created, then the problem cannot be solved within reasonable amounts of time. Structure generation needs to generate an output file that is not empty, and the processing time for structure generation should not be too long. If the generation time is acceptable, then checking the different structural hypotheses is rarely an issue.

An examination of the distribution of problems in regards to the time required for structure generation showed that in the overwhelming majority of cases (75%), structure generation was completed in less than one minute. As a comparison of structure generation speeds, the *StrucEluc* and COCON systems have been compared using materials published by Junker et al.<sup>36</sup> They described the structure elucidation of ascomicyne, **22**, ( $C_{43}H_{69}NO_{12}$ ) using COSY and HMBC spectral data:



In the *common mode*, with default atom properties set, spectral filtering and application of APCT during structure generation, *StrucEluc* generates only one correct structure within 2 s. This result was obtained using a Celeron processor at a clock speed of 300 MHz. When atom hybridizations were set, but checks by heteroatom neighboring and spectral filter were switched off, the results gave  $k = 1926$ ,  $t_g = 56$  s,  $r_A = r_F = 1$ ,  $d_A(1) = 0.95$  ppm, and  $\Delta_{(2-1)} = 0.64$  ppm. Spectral filtering of the output file reduced the output file:  $k = 1926 \rightarrow 321$ . Under similar conditions, the COCON program produced the following results:  $k = 350$ ,  $t_g = 48$  min,  $r = 2$ ,  $d(1) = 3.51$  ppm, and  $d(2) = 3.56$  ppm. These results were obtained using an SGI R10000 running at a clock speed of 195 MHz. Under essentially the same conditions, the *StrucEluc* structure generator runs markedly faster, and, in this example, the correct structure selection proved to be more successful ( $r = 1$ ).

This work emphasizes that only expert systems based on the application of 2D NMR data allows the elucidation of complex molecules and, as evidenced in this work, in particular natural products. In this regard we argue with Meiler et al.<sup>37</sup> who suggest using the 1D  $^{13}C$  NMR spectrum as an alternative to 2D NMR. That group suggests<sup>37</sup> that 1D  $^{13}C$  NMR could be successfully used when a genetic algorithm for structure generation is employed. The authors argue<sup>37</sup> that one of the benefits of 1D  $^{13}C$  NMR is the fact that “the time-consuming determination of connectivity information... by 2D NMR spectroscopy is replaced by the much easier and rapidly obtainable chemical shift value”. While this may be true for small organic molecules where there is an abundance of sample and where the  $^{13}C$  NMR spectra are not crowded, this is certainly not true in many structural elucidation problems. In particular, when only small quantities of natural products, metabolites, forensic samples, etc. are isolated from their respective source, the possibility of efficiently obtaining a 1D  $^{13}C$  NMR spectrum is minimal even with recent advances in hardware technology. The problems presented in the article<sup>37</sup> are limited to molecules where the number of skeletal atoms is less than 20. Generally speaking, the amount of structural information described within a heteronuclear shift correlation 2D NMR spectrum is much higher than that extracted from a 1D  $^{13}C$  NMR. It appears that the genetic algorithm is unable to compensate for a lack of structural information especially in regards to larger molecules.

For choosing the most probable structure in *StrucEluc*, the procedure consists of four steps which are fulfilled along with removing duplicate structures: (1) fast prediction of  $^{13}C$  NMR spectra for all structures using the fast increment method; (2) preliminary ranking of the structures in order of increasing  $d_F$  value; (3) accurate prediction of  $^{13}C$  NMR spectra for the first 10–20 structures within a ranked file; and (4) rank ordering based on increasing  $d_A$  values. Statistical analysis of these results shows that for a reliability estimation of the most probable structure in the output file, the value  $\Delta_{(2-1)} = d_A(2) - d_A(1)$  can be used. If  $\Delta_{(2-1)} \geq 1$  ppm, the first structure of the output file ranked in order of increasing  $d_A$  value is, as a rule, correct. Using this approach, it is possible to distinguish the correct structure even in those cases when the structural file contains thousands of structures. As the procedure of preferable structure selection is by necessity routine, it is completely automated in the *StrucEluc* system.



During this work, the calculation of MS match factors,  $m_i$ ,<sup>38</sup> helped to confirm the validity of the most probable structure. The structure of *tasiamide* (C<sub>42</sub>H<sub>67</sub>N<sub>7</sub>O<sub>10</sub>) was elucidated using 2D NMR data from the article by Williams et al.<sup>39</sup> The top ranked structures had very close  $d_A$  deviations:  $d_A(1) = 1.48$ ,  $d_A(2) = 1.64$ ,  $d_A(3) = 1.91$  ppm, and  $\Delta_{(2-1)} = 0.16$  ppm. The priority of the correct structure was confirmed by the following MS match factors calculated for the ranked structures:  $m_1 = 97\%$ ,  $m_2 = 78\%$ , and  $m_3 = 48\%$ .

The possibility of utilizing substructural fragments during the elucidation process using 2D NMR data has been described. The methodology and corresponding algorithms have been developed to employ both KB fragments as well as those introduced by the user. The value of this approach has been corroborated by solving a number of problems.

As a result of testing this approach, it has been concluded that the utilization of found and user fragments is profitable in the following situations:

When the number of experimentally available 2D NMR correlations is markedly less than the number of theoretical correlations for a given structure. This situation can occur due to unfortunate experimental parameter choices and/or sample induced limitations such as a low  $S/N$  ratio.

When the number of 2D correlations is small due to a deficit in the number of hydrogen atoms and/or specific stereochemical factors.

When a molecule under investigation is so large and the spectral data so complex that even rich 2D NMR data can lead to a long structure generation time.

The application of found fragments frequently allows the solution of a problem either with minimum user intervention or in a fully automated fashion. The precision of the accurate method of <sup>13</sup>C NMR spectrum prediction is more than enough to provide high quality user fragments to the program. Moreover, the user-defined fragments can be utilized in the creation of molecular connectivity diagrams from found fragments containing the given user fragments.

This investigation has allowed the identification of a number of criteria for estimating the appropriateness of a user fragment. A user fragment, UF, should be selected such that the UF would represent a terminal or external part of a molecule. If a fragment belongs to the internal core of a molecule, the accurate <sup>13</sup>C NMR spectrum calculation turns out to be too approximate. In this case, when the discrepancy between calculated and experimental chemical shifts is considerable, the program is usually not capable of accepting the fragment for the creation of molecular connectivity diagrams. Moreover, the introduction of such UF's that have two or more free bonds from a terminal atom should be avoided. In this case, the difference between the experimental and calculated chemical shifts generally exceeds the default value of 12 ppm adopted for the terminal atoms in *StrucEluc-2*.

The application of found fragments requires the following three operations: (1) searching for the fragments in the database; (2) creating a set of molecular connectivity diagrams from the found fragments; and (3) checking the diagrams for contradictions. All of these procedures can take some time, and in some cases the total time required for problem solving may be of the same order or even longer than the *common* mode approach. Occasionally the most

time-consuming procedure can be the process of choosing the valid  $E$  value by gradually increasing its magnitude. It should be emphasized that when fragments are used, the structure generation is performed most frequently not from a single MCD, as in the case of *common* mode, but from a set of MCDs. Therefore, it is necessary to resort to the fragment approach only when the *common* mode fails or a significant number of assumptions are required to solve the problem in a reasonable time.

The procedure of creating MCDs from found fragments can be made more effective by employing a user defined BADLIST for filtering the list of FFs. For example, the absence of X-H groups, where X is a heteroatom, for a molecule under analysis can generally be determined from either NMR or IR spectra. These groups can then be included into the BADLIST. All found fragments that have the mentioned groups may then be rejected by filtering the FFs file with the BADLIST. As a result, incorrect fragments are excluded from the process of forming the MCDs. Depending on the nature of the molecular formula, the number of found fragments,  $L$ , can be very large. For a compound with molecular formula C<sub>52</sub>H<sub>80</sub>N<sub>8</sub>O<sub>6</sub>S (section 5, *Apramide G*) the number of found fragments turned out to be ~14 000. Application of found fragment filtering with a reasonably formed BADLIST usually leads to a considerable decrease in the number of FFs, occasionally by 30–40%.

This study has shown that there is a need for utilizing the fragment approach in expert systems based on 2D NMR data. In particular, the *StrucEluc-2* program has already been successfully applied to the automated structure elucidation of a series of cryptolepine derivatives, the identification of degradants, e.g., **11**, of the complex alkaloid cryptospirolepine (**10**) using NMR cryoprobe technology,<sup>21,22,30</sup> determination of products obtained in a reaction of an  $\alpha,\beta$ -unsaturated pyruvate,<sup>31</sup> and in a series of investigations not yet published.

Despite the promising results reported here, continued development and optimization of the system is necessary. The program needs to be enhanced to allow new 2D NMR techniques that are presently in development to be used. There is also a necessity to enhance the system with new algorithms to allow the identification of stereochemical information from recently developed 2D NMR experiments. The ability of the algorithms utilized to detect the presence of contradictions in 2D NMR data also needs to be further refined to make the process still more robust. We have already demonstrated that in 90% of the problems that were extremely challenging due to the presence of contradictions, the latter were automatically revealed by the program.<sup>9</sup>

**List of Abbreviations:** KB, knowledge base, containing full structure library and fragment library; FL, fragment library; LSC, library of spectrum-to-structure correlations; TFGL, typical functional group library; APCT, atom property correlation table; MCD, molecular connectivity diagram; FF, fragment found in FL from the 1D <sup>13</sup>C NMR spectrum; UF, user fragments involved by the investigator in a structure elucidation process; MF, molecular formula;  $E$ , the maximum difference allowed between the chemical shifts of the fragment carbons and the corresponding values observed in the experimental spectrum under study;  $L$ , number of FFs chosen as a result of fragment search in FL from the <sup>13</sup>C NMR spectrum;  $l$ , number of found fragments which will



be used for the creation of MCDs ( $l \leq L$ );  $n_f$ , the minimal number of fragments that must be present in each MCD;  $q$ , the minimum percentage of all skeletal atoms that must be absorbed by the fragments present in each MCD;  $n_{\text{MCD}}$ , number of MCDs created using fragments;  $k$ , number of structures generated;  $t_g$ , structure generation time;  $d_A$ ,  $d_F$ , average deviations of predicted  $^{13}\text{C}$  NMR chemical shifts from the experimental ones calculated by “accurate” (A) and “fast” (F) methods correspondingly;  $d_H$ , average deviation of predicted  $^1\text{H}$  NMR chemical shifts from the experimental ones;  $r_A$ ,  $r_F$ ,  $r_H$ , positions of the correct structure in an output structural file ranked by corresponding average deviations.

### ACKNOWLEDGMENT

The authors are grateful to Sherry Peach for a thorough review and proofreading of this manuscript.

### REFERENCES AND NOTES

- Elyashberg, M. E.; Blinov, K. A.; Martirosian, E. R. A new approach to computer-aided molecular structure elucidation: the expert system *Structure Elucidator*. *Autom. Inf. Manage.* **1999**, *34*, 15–30.
- Blinov, K. A.; Elyashberg, M. E.; Molodtsov, S. G.; Williams, A. J.; Martirosian, E. R. An expert system for automated structure elucidation utilizing  $^1\text{H}$ - $^1\text{H}$ ,  $^{13}\text{C}$ - $^1\text{H}$  and  $^{15}\text{N}$ - $^1\text{H}$  2D NMR correlations. *Fresenius J. Anal. Chem.* **2001**, *369*, 709–714.
- Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Molodtsov, S. G.; Martirosian, E. R. Application of a new expert system for the structure elucidation of natural products from their 1D and 2D NMR data. *J. Nat. Prod.* **2002**, *65*, 693–703.
- Elyashberg, M. E.; Martirosian, E. R.; Karasev, Yu. Z.; Thiele, H.; Somberg, H. X-PERT: a user-friendly expert system for the molecular structure elucidation by spectral methods. *Anal. Chim. Acta* **1997**, *337*, 265–286. Expert systems as a tool for the molecular structure elucidation by spectral methods. Strategies of solution to the problems. *Anal. Chim. Acta* **1997**, *348*, 443–463.
- Funatsu, K.; Miyabayashi, N.; Sasaki, S. Further development of structure generation in the automated structure elucidation system, CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18–28.
- Christie, B. D.; Munk, M. E. Structure elucidation by reduction — a new strategy for computer-assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87–93.
- Will, M.; Fachinger, W.; Richtert, J. R. Fully automated structure elucidation—a spectroscopist’s dream comes true. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221–227.
- Elyashberg, M. E.; Karasev, Yu. Z.; Martirosian, E. R. Spectroscopic determination of elemental composition of organic compounds with the aid of the X-PERT system. *Anal. Chim. Acta* **1999**, *388*, 353–363.
- Molodtsov, S. G.; Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Martin, G. M. Detection and removal of contradictions in the 2D NMR data during structure elucidation by the *StrucEluc* expert system. *J. Chem. Inf. Comput. Sci.* Submitted for publication.
- Molodtsov, S. G. Computer-aided generation of molecular graphs. *Commun. Math. Chem. (MATCH)* **1994**, *302*, 213–224.
- Molodtsov, S. G. Generation of molecular graphs with a given set of nonoverlapping fragments. *Commun. Math. Chem. (MATCH)* **1994**, *30*, 203–212.
- Molodtsov, S. G. The generation of molecular graphs with obligatory, forbidden and desirable fragments. *Commun. Math. Chem. (MATCH)* **1998**, *37*, 157–162.
- Bremser, W. HOSE — a novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355–365.
- Willett, P.; Barnard, J.; Downs, G. J. Chemical similarity searching. *Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- Gonzales, A. G.; Leon, F.; Sanchez-Pinto, L.; Pardon, J. I.; Bermejo, J. Phenolic compounds of dragon’s blood from *Dracaena draco*. *J. Nat. Prod.* **2000**, *63*, 1297–1297.
- Gribov, L. A.; Elyashberg, M. E.; Serov, V. V. On the solution of the one classical problem in vibrational spectroscopy. *J. Mol. Struct.* **1978**, *50*, 371–387.
- Luesch, H.; Yoshida, W. Y.; Moore, R. E.; Paul, V. J.; Mooberry, S. L. Isolation, structure determination, and biological activity of lingbyabellin A from the marine cyanobacterium *Lyngbya majuscula*. *J. Nat. Prod.* **2000**, *63*, 611–615.
- Ohta, T.; Maruyama, T.; Nagahashi, M.; Miyamoto, Y.; Hosoi, S.; Kiuchi, F.; Yamazoe, Y.; Tsukamoto, S. Paradisin C: a new CYP3A4 inhibitor from grapefruit juice. *Tetrahedron* **2002**, *58*, 6631–6635.
- Luesch, H.; Yoshida, W. Y.; Moore, R. E.; Paul, V. J. Apramides A–G, novel lipopeptides from the marine cyanobacterium *Lyngbya majuscula*. *J. Nat. Prod.* **2000**, *63*, 1106–1112.
- Marquez, B.; Watts, K. S.; Yokochi, A.; Roberts, M. A.; Verdier-Pinard, P.; Jimenez, J. I.; Hamel, E.; Scheuer, P. J.; Gerwick, W. H. Structure and absolute stereochemistry of hectochlorin, a potent stimulator of actin assembly. *J. Nat. Prod.* **2002**, *65*, 866–871.
- Martin, G. E.; Hadden, C. E.; Kaluzny, B. D.; Russell, D. J.; Stiemsma, B. A.; Thamann, T. J.; Crouch, R. C.; Blinov, K. A.; Elyashberg, M. E.; Williams, A. J.; Martirosian, E. R.; Schiff, P. L., Jr. Identification of degradants of a complex alkaloid using NMR cryptoprobe technology and ACD/Structure Elucidator. *J. Heterocycl. Chem.* **2002**, *39*, 1241–1250.
- Blinov, K. A.; Carlson, D.; Elyashberg, M. E.; Martin, G. E.; Martirosian, E. R.; Molodtsov, S. G.; Williams, A. J. Computer assisted structure elucidation of natural products with limited 2D NMR data: application of the *StrucEluc* system. *Magn. Reson. Chem.* **2003**, *41*, 359–372.
- Ablordepey, S. Y.; Hufford, C. D.; Borne, R. F.; Dwumu-Badu, D.  $^1\text{H}$  NMR and  $^{13}\text{C}$  NMR assignment of cryptolepine, a 3: 4-benz- $\delta$ -carboline derivative isolated from *Cryptolepis sanguinolenta*. *Planta* **1990**, *56*(4), 416–417.
- Hadden, C. E.; Duholke, W. K.; Guido, J. E.; Robins, R. H.; Martin, G. E.; Sharaf, M. H. M.; Schiff, P. L., Jr. Structural characterization of components of a mixture — characterization of a facile oxidation product of cryptolepine in situ. *J. Heterocycl. Chem.* **1999**, *36*, 525–531.
- Spitzer, T. D.; Crouch, R. C.; Martin, G. E.; Sharaf, M. H. M.; Schiff, P. L., Jr.; Tackie, A. N.; Boye, G. L. Total assignment of the proton and carbon NMR spectra of the alkaloid quindoline — utilization of HMQC-TOCSY to indirectly establish protonated carbon-protonated carbon connectivities. *J. Heterocycl. Chem.* **1991**, *28*, 2065–2070.
- Sharaf, M. H. M.; Schiff, P. J., Jr.; Martin, G. E.; Phoebe, C. H., Jr.; Tackie, A. N. Two new indoloquinoline alkaloids from *Cryptolepis sanguinolenta*: cryptosanguinolentine and cryptotackieine. *J. Heterocycl. Chem.* **1996**, *33*, 239–243.
- Sharaf, M. H. M.; Schiff, P. J., Jr.; Tackie, A. N.; Phoebe, C. H., Jr.; Davies, A.; Andrews, C. W.; Crouch, R. C.; Martin, G. E. Isolation and elucidation of the structure of homocryptolepine. *J. Heterocycl. Chem.* **1995**, *32*, 1631–1636.
- Sharaf, M. H. M.; Schiff, P. J., Jr.; Tackie, A. N.; Phoebe, C. H., Jr.; Johnson, R. L.; Minick, D.; Andrews, C. W.; Crouch, R. C.; Martin, G. E. The isolation and structure determination of cryptomisine, a novel indolo[3,2-b]quinoline dimeric alkaloid from *Cryptolepis sanguinolenta*. *J. Heterocycl. Chem.* **1996**, *33*, 789–797.
- Crouch, R. C.; Davies, A.; Spitzer, T. D.; Martin, G. E.; Sharaf, M. H. M.; Schiff, P. J., Jr.; Phoebe, C. H., Jr.; Tackie, A. N. Elucidation of the structure of quindolinone, a minor alkaloid of *Cryptolepis sanguinolenta*: submilligram  $^1\text{H}$ - $^{13}\text{C}$  and  $^1\text{H}$ - $^{15}\text{N}$  heteronuclear shift correlation experiments using micro inverse-detection. *J. Heterocycl. Chem.* **1995**, *32*, 1077–1080.
- Tackie, A. N.; Boye, G. L.; Sharaf, M. H. M.; Schiff, P. L.; Crouch, R. C.; Spitzer, T. D.; Johnson, R. L.; Dunn, J.; Minick, D.; Martin, G. E. Cryptospirolepine, a unique spiro-nonacyclic alkaloid isolated from *Cryptolepis sanguinolenta*. *J. Nat. Prod.* **1993**, *54*, 653–670.
- ACD/SpecManager program. This program is capable of reading data presented in practically all formats common in NMR spectroscopy. [http://www.acdlabs.com/products/spec\\_lab/exp\\_spectra/](http://www.acdlabs.com/products/spec_lab/exp_spectra/).
- Blinov, K. A.; Elyashberg, M. E.; Martirosian, E. R.; Molodtsov, S. G.; Williams, A. J.; Sharaf, M. H. M.; Schiff, P. L., Jr.; Crouch, R. C.; Martin, G. E.; Hadden, C. E.; Guido, J. E. Quindolinocryptotackieine: the elucidation of a novel indoloquinoline alkaloid structure through the use of computer-assisted structure elucidation and 2D-NMR. *Magn. Reson. Chem.* **2003**, *41*, 577–584.
- Sharman, G. J.; Jones, I. C.; Parnell, M. J.; Willis, M.; Carlson, D. V.; Williams, A.; Elyashberg, M. E.; Blinov, K. A. and Molodtsov, S. G. Automated structure elucidation of two unexpected reaction products in a reaction of  $\alpha,\beta$ -unsaturated pyruvate. *Magn. Reson. Chem.* In press.
- Jayasuriya, H.; Lingham, R. B.; Graham, P.; Quamina, D.; Herranz, L.; Genilloud, O.; Gagliardi, M.; Danzeisen, R.; Tomassini, J. E.; Zink, D. L.; Guan, Z.; Singh, S. B. Durhamycin A, a potent inhibitor of HIV tat transactivation. *J. Nat. Prod.* **2002**, *65*, 1091–1095.

- (35) Steinbeck, C. J. SENECA: a platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. *Chem. Inf. Comput. Sci.* **2001**, *41*, 1500–1507.
- (36) Junker, J.; Maier, W.; Lindel, T.; Koeck, M. Computer-assisted constitutional assignment of large molecules: COCON analysis of ascomycin. *Org. Lett.* **1999**, *1*, 737–740.
- (37) Meiler, J.; Will, M. C. Automated structure elucidation of organic molecules from  $^{13}\text{C}$  NMR spectra using genetic algorithms and neural networks. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1535–1546.
- (38) Williams, A. J.; Lee, M. S.; Lashin, V. An integrated desktop mass spectrometry. Processing and molecular structure management system. *Spectroscopy* **2001**, *16*, 38–49.
- (39) Williams, P. G.; Yoshida, W. Y.; Moore, R. E.; Paul, V. J. Tasiamide, a cytotoxic peptide from the marine cyanobacterium *Symploca* sp. *J. Nat. Prod.* **2002**, *65*, 1336–1339.

CI0341060