

Are Deterministic Expert Systems for Computer-Assisted Structure Elucidation Obsolete?

Mikhail E. Elyashberg,[†] Kirill A. Blinov,[†] Antony J. Williams,^{*,‡} Sergey G. Molodtsov,[§] and Gary E. Martin^{||,⊥}

Advanced Chemistry Development, Moscow Department, 6 Akademik Bakulev Street, Moscow 117513, Russian Federation, Advanced Chemistry Development, Inc., 110 Yonge Street, 14th floor, Toronto, Ontario, Canada M5C 1T4, Novosibirsk Institute of Organic Chemistry, Siberian Branch of Russian Academy of Science, Lavrentiev Avenue 9, Novosibirsk 630090, Russia, and Michigan Structure Elucidation Group, Pfizer Global Research and Development, Kalamazoo, Michigan 49001-0199

Received October 24, 2005

Expert systems for spectroscopic molecular structure elucidation have been developed since the mid-1960s. Algorithms associated with the structure generation process within these systems are deterministic; that is, they are based on graph theory and combinatorial analysis. A series of expert systems utilizing 2D NMR spectra have been described in the literature and are capable of determining the molecular structures of large organic molecules including complex natural products. Recently, an opinion was expressed in the literature that these systems would fail when elucidating structures containing more than 30 heavy atoms. A suggestion was put forward that stochastic algorithms for structure generation would be necessary to overcome this shortcoming. In this article, we describe a comprehensive investigation of the capabilities of the deterministic expert system Structure Elucidator. The results of performing the structure elucidation of 250 complex natural products with this program were studied and generalized. The conclusion is that 2D NMR deterministic expert systems are certainly capable of elucidating large structures (up to about 100 heavy atoms) and can deal with the complexities associated with both poor and contradictory spectral data.

1. INTRODUCTION

Expert systems allowing computer-assisted structure elucidation based on spectral data, including multidimensional NMR data, have proven valuable in the elucidation of complex organic molecules. Articles published in recent years have reported on systems capable of elucidating the chemical structures of even complex natural products.^{1–9} Some studies^{7,10–21} report the application of expert systems to the elucidation of novel structures of natural compounds. Until recently, most publications presented examples that showed the capabilities of the expert systems in terms of solving tasks that would commonly be dealt with by qualified spectroscopists. In one of our recent studies,¹¹ an expert system elucidated the structure of a complex alkaloid C₃₁H₂₀N₄ that could not be identified by a highly experienced spectroscopist, even after a decade of effort. The success of this work¹¹ highlights the capabilities of the Structure Elucidator system, referred to as StrucEluc for the remainder of this article,^{7–9} developed as a result of tenacious efforts to develop algorithms and databases that can be applied to structure elucidation tasks in a general manner.

The successes of StrucEluc are dependent on the impressive achievements and developments of 2D NMR spectroscopy

that are now both routine and widespread. The variety and quality of the data inputs available from 2D NMR data have allowed the development of expert systems that are able to elucidate structures of very complex organic molecules. Until recently, systems were deterministic, using discrete mathematical methods such as logic, combinatorial analysis, and graph theory combined with heuristic approaches for the exhaustive generation of all isomeric structures satisfying a set of structural constraints.

Experience accumulated as a result of the application of expert systems to many examples has demonstrated that such systems derive direct value from the huge amount of chemical structure information contained within spectral data. Figure 1 displays the structures of a series of small natural product molecules elucidated from their 2D NMR spectra, as reported in the literature, and analyzed by our research group using the StrucEluc expert system.^{7,8}

In Figure 1, N is the number of structural isomers that were generated using the structure generator within the StrucEluc system. Even the simplest structures can have hundreds of millions if not billions of isomers, and a successful result depends on the screening and rejection of $N - 1$ structural formulas that do not comply with the experimental data and systematic constraints applied. The number of isomers associated with the structures of medium-sized complex organic molecules can be estimated at about 10^{20} – 10^{30} isomers. For comparison purposes, the number of stars in our galaxy is “only” 10^{11} , so it is appropriate to comment that astronomical numbers and “isomeric” numbers are comparable.

* Corresponding author phone: 919-341-8375; fax: 425-790-3749; e-mail: tony@acd labs.com.

[†] Advanced Chemistry Development, Moscow Department.

[‡] Advanced Chemistry Development, Inc.

[§] Siberian Branch of Russian Academy of Science.

^{||} Pfizer Global Research and Development.

[⊥] Present address: Schering-Plough Corp., 2000 Galloping Hill Rd, K-11-3 L5, Kenilworth, NJ 07033.

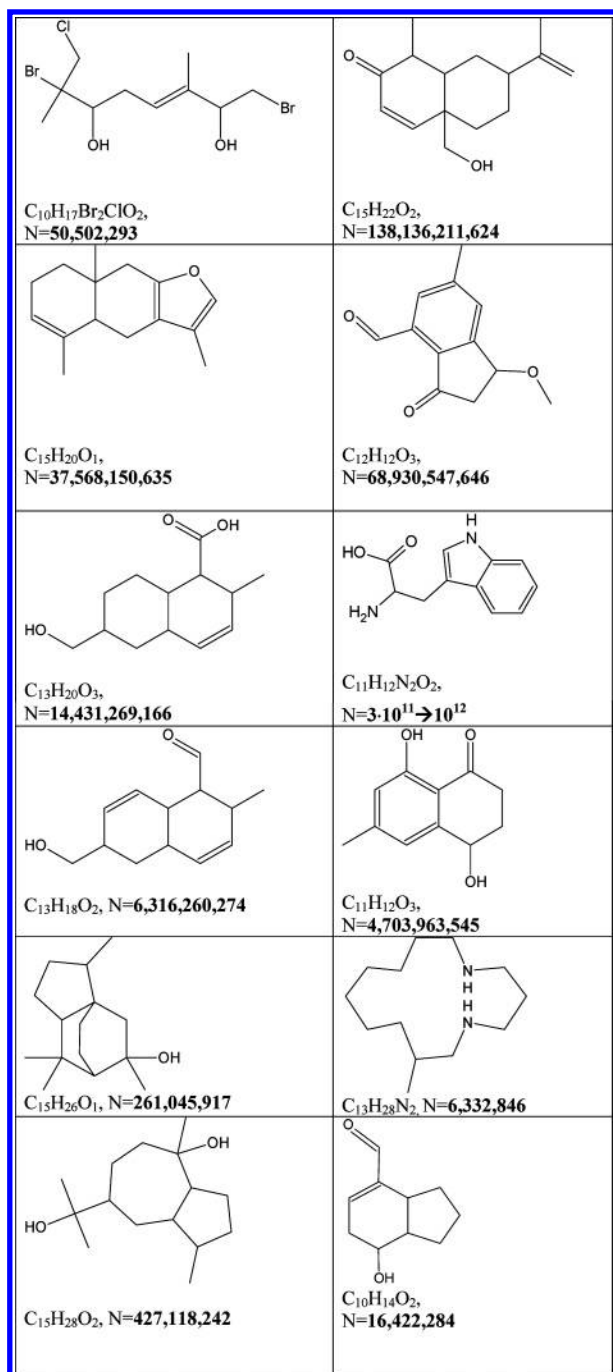


Figure 1. Structures of a number of natural products including their associated molecular formulas and the numbers of mathematically possible isomers.

A huge number of isomers corresponding to molecules containing 30–100 skeletal atoms resulted in some researchers^{22–24} declaring that, in principle, deterministic systems would not be able to analyze the entire space of potential isomers and, therefore, would fail in the elucidation of structures. As an alternative, they suggested stochastic algorithms for structure generation, specifically simulated annealing,^{22,23} and genetic algorithms.^{24,25} In a series of reports,^{23,24,26} Steinbeck et al. limited the number of skeletal atoms in a molecule that could be identified by deterministic systems to 30 or less. Examples of large molecules that were successfully elucidated and reported, but above the limit of 30 atoms, were considered to be exceptions.²³ One would

expect that, in order to support their position, the proponents of stochastic algorithms would cite successful examples of stochastic algorithms applied to molecules with more than 30 skeletal atoms. This has not happened to the best of our knowledge. Steinbeck et al.^{23,24} have shown that stochastic methods can elucidate structures containing close to 30 skeletal atoms.

Stochastic methods of structure generation are of interest because any such algorithms should be comprehensively studied in terms of their potential advantages in expert systems. It is not improbable that the use of techniques peculiar to deterministic systems in combination with stochastic algorithms could lead to the creation of *hybrid* systems that, because of synergistic effects, could successfully compete with purely deterministic systems. While this interest exists, we question the validity of the statement that deterministic systems are already obsolete and are limited to the elucidation of structures of molecules containing less than 30 skeletal atoms. During the past 35 years of development of deterministic systems, a large number of methods have been elaborated to overcome the “problem of dimensionality”.^{1–9} It will be shown below that molecules with 40–80 skeletal atoms are fairly typical examples in computer-assisted structure elucidation using deterministic systems. With the effective application of the structural constraints provided by 2D NMR spectra, as well as the use of sophisticated databases and spectra-structural correlations accumulated by several generations of spectroscopists, it is possible to significantly reduce the resulting output set of a deterministic system. The sum total of all of these innovations introduces elements of artificial intelligence into expert systems, thereby supplying many abilities that are absent from a human expert. The program starts with a huge but restricted space for all possible isomers containing, of course, structures complying with the constraints imposed by spectral data or introduced as additional information. The challenge is to reveal these selected structures and select the most probable.

In this work, we investigate the properties, strategies, and techniques for elucidating the structure of an unknown compound using the deterministic system StrucEluc^{7–9,11,27–30} and on the basis of an input data set including 2D NMR spectra.

The principal scheme illustrating the operation of StrucEluc is shown in Figure 2.

Conceptually, the system operates in two modes—the so-called common and fragment modes. The common mode implies the utilization of 2D NMR data without the application of substructural fragments either suggested by the chemist or identified within the system knowledge base. The fragment mode is initiated when user-defined fragments are used or fragments are found as a result of a ¹³C NMR search in the fragment library (FL) containing ca. 1.5 million fragments and their associated ¹³C NMR subspectra. The final step in the elucidation is the selection of the most probable structure and determination of its likely relative stereochemistry and 3D geometry.

In this study, the solutions of over 250 complex structure elucidation problems, specifically, natural products, were comprehensively analyzed. The reliability of the system's knowledge base was examined together with the characteristics of the structures which were elucidated. Also, the

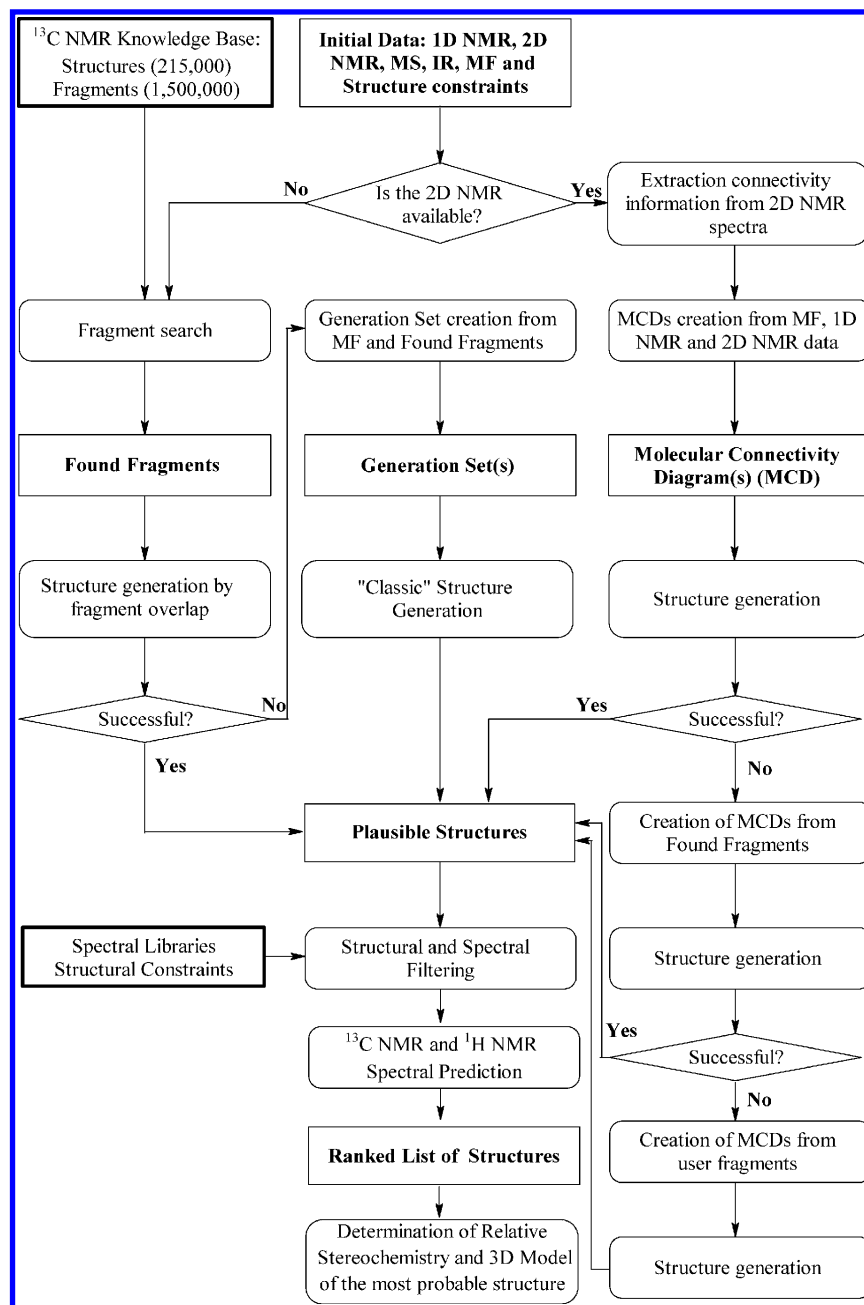


Figure 2. Flow diagram describing StrucEluc.

properties of the 2D NMR data used as the basis of analysis were studied to determine what specific features would be problematic and would limit the successful application of a computer-assisted structure elucidation system. Attention was given to the possibility of identifying the most probable structure in a large results output file and specifically to solving problems in the presence of either correlation spectroscopy (COSY) or heteronuclear multiple-bond correlation (HMBC) correlations of “nonstandard”²⁷ length (nonstandard correlations, NSC), that is, where correlations are characterized by $^nJ_{HH}$ and $^nJ_{CH}$ with $n > 3$.

Extensive research has resulted in the development of NMR spectral prediction methods that enable the selection of the most probable structure from an output set (for instance, refs 31–36). To the best of our knowledge, no tests have been performed on a wide range of real-world tasks to validate the effectiveness of these methods. In this article, a study of the discriminating ability of the prediction methods

described previously in our other works^{7–9} has been conducted. The method includes several stages and is based on the prediction of both 1H and ^{13}C NMR spectra using ACD/NMR prediction software.³²

As a result of our studies, we conclude that deterministic expert systems are certainly not yet obsolete and, moreover, it is likely that the near future will see an increasing application of such software programs to the elucidation of even more highly complex chemical structures.

2. RESULTS AND DISCUSSIONS

2.1. Analyzing the Relevance of the System Knowledge Base. It may seem that 2D NMR data allow the elucidation of a molecular structure as if it is an ab initio approach (*common mode*). Structure elucidation is, however, an inverse problem^{37,38} by nature. To obtain a unique structure as an output file, spectral data provided in various forms are

commonly necessary (including MS, NMR, IR, Raman, UV spectra, databases containing spectrum-structural information, spectral calculation outputs, etc.), and this problem can therefore be related to a class of so-called *ill-posed problems*.^{37,38} A consequence is that arbitrarily small changes in the input experimental data or in the knowledge base of a system may lead to arbitrarily large changes in the solution. With this in mind, it can be concluded that automated or computer-assisted structure elucidation is possible only in those cases when the data are correct, consistent, and comprehensive. The accuracy of the input spectral data should be confirmed by the spectroscopist. If these data are correct and consistent, the result of computer-aided structure elucidation will correspondingly depend on the quality and reliability of the system knowledge base. The following types of data are used in the knowledge base of StrucEluc:

1. A database of structures and their corresponding assigned ^1H and ^{13}C NMR spectra (>215 000 pairs) is used as the basis of the search for structures according to their ^{13}C NMR spectra and for the creation of the fragments database described below.

2. A fragments database generated from the structures database containing the assigned ^{13}C NMR subspectra (>1 500 000 fragments) is used to generate structures based on the 1D ^{13}C NMR spectra and to aid in the solution of tasks using 2D NMR in those cases when the 2D NMR data are not definitive enough to resolve the problem (the molecules may be too large, there may be extensive resonance overlap in one or both dimensions, or there may be a dearth of hydrogen atoms resulting in a scarcity of cross-peaks, etc.^{7,8}).

3. A database of spectra-structure correlations for ^1H NMR, ^{13}C NMR, and IR spectra is used for the spectral filtering of selected fragments and their generated structures.

4. A set of correlation tables containing atom-centered fragments and their spectral characteristics in the ^1H and ^{13}C NMR spectra (atom property correlation table, APCT) is used. These correlation tables are designed to allow automated determination of the properties of carbon atoms, specifically the hybridization state and the probability of certain neighboring heteroatoms. These correlations are used to create molecular connectivity diagrams (MCD)^{7,8} and to reduce the expanse of the search during the process of structure generation.

5. A set of special databases containing structures and their assigned ^1H and ^{13}C NMR spectra is used to assist in the prediction of the NMR spectra of structures contained in the resultant output file. A comparison between the experimental data and the predicted data is used to help identify the most probable structure.

The properties of the knowledge base of StrucEluc were examined and optimized using large files of reference data.

The structure database was produced from literature encompassing numerous journals publishing the assigned NMR spectra of synthetic substances and natural products. The structure database is therefore characterized by a high level of diversity, and the fragments excised from the structures will therefore also be diverse and available for the analysis of a wide range of classes of organic molecules.

To optimize the chemical shift intervals assigned to the fragments present in the correlation tables and tables used for setting the atom properties (APCT), all 215 000 structures

contained within the database of assigned ^1H and ^{13}C NMR spectra were used. The database structures were filtered with the help of the fragment libraries. If any contradictions were detected between a structure and the characteristic spectral intervals associated with a fragment, the program provided a corresponding message. On the basis of these messages, the intervals were modified to resolve the contradiction. Some fragments characterized by specific chemical shift values were placed into a *library of exceptions* that was applied as part of the structure filtration process using a specially derived algorithm.

Analysis of the fragment tables revealed that spectral filter libraries and correlation tables (APCT) allowed the solution of tasks with a negligible risk of overlooking the correct structure. A total of 96% of the structures present in the system database withstood a verification challenge by both ^1H and ^{13}C NMR spectra using both spectral filters and ACPT. When the high degree of diversity of the structure library is taken into account, it can be expected that the spectral filtering of the output file is a procedure that offers only a small risk of losing the actual structure. To validate the spectral filters as applied to natural products, we chose ca. 13 500 natural product compounds from the Full Structure database. It was found that 99.8% of them withstood spectral filtering.

As mentioned previously, a database containing ca. 1.5 million fragments was produced from the structures in the Full Structure database and forms the system fragment library. The properties of the fragment library were scrutinized as a result of challenging the StrucEluc system with a set of 250 problems. During the period over which the StrucEluc system has been developed, 155 of the 250 solved problems (set A) have been subsequently included into the Full Structure database, and as a result, these structures were involved in producing the current fragment library. Set B is identified as the set of remaining 95 problems whose structures are absent from the database. The full set of solved problems is referred to as P where $P = A \cap B$.

For all 250 problems, fragments with ^{13}C NMR spectra were searched in the fragment database. The sizes of the identified fragment files were generally $1000 < x < 2000$ structures, though in rare cases, they can be significantly larger (up to 20 000). It is noteworthy that the procedure of searching the fragments using ^{13}C NMR spectra is rather fast and generally does not exceed 30 s on a standard desktop computer (Pentium IV, 2.8 MHz).

The amount of information that can be elucidated from the system fragment library about an unknown structure is of interest. All structures included in the set P of "unknown" structures were filtered using the FL, and the *real fragments* (RFs) identified in each molecule were saved. These fragments, if found, offer the possibility of forming "good" MCDs from which a correct solution can be generated.

The main advantage of using fragments found in the database is that their carbon atoms are associated with chemical shifts close or equal to the experimental values observed in the ^{13}C NMR spectrum of the unknown substance. StrucEluc includes a procedure^{7,8} that will automatically assign the experimental chemical shifts to the carbon atoms of the fragments used to form the MCDs. Each assignment is then checked by the program for agreement with the 2D NMR connectivities. Other investigators^{26,39} have

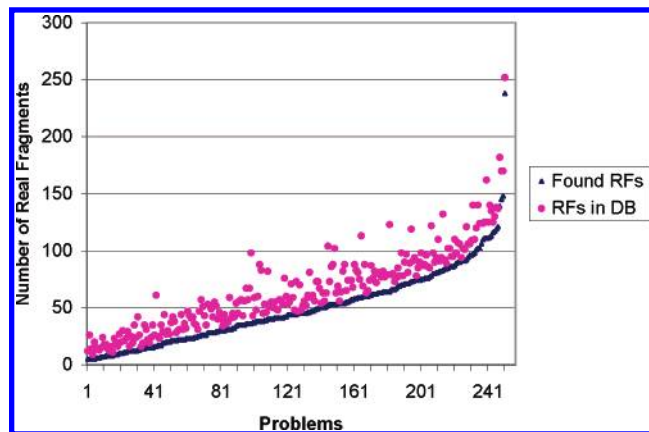
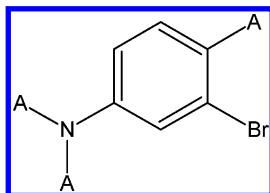


Figure 3. Numbers of real fragments (RFs) existing in the database (violet circles) compared with the numbers of *found* RFs (blue triangles). Not all of the RFs were detected during the fragment search.

reported the application of user-defined fragments for structure generation from 2D NMR data but gave no information about the method employed for chemical shift assignment of the fragment carbon atoms. We suspect that the authors simply set the chemical bonds between definite atoms, but we consider this an artificial approach. For instance, let the ^{13}C NMR spectrum of an unknown contain the following chemical shifts: 115 (d), 116 (d), 121 (d), 127 (d), 126 (s), 131 (s), 132 (s), 145 (s), and 150 (s) ppm. The user defined fragment is



A total of 2880 combinations of the chemical shift assignments should be produced and checked to select the single correct combination that will lead to the generation of the correct structure corresponding to all of the 2D NMR data. Obviously, it is impossible to guess the appropriate shift assignment without checking all of the combinations.

In reality, not all real fragments present in the database are selected during the fragment search using the ^{13}C NMR spectrum of an unknown structure. This is illustrated in Figure 3 where the numbers of real fragments existing in the database are measured by tens or even exceed hundreds, while the numbers of *found* real fragments (blue triangles) are always less than the corresponding possible values (violet circles).

An investigation was performed to reveal how the presence of an unknown molecule in the Full Structure database influences the result of a fragment search. The numbers of RFs $n_i(\text{A})$ and $n_j(\text{B})$ ($i = 1/155$, $j = 1/95$) were calculated for both subsets of P. The corresponding values $n_i^{\text{F}}(\text{A})$ and $n_j^{\text{F}}(\text{B})$ were determined to show how many RFs were found as a result of library searching using ^{13}C NMR spectra. In accord with Figure 3, it was shown that in all cases $n_i^{\text{F}}(\text{A}) < n_i(\text{A})$ and $n_j^{\text{F}}(\text{B}) < n_j(\text{B})$. The average values of $n_i(\text{A})$ and $n_j(\text{B})$ are equal to 76 and 49, correspondingly, and for $n_i^{\text{F}}(\text{A})$ and $n_j^{\text{F}}(\text{B})$, the values are 59 and 34, respectively. This means that the numbers of real fragments $n_j(\text{B})$ and $n_j^{\text{F}}(\text{B})$ are

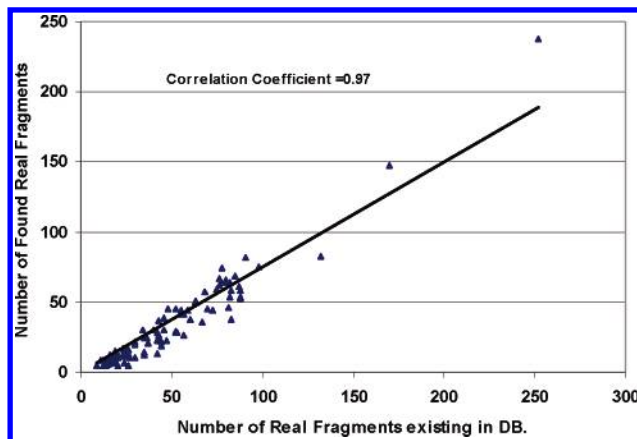


Figure 4. Dependence of $n_j^{\text{F}}(\text{B})$ on $n_j(\text{B})$. The correlation coefficient is 0.97.

therefore large enough even in those cases when the unknown structures are absent from the Full Structure library and, consequently, when they were not used in the production of the fragment library. The dependence of $n_j^{\text{F}}(\text{B})$ on $n_j(\text{B})$ is presented in Figure 4, which shows that the relationship between these values is linear in nature and has a correlation coefficient of 0.97.

The analysis described here leads us to conclude that fragments contained within the unknown molecule usually make up only a small amount of the fragments found as a result of a database search. To further improve the efficiency of employing fragments in the structure elucidation process, it is necessary to raise the selectivity of the fragment search and perfect methods that position the largest real fragments at the top of the list of found fragments.

2.2. General Characteristics of Identified Structures.

During the process of developing the StrucEluc system, more than 250 challenges were undertaken to elucidate the structures of recently isolated materials; almost all of the 250 compounds were natural products because nature delivers far more structural diversity than that generally observed via laboratory-based synthesis. Each of the problems was solved using spectral data obtained from two sources. The first source was literature articles spanning the years 2000–2004 extracted from the *Journal of Natural Products*, the *Journal of Organic Chemistry*, and *Magnetic Resonance in Chemistry*. The only criterion that influenced the choice of articles for the tasks was the presence of tables containing ^1H NMR, ^{13}C NMR, or HMQC or HSQC, and, optionally, COSY spectral data. The second source was raw NMR data acquired by users of the StrucEluc system whose interests included the identification of novel natural products. The diversity of structures identified in the 250 challenges is very likely representative of the diversity of natural products isolated and characterized during the identified period. For example, the selected compounds were identified to be members of the following classes: steroids, terpenoids, acyclic and cyclic peptides, glycosides, various alkaloids and heterocyclic compounds, and so forth.

The problems extracted from the literature were solved immediately after the corresponding publication appeared, thereby ensuring the exclusion of the new compounds from the system databases. If there was a need to solve the task later for revalidation as the algorithms were improved, the unknown structure was initially preliminarily searched in the

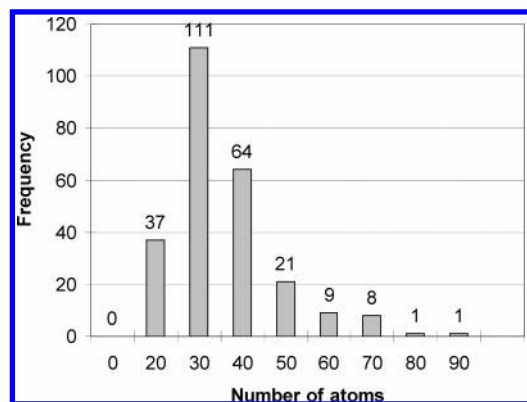


Figure 5. Distribution of structures relative to the number of skeletal atoms across the 250-example problem set. The bars correspond to the number of problems in a given range. The first bar comprised of 37 problems involved molecules with 1–20 atoms. The second bar with 111 problems involved molecules with 21–30 atoms, etc.

system database in an automated fashion. If it was found in the system databases, it was temporarily removed from the library for the process of revalidation. This allowed the deliberate exclusion of any influence of the presence of the molecule under study consideration during both the fragment search process and the prediction of both ^1H and ^{13}C NMR spectra.

The distribution of structures relative to the number of skeletal atoms is shown in Figure 5.

From the histogram in Figure 5, it is seen that the most typical structures are those where the number of skeletal atoms varies from 20 to 30 (typical molecular weights lie in the range of 300–600 Da). At the same time, more than 100 molecules contain from 30 to 90 skeletal atoms, which immediately invalidates the limit declared by others for deterministic expert systems.^{22,23} Almost all of the compounds contained oxygen or nitrogen atoms, up to a total of 30 such heteroatoms in a single molecule.

For the majority of structures, the double-bond equivalent (DBE) value varies between 5 and 15 and reached a maximum value of 25. The number of cycles varies from two to eight, with the majority of molecules containing three to four cycles. As expected, the molecules contained mostly five- and six-membered cycles. There were ~35 structures containing seven-membered cycles. Other researchers suggested that structures containing n -membered cycles with $n > 7$ are uncommon.⁴⁰ However, in the series of structures under consideration, there were about 60 compounds containing rings of this size. No structures containing four-membered cycles were detected.

Some examples of structures identified with the aid of StrucEluc operating in the common mode (i.e., without fragments found in the knowledge base and user-defined fragments) are presented in Table 1. The structures are accompanied by the main parameters, allowing the evaluation of the system to perform efficiently as a CASE tool.

Listed in the table of elucidated chemical structures is the following information: a reference to the article from which the NMR data were extracted; the molecular formula (MF) and molecular weight; the number of skeletal atoms, n ; the number of connectivities in the COSY and HMBC spectra; the number of nonstandard correlations in the 2D NMR data; k , the number of structures in the output file before and after

the removal of duplicates; t_g , the time of structure generation; T , the total time for solving the problem; and r_A , r_F , r_H , and r_Σ , the position of the correct structure in the output file ranked by the deviations calculated using different computational methods (see refs 7 and 8). Calculations were performed on a Pentium IV, 2.8 MHz PC.

The data allow us to conclude that the molecules used in the investigation discussed in this article were complex and characterized by the diversity of both their topology and chemical composition. The MF, molecular mass (M), and DBE averaged across all of the problems are represented by $\text{MF} = \text{C}_{25}\text{H}_{33}\text{O}_5\text{N}_1$, $M = 427$, and $\text{DBE} = 10$.

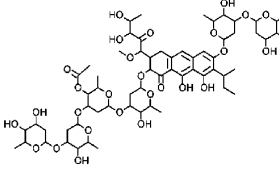
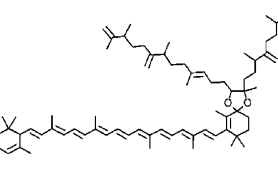
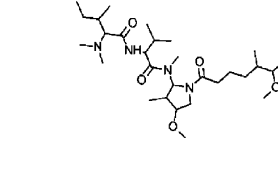
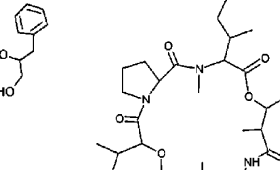
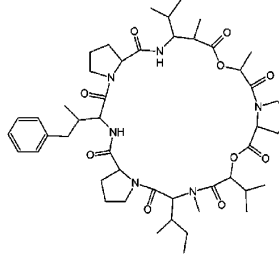
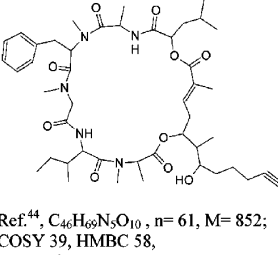
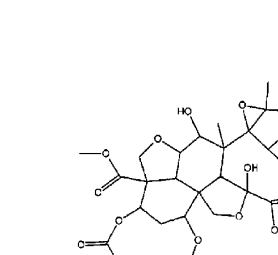
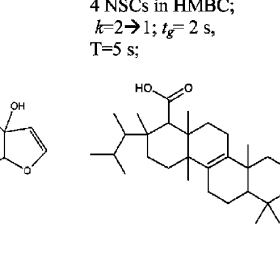
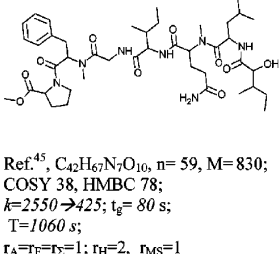
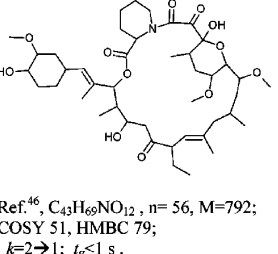
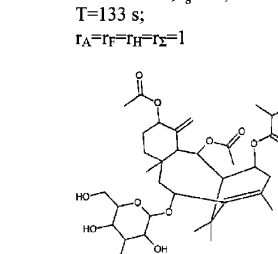
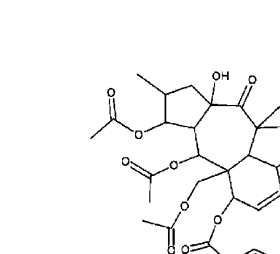
A comparison of Figure 5 with the results of statistical analysis on approximately 3.5 million structures published by Petitjean and Dubois⁵³ allows us to conclude that, when the size of structures successfully elucidated by the StrucEluc system is taken into account, this system can be, in principle, applied to at least 98% of all organic compounds encountered in chemical investigations.

2.3. Peculiarities of 2D NMR Spectral Data. Our experience has shown that the successful elucidation of a chemical structure by the StrucEluc system depends not only on the size of the structure but on the information contained within the experimental 2D NMR spectra and on the number of observed correlations relative to the theoretically expected value. The degree of overlap of the signals is, of course, a major consideration. The key point in the operation of the system is the generation of structures with the restrictions arising from both the HMBC and COSY data. The more correlations that are observed in the 2D NMR spectra, the correspondingly more restrictions that can be imposed during the generation process, thereby increasing the likelihood that the problem can be solved in a reasonable period of time.

The number of observed 2D NMR correlations depends on a series of complicating factors including not only the experiment and the parameters used when the experiment is performed but the dispersion of the proton and carbon NMR spectra, in addition to spatial effects. The number of correlations observed experimentally commonly turns out to be less than the number expected theoretically. This number is calculated on the basis of the assumption that all correlations corresponding to the coupling constants expressed as $^2,3J_{\text{HH}}$ and $^2,3J_{\text{CH}}$ should be observed in the COSY and HMBC spectra. In our previous works,^{7,8,27} we defined such correlations as “standard”. In those cases where there are $^{n>3}J_{\text{HH/CH}}$ couplings, the number of experimentally observed peaks can be greater than theoretically possible on the basis of the defined constraints.

Efforts have been made to determine the connection between the numbers of observed correlations by analyzing the results of problems studied to date. For this purpose, both theoretical COSY and HMBC spectra were generated for the entire set of problems and the MCDs were created from the calculated spectra. All problems were then solved in a batch mode. Under the specified conditions, most of the problems resulted in only one structure being generated. For 20 of the tasks, two to four structures were generated, and only 4 out of 250 tasks gave output files containing several tens of structures. The generation time for most of the tasks did not exceed 2–3 s (Pentium IV, 2.8 MHz). In an ideal case, information that can be obtained from 2D NMR spectra

Table 1. Examples of Solutions to Problems Solved Using the Common Mode

 <p>Ref.⁴¹, C₆₂H₉₂O₂₈, n=90, M=1285; COSY 37, HMBC 134 k=472→59; t_g=614 s; T=1500 s r_A=1, r_F=21, r_H=2, r_Σ=1</p>	 <p>Ref.⁴², C₇₄H₁₁₂O₂, n=86, M=1033; COSY 4, HMBC 224, 1 NSC in HMBC; k=4→1, t_g=24 s, T=37 s; r_A=r_F=r_H=r_Σ=1</p>	 <p>Ref.⁴⁷, C₄₀H₆₈N₄O₈, n=52, M=732; COSY 35, HMBC 76; k=224→224; t_g=8 s; T=184 s; r_A=r_F=r_Σ=1, r_H=2</p>	 <p>Ref.⁴⁸, C₄₁H₆₀N₄O₈, n=53, M=736; HMBC 113, 4 NSCs in HMBC; k=2→1; t_g=2 s, T=5 s;</p>
 <p>Ref.⁴³, C₄₈H₇₂N₆O₁₀, n=64, M=893; COSY 3, HMBC 88 k=5804→2882, t_g=784 s, T=3500 s; r_A=r_F=r_H=r_Σ=1</p>	 <p>Ref.⁴⁴, C₄₆H₆₈N₅O₁₀, n=61, M=852; COSY 39, HMBC 58, 3 NSCs in COSY; k=4→4, t_g=12 s, T=31 s; r_A=r_F=r_Σ=1, r_H=2</p>	 <p>Ref.⁴⁹, C₃₅H₄₄O₁₆, n=51, M=720; COSY 15, HMBC 66; k=105→105, t_g=4 s, T=133 s; r_A=r_F=r_H=r_Σ=1</p>	 <p>Ref.⁵⁰, C₃₈H₆₂O₁₀, n=48, M=678; HMBC 101 k=228→38; t_g=5 s; T=234 s; r_A=r_Σ=1, r_H=4, r_F=2</p>
 <p>Ref.⁴⁵, C₄₂H₆₇N₇O₁₀, n=59, M=830; COSY 38, HMBC 78; k=2550→425; t_g=80 s; T=1060 s; r_A=r_F=r_Σ=1; r_H=2, r_{MS}=1</p>	 <p>Ref.⁴⁶, C₄₃H₆₉NO₁₂, n=56, M=792; COSY 51, HMBC 79; k=2→1; t_g<1 s, T=7 s;</p>	 <p>Ref.⁵¹, C₃₅H₅₄O₁₂, n=47, M=666; COSY 19, HMBC 67, 1 NSC in HMBC; k=230→190, t_g=4 s, T=171 s; r_A=r_F=r_H=r_Σ=1</p>	 <p>Ref.⁵², C₃₃H₄₀O₁₁, n=44, M=612; HMBC 66, 1 NSC in HMBC; k=9→7, t_g=75 s, T=106 s; r_A=r_Σ=r_H=1, r_F=2,</p>

is sufficient to elucidate the structure of natural products of fairly high complexity.

When solving real problems, difficulties occur primarily due to the sparse numbers of experimentally observed correlations, the presence of extraneous correlations of "nonstandard" length (${}^nJ_{\text{HH/CH}}$, $n>3$), and the overlap of signals that leads to the appearance of ambiguous correlations.

Parts A and B of Figure 6 show the relationship between the numbers of observed and theoretically possible correlations in both the HMBC and COSY spectra. When creating Figure 6B, only those problems with available COSY spectra were used.

It is generally believed by others in this field that correlations of nonstandard length are observed only rarely.^{23,24,26} In this work, 114 out of 250 tasks (i.e., 45%) contained nonstandard correlations in the 2D NMR data. Figure 7 illustrates the distribution of the number of nonstandard correlations present across the 2D NMR data sets.

Obviously, the presence of nonstandard correlations in 2D NMR data is not as rare of an occurrence as some workers have suggested. In the examples cited in this work, a maximum of 15 nonstandard correlations was observed for a single structure. For the problem data sets as a whole, nonstandard correlations were observed in the COSY data in 52 cases and in the HMBC data in 85 cases. Indeed, it is quite likely with the higher sensitivity afforded by cryogenic NMR probes that the frequency with which nonstandard correlations will be observed will increase as access to spectrometers equipped with cryogenic probe capabilities broadens.⁵⁴

Numerous examples of structures whose HMBC spectra contain nonstandard correlations have been reported.⁵⁵ From our work, it is evident that spectral data can not only contain a large number of nonstandard correlations but also the deviation from the standard connectivity length can be up to three bonds to provide a ${}^6J_{\text{XH}}$ coupling constant. It is evident that the method reported in some works³⁹ to

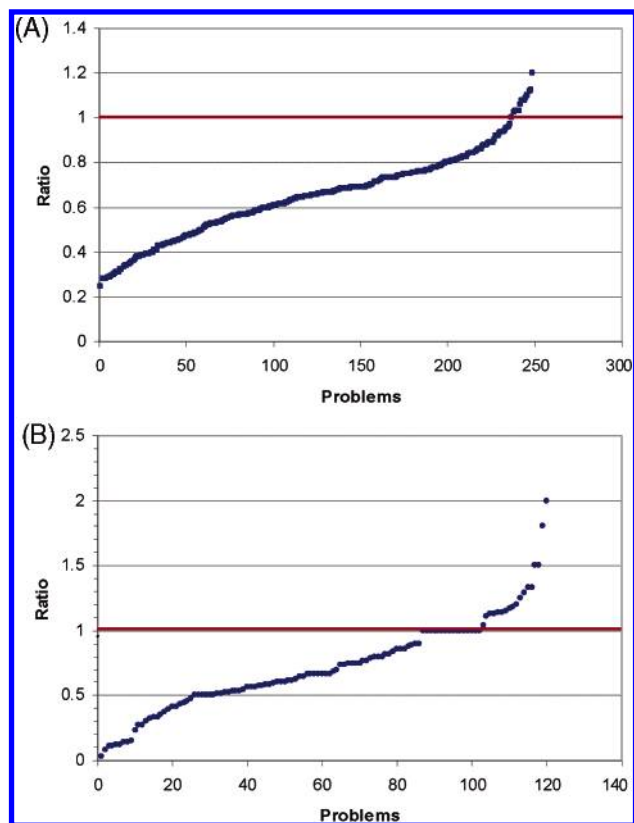


Figure 6. (A) Ratio of the number of experimental HMBC correlations vs the number of theoretical correlations across the 250 example problems. The problems are ordered in ascending order of the ratio. (B) Ratio of the number of COSY correlations vs the number of theoretical correlations across the 120 example problems that included COSY data. The problems are ordered in ascending order of the ratio.

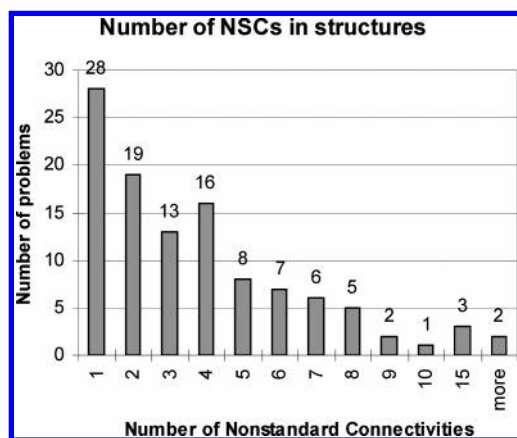


Figure 7. Histogram illustrating the number of nonstandard connectivities contained within the 114 problems.

overcome contradictions in 2D NMR data by lengthening all of the correlations by one bond (using couplings $^{2-4}J_{\text{HH}}$ and $^{2-4}J_{\text{CH}}$ by default) would be insufficient to elucidate structures in the presence of correlations characterized by $^{5,6}J_{\text{XH}}$ coupling constants. In our work,²⁷ we suggested approaches for addressing this issue in such situations. Experience has shown that the application of *fuzzy* structure generation²⁷ can deliver great value, and the methodology and advantages of this approach will be discussed in a separate publication.

2.4. Some Important Properties of Solutions to Problems. 2.4.1. Time of Problem Solving.

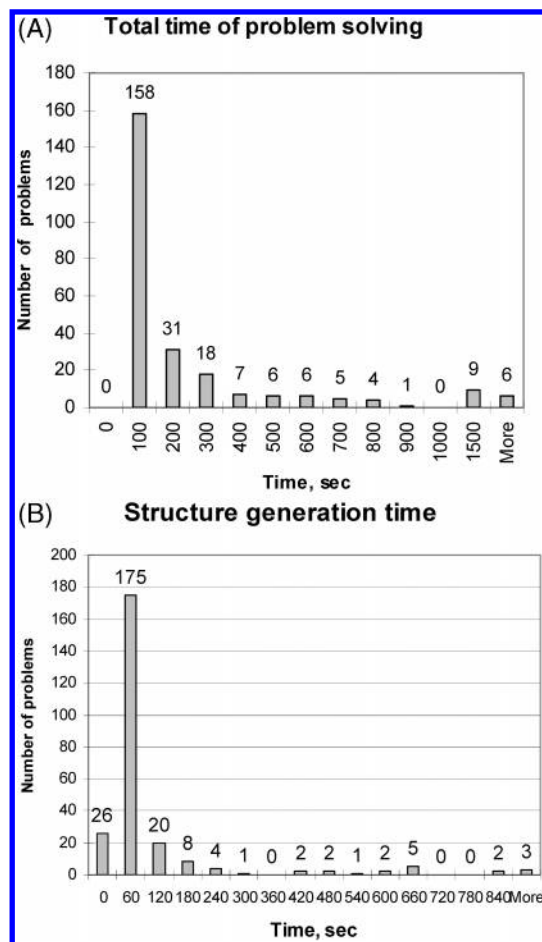


Figure 8. (A) Histogram showing the total time required to elucidate the structures within the problem set. (B) Distribution of structure generation time across the entire problem set.

necessary to solve a problem, T_{tot} , can be represented as a sum of the following components:

$$T_{\text{tot}} = t_{\text{MCD}} + t_g + t_{\text{flt}} + t_{\text{spr}}$$

where t_{MCD} is the time required to create the MCD (when the problem is solved in the common mode,^{7,8} this value is practically zero); t_g is the time of structure generation; t_{flt} is the time consumed in performing spectral filtering on the generated structural file (a process applied just after structures are generated in to save free disk space) and the removal of duplicates; t_{spr} is the time necessary to perform spectrum prediction on the output file and the ranking of the structures in ascending order of d_A deviation (see below and refs 7 and 8). The histogram presented in Figure 8A shows that the majority of problems were solved in less than 2 min and that, as a rule, this time does not exceed 20 min.

Structure generation times are shown as a histogram in Figure 8B. These data demonstrate that t_g comprises only a small part of the T_{tot} value. In particular, for 200 out of 250 problems, t_g is less than 1 min.

2.4.2. Efficiency of Filtering Generated Structures.

Figure 9A shows a histogram plot illustrating the number of structures output into the resultant file prior to filtering and the removal of duplicates.

Analysis of Figure 9A allows us to conclude that, for 75% of the problems, the output file of generated structures contains less than 1000 structures and only in 2% of the cases

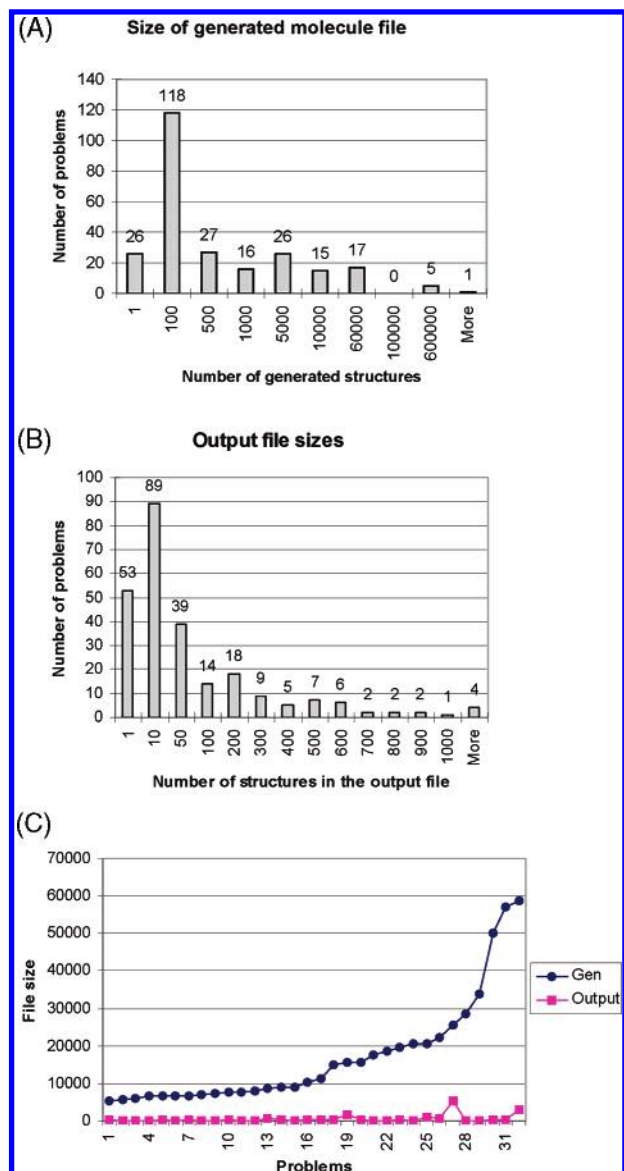


Figure 9. (A) Histogram of the number of structures in the output file prior to filtering and removal of duplicates. (B) Distribution of structures in the reduced output files resulting from filtering and removal of duplicates. (C) Reduction in the number of generated structures after spectral and structural filtering. The data are provided for a subset of problems with output files containing between 5000 and 50 000 structures.

does the output file grow to around half a million structures. Recalling that the number of potential isomers for a molecular formula comprising a natural product can be astronomical (see Figure 1), the results obtained can be considered as particularly successful. The results again provide evidence that the statement made by other workers^{22,23} regarding the limited possibilities of deterministic expert systems because of the potential threat of a combinatorial explosion is in fact unjustified.

Figure 9B illustrates the distribution of the size of the output file obtained following both spectral and structural filtering of the initial file of generated structures followed by the removal of duplicates. A comparison of the distributions presented in Figure 9A,B illustrates the high efficiency of the filtering process because the number of generated structures is reduced dramatically as a result of the application of the filtering procedures. The filtering efficiency is

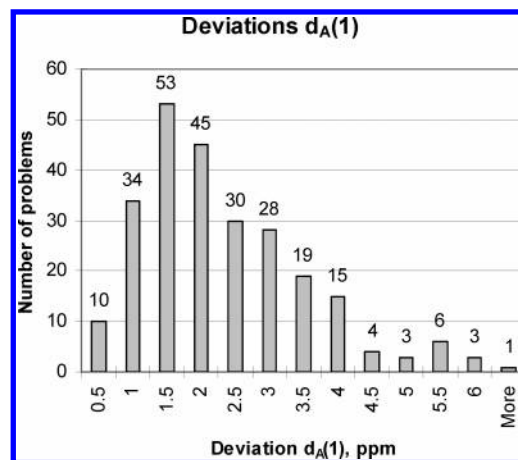


Figure 10. Histogram illustrating the distribution of problems with deviations corresponding to the first-ranked structures, $d_A(1)$ values.

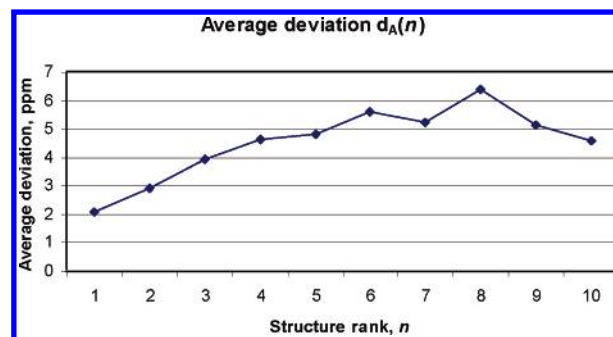


Figure 11. Average d_A magnitudes corresponding to the first 10 structures of the ranked output file.

obvious in Figure 9C where the initial and final output files are shown for problems with output files containing between 5000 and 50 000 structures.

2.4.3. Evaluation of Methods for Selection of the Most Probable Structure.

A three-step method for selection of the most probable structure contained within the output file obtained using StrucEluc has been fully described previously.^{7,8} The d_A values represent the average deviations of the “accurately” calculated ^{13}C NMR chemical shifts relative to the experimental shifts. These values play a decisive role in the process of selecting the correct structure. It is generally assumed that the first-ranked structure with the minimum d_A value is the most probable structure match. For the problem sets examined in this report, for 93% of the problems, d_A ranking placed the correct structure in the first position. The distribution of problems with first-order rankings is presented in Figure 10, where $d_A(1)$ values, the deviations corresponding to the first-ranked structures, are presented. In about 60% of the cases, the $d_A(1)$ value is less than 2 ppm, and the average value is 2.09 ppm. To reveal the general trend of d_A values within a ranked output file, the average magnitudes corresponding to the 10 first structures in the ranked output file were calculated (see Figure 11).

The figure shows that d_A deviations increase for the first four structures by an increment of approximately 1 ppm from each structure to the next. Average similarity coefficients were also considered for the 10 structures as represented in Figure 12. The similarity coefficients were calculated by comparing all structures with that ranked first. Recall that the first-ranked structure was correct in 93% of the cases. Figure 12 shows that the average similarity coefficients drop

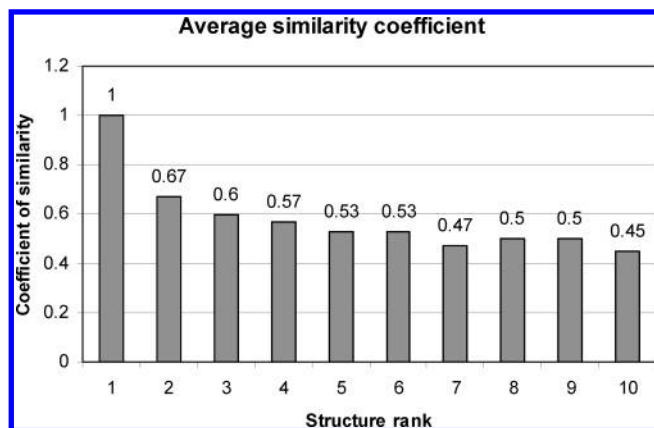


Figure 12. Average similarity coefficients of the first 10 ranked structures. The first structure was the reference structure for similarity.

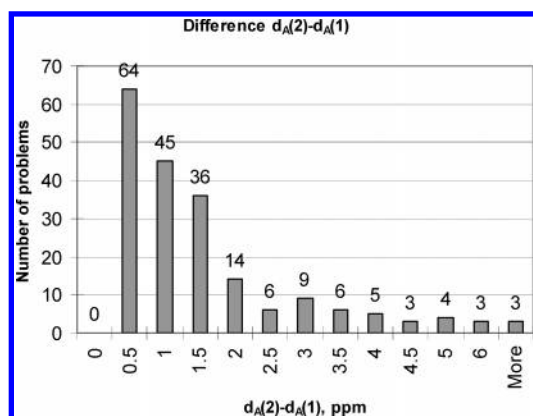


Figure 13. Distribution of problems as a function of the difference in deviations between the first and second ranked structures, $d_A(2) - d_A(1)$ values.

slowly starting from the second structure, whose average similarity coefficient is equal to 0.67. The observed dependence confirms that, in general, the main assumption common to molecular spectroscopy is that similar structures have similar spectra.

Experience has shown that the probability of coincidence of the first-ranked structure with the correct one sufficiently depends on the difference Δ_{2-1} between $d_A(2)$ and $d_A(1)$: the greater the difference, the greater the probability that the first-ranked structure is the target structure. The distribution of Δ_{2-1} values for output files containing more than one structure across the problem set is presented in Figure 13. Investigations performed on a large number of problems confirmed the following empirical fact, which was established by us earlier:⁵⁶ if the value $\Delta_{2-1} > 1$ ppm, then the structure ranked first has a significant probability of being the correct solution to the problem. At the same time, it was also found that the correct structures were distinguished by structural ranking in many cases, even when the difference Δ_{2-1} was very small. It turned out that only in ca. 20 problems was the correct structure not placed in the first position (see the problem distribution in Figure 14).

Figure 14 illustrates for a series of examples that the correct structure was ranked as either second or third for eight problems in each case. Analysis of the improperly ranked solutions establishes that the cause of the incorrect structure ranking is most frequently the absence of appropriate structures from the database that allow precise prediction

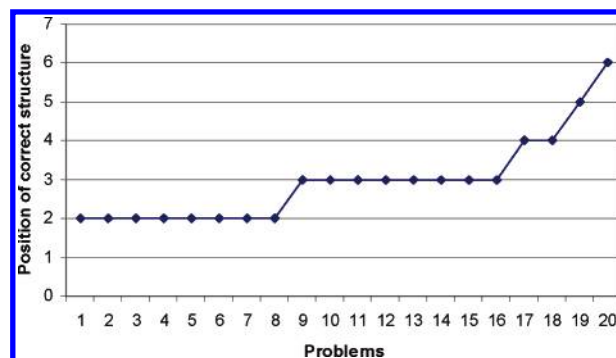


Figure 14. Position of the correct structure in output files for 20 problem sets where the first-ranked structure was *not* the correct solution.

of the chemical shifts of certain carbon atoms characterized by unusual environments. In other cases, the carbon atom environment was not poorly represented per se, but the chemical shift of the given atom was unusual because of some particular spatial effects. The following empirical observation was deduced: for the majority of *improperly ranked* structures, the difference Δ_{2-1} is less than 1. A large $d_A(1)$ value in combination with a small Δ_{2-1} value is suggestive of either improper structural ranking or incorrect ^{13}C chemical shift assignment, the latter due to misassignment by the investigator.

Traditionally, the developers of expert systems aspire to minimize the structure generation time and the size of the output file. However, the results presented above indicate that an output file containing many thousands of structures cannot seriously hamper the selection of the most probable structure because the methods described push the actual structure to the top of the rank-ordered file.

2.7. Deterministic versus Stochastic Systems. In his paper,²³ Steinbeck compares the results of structure elucidations performed by the deterministic system LUCY³ and the “stochastic” system SENECA on the basis of a simulated annealing (SA) structure generation algorithm. Both systems were developed by the author.

Steinbeck concluded that “the data in Table 2 are in agreement with a polynomial behavior, which, in contrast to the exponential growth of time for the LUCY calculation, strongly suggests that the SENECA system will be able to tackle problems too large for deterministic algorithms”.

We argue against this conclusion for the following reasons:

(a) The comparison is not with deterministic systems but rather for a series of runs for the LUCY system. In the case of polycarpol, the generation time was 120 times higher than that of monochaetin, and we believe this reflects a peculiarity within the system. Our calculations show that, with the StrucEluc system,^{7,8} the comparative generation times were only fractions of a second for tasks 1–3 given in Table 2 when executed on a 500 MHz Pentium III computer in an automated mode and without the interference of a user. The data obtained using other deterministic systems^{1,6,14,46} also contradict Steinbeck’s conclusion. As demonstrated above for deterministic systems, the structure generation time not only depends on the number of heavy atoms in a molecule but also, and to a higher degree, on the nature of the information captured within the 2D NMR spectra.

(b) From the examples discussed (with molecular formulas of $\text{C}_{15}\text{H}_{28}\text{O}_2$, $\text{C}_{18}\text{H}_{20}\text{O}_5$, and $\text{C}_{30}\text{H}_{48}\text{O}_2$), it cannot be concluded

Table 2. Calculation Time and Number of Iterations Required for Three Example Compounds^a Studied with SENECA

name	molecular formula	calculation time		number of SA iterations performed
		LUCY	SENECA	
eurabidiol	C ₁₅ H ₂₈ O ₂	29 s	5 min	90 000
monochaetin	C ₁₈ H ₂₀ O ₅	16 s	9 min	250 000
polycarpol	C ₃₀ H ₄₈ O ₂	33 min	12 min	350 000

^a The calculation was performed using desktop PCs operating at processor speeds of 600 MHz.

that the “SENECA system will be able to tackle problems too large for deterministic algorithms”. This conclusion is purely speculative and has not been proven by any examples reported to date. The simulated annealing algorithm has not been successfully applied to the analysis of problems that could not be solved by deterministic systems.

(c) The comparison of the number of structures generated by the SENECA simulated annealing program and the StrucEluc deterministic system shows that, in the case of monochaetin, StrucEluc produces 18 structures at the first stage and, in the case of polycarpol, only 6. The first stage is before spectral filtering is performed and identical structures are removed. This demonstrates that the exhaustive search in StrucEluc is performed within a highly restricted area of the full isomer space, while SENECA searches through hundreds of thousands of structures.

(d) The underestimation of the performance of deterministic expert systems is likely explained by the fact that skeptics do not take into account that these systems are intended to mimic an expert's reasoning. Experienced spectroscopists can determine the structures of molecules with formulas capable of generating 10²⁰–10³⁰ isomers. In doing so, a spectroscopist does not search the full isomer space but identifies the probable structures by “tunneling”. The possibility of performing computer-assisted structure elucidation using such a “tunneling” approach is the most important capability of a well-designed expert system.

Steinbeck has also expressed an opinion²³ that the scoring function itself makes the simulated annealing algorithm more flexible by honoring the exceptions to the rules of interpreting cross signals in 2D NMR spectra. While considering structure generation in the presence of connectivities of nonstandard lengths, the author concluded that deterministic algorithms will fail and CASE systems will not find the correct structure. As was mentioned above, structure generation tools capable of coping with nonstandard connectivity lengths can be dealt with in a deterministic system, and this considerably reduces the structure space search area and generation time. Steinbeck rightly notes, however, that the number of structures and the generation time will grow and suggests that deterministic algorithms will not honor the rareness of four-bond correlations in HMBC spectra, while “in SENECA, this problem can be easily overcome by the scoring system.” Steinbeck's declaration that four-bond HMBC correlations are a rare phenomenon is arguable. According to our observations described above, four-bond or even higher-order correlations in both HMBC and COSY spectra occur quite regularly and will likely be encountered even more often as investigators have increasing access to instruments equipped with cryogenic NMR probes.⁵⁴ The computer-assisted structure elucidation of 250 natural products showed that nonstandard

correlations were observed in 2D NMR spectra in almost half of the tasks examined (see Figure 7). To the best of our knowledge, the statement that “in SENECA, this problem can be easily overcome by the scoring system” has not yet been proven.

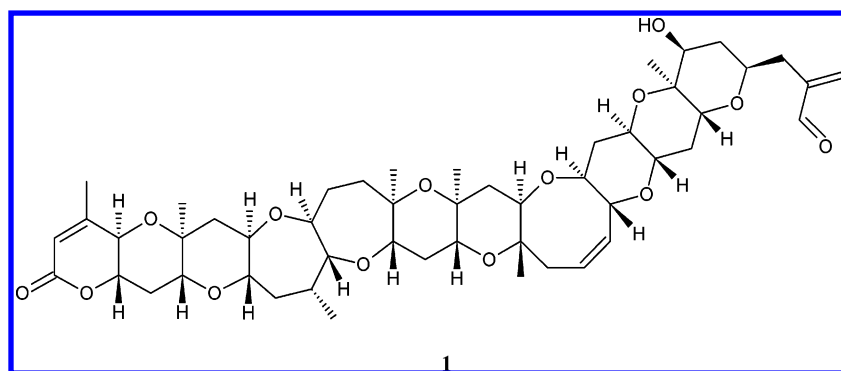
Faulon elaborated a stochastic structure generator called SIGNATURE.⁵⁷ He declared that “for large molecular compounds, deterministic techniques are not applicable to resolve the problem of structure elucidation. In such an instance one has to use a stochastic structure generation.” This suggestion has also not been confirmed experimentally. To illustrate the superiority of stochastic generators over deterministic ones, Faulon compared the efficiency of a very slow and incorrectly working deterministic structure generator with that of SIGNATURE.⁵⁷ For the molecular formula C₈H₁₀, Faulon's deterministic structure generator produced 4008 isomers in 353 s on an SGI Personal Iris Workstation. He commented, “For comparison, the stochastic version of the structure generator estimated the number of isomers to be 3,399. To calculate this number the stochastic generator was run for 16.5 s CPU time ...”. By comparison, StrucEluc has generated 4679 isomers with the C₈H₁₀ composition in 0.023 s (Pentium IV, 2.8 MHz). The results described require little additional commentary.

Despite these results, the potential use of stochastic algorithms remains of interest. The successful selection of a scoring function allows the algorithm to effectively direct the search to an isomer subspace where the structures comply with the constraints imposed by the 2D NMR data. An advantage of the approach is that very few assumptions are made prior to the structure elucidation run. Large fragments do not need to be deduced from the spectral data in the preprocessing run, and no particular hybridization states are assumed. This does not indicate that additional data may not be necessary in certain situations, for example, when a large molecule has a structure with a deficit of hydrogen atoms and a significant number of heteroatoms. Deterministic systems can successfully solve such tasks as evidenced by Table 1 and the materials presented in our previous publications.^{7–12,27–29}

In the examples given by Steinbeck,²³ the number of heteroatoms varies only from two to five nuclei, while the problems solved with the use of the deterministic StrucEluc^{7,8} system contain structures where the number of various heteroatoms can be up to 30 nuclei.

Another report published by Han and Steinbeck²⁴ demonstrates that the SENECA system has markedly expanded the size of the molecular structures (20 heavy atoms) that were examined with the genetic algorithms of Meiler and Will²⁵ using 1D NMR. This was achieved using 2D NMR spectra and a strategy for structure generation that allows for the application of a series of control parameters while simultaneously having a well-selected fitness function. The authors again did not succeed in demonstrating that the stochastic algorithms could solve tasks of similar or greater complexity than those solved by deterministic systems. The results do suggest that this approach may indeed find increasing applications in spectroscopic analysis, and the task remains to validate the performance and reliability of the approach. Even for small molecules with ~15 skeletal atoms (see Figure 1), the number of possible isomers is immense (>10¹²), and the ability of the genetic algorithm to identify

Chart 1



the target structure in a short time is promising. It is quite possible that future deterministic and stochastic expert systems will become complementary in the solution of specific structural problems. It can be expected that, if certain components of a deterministic system are implemented in a system utilizing a stochastic algorithm, then overall the effectiveness of the system may be enhanced.

There is, however, a stage of molecular structure elucidation that has been shown to demand the application of a genetic algorithm. The relative stereochemistry of the most probable structure and calculation of its 3D molecular model has been addressed in this manner. Generally, the final step in contemporary structure characterization efforts is to define the relative and, if possible, absolute stereochemistries of the individual centers. NMR-based determination of relative stereochemistry is based on the nuclear Overhauser effect (NOE), which is dependent on the distance separating the cross-relaxing nuclides.⁵⁸ Typically, NOESY or ROESY two-dimensional NMR experiments or their selective 1D analogues are used to provide the data for analysis. In a recent study, we extended the capabilities of StrucEluc to the derivation of multiple relative stereocenters in complex structures such as taxol and brevetoxin. This work has been described in detail elsewhere.³⁰ An appropriate penalty function was suggested³⁰ that can be minimized by calculation for *all rigid* stereoisomeric structures or by using a genetic algorithm to limit the number of stereoisomers that need to be investigated. To improve genetic algorithm convergence, efficient methods of parameter optimization were suggested and compared.

To challenge the system, the structure of *brevetoxin B* (**1**) was examined (Chart 1). This remarkable structure includes 11 rings, 23 stereogenic centers, and 3 carbon-carbon double bonds. The CPU time necessary for running over all ~8.4

million stereoisomers corresponding to this structure was estimated to be about 1 month for a Pentium IV computer.

In Figure 15, the structure of brevetoxin B from X-ray studies (yellow) and the structure obtained from the best chromosome in the final pool for our system for stereochemistry determination (blue) are shown as superimposed structures. In this case, the configurations of all of the stereocenters in both structures are the same. Even the conformations of the most “flexible” seven- and eight-membered rings are similar. This demonstrates the power of the approach we have described to facilitate the identification of relative stereochemistry in a complex molecule containing multiple stereocenters.

3. CONCLUSIONS

This work was conducted for the purpose of evaluating the capabilities of deterministic expert systems for elucidating molecular structures from 2D NMR data. To achieve this aim, a comprehensive analysis of the databases making up the advanced expert system StrucEluc was performed. The structural features of 250 molecules belonging to the class of natural products that were identified with the aid of the system under study were investigated. Simultaneously, the influence of the nature of the correlations contained within the 2D NMR spectral data used for elucidation was studied. The program is shown to cope with those problems where 2D NMR data were assumed to contain an unknown number of nonstandard (corresponding to ${}^nJ_{\text{HH,CH}}$, $n > 3$) correlations of unknown lengths.

It was conclusively demonstrated that the deterministic expert system StrucEluc possessed algorithms allowing the program to elucidate the chemical structures of complex natural products containing at least 100 skeletal atoms. The

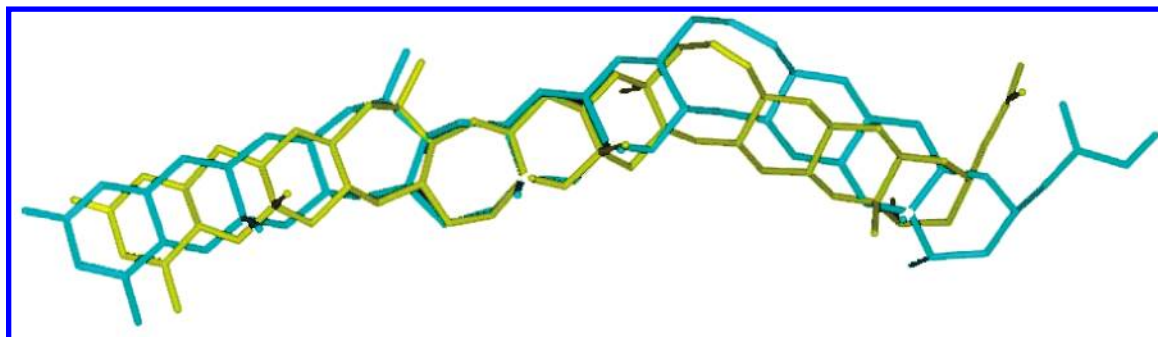


Figure 15. X-ray crystal structure of brevetoxin B (yellow) and the 3D model of the best stereoisomer calculated by StrucEluc. Small differences in the bond angles of some of the more flexible rings are present, but all stereocenters have been properly oriented.

deterministic algorithm for structure generation in combination with an intelligent strategy of the most probable structure selection provides complete solution of a problem in a short time—from several seconds up to several minutes. This result alone disproves the beliefs of other workers that molecules containing 30 skeletal atoms were an upper limit for study by deterministic expert systems and only stochastic algorithms of molecular structure generation—simulated annealing and genetic algorithms—will allow solving real problems of great dimensionality.

At the same time, it was shown that genetic algorithms could be efficiently applied within a deterministic expert system. Recently, we extended our system³⁰ to make it capable of determining the relative stereochemistry of complex rigid structures or molecules containing rigid substructures that contain a large number of stereocenters. To solve this problem, a genetic algorithm was used. Including stereochemical assignment capabilities into StrucEluc allowed the automation of all stages of identification, from the spectra representing an unknown compound to elucidation, stereocenter identification, and 3D model generation.

REFERENCES AND NOTES

- (1) Munk, M. E. Computer-Based Structure Determination: Then and Now. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 997–1009.
- (2) Peng, C.; Yuan, S.; Zheng, C.; Hui, Y. J. Efficient Application of 2D NMR Correlation Information in Computer-Assisted Structure Elucidation of Complex Natural Products. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 805–813.
- (3) Steinbeck, C. Lucy – A Program For Structure Elucidation From NMR Correlation Experiments. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 1984–1986.
- (4) Nuzillard, J.-M.; Massiot, G. Logic for Structure Determination. *Tetrahedron* **1991**, *47*, 3655–3664.
- (5) Funatsu, K.; Susuta, Y. S.; Sasaki, S. Introduction of Two-Dimensional NMR Spectral Information to an Automated Structure Elucidation System CHEMICS. Utilization of 2D-INADEQUATE Information. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 6–11.
- (6) Lindel, T.; Junker, J.; Köck, M. 2D-NMR-Guided Constitutional Analysis of Organic Compounds Employing the Computer Program Cocon. *Eur. J. Org. Chem.* **1998**, *3*, 573–577.
- (7) 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 Data: Application of the StrucEluc System. *Magn. Reson. Chem.* **2003**, *41*, 359–372.
- (8) Elyashberg, M. E.; Blinov, K. A.; Molodtsov, S. G.; Williams, A. J.; Martin, G. E. *Structure Elucidator: A Versatile Expert System for Molecular Structure Elucidation from 1D and 2D NMR Data and Molecular Fragments*. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 771–792.
- (9) Blinov, K. A.; Elyashberg, M. E.; Molodtsov, S. G.; Williams, A. J.; Martirosian, E. R. Application of ¹H–¹H, ¹³C–¹H and ¹⁵N–¹H 2D NMR Correlations for the Complex Molecular Structure Elucidation by Means of an Expert System. *Fresenius' J. Anal. Chem.* **2001**, *369*, 709–714.
- (10) Martin, G. E.; Hadden, B. D.; Russell, C. E.; Kaluzny, D. J.; Guido, J. E.; Duholke, W. K.; Stiemsma, B. A.; Thamann, T. J.; Crouch, R. C.; Blinov, K. A.; Elyashberg, M. E.; Martirosian, E. R.; Molodtsov, S. G.; Williams, A. J.; Sharif, P. L., Jr. Identification of Degradants of a Complex Alkaloid Using NMR Cryoprobe Technology and ACD/Structure Elucidator. *J. Heterocycl. Chem.* **2002**, *39*, 1241–1250.
- (11) Blinov, K. A.; Elyashberg, M. E.; Martirosian, E. R.; Molodtsov, S. G.; Williams, A. J.; Sharif, P. L.; Schiff, P. L., Jr.; Crouch, R. C.; Martin, G. E.; Hadden, C. E.; Guido, J. E.; Mills, K. A. Quindolinocryptotackeine: 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.
- (12) Sharman, G. J.; Jones, I. C.; Parnell, M. J.; Willis, M.; Carlson, D. V.; Williams, A.; Elyashberg, M. E.; Blinov, K. A.; Molodtsov, S. G. Automated Structure Elucidation of Two Unexpected Reaction Products in a Reaction of an α,β -Unsaturated Pyruvate. *Magn. Reson. Chem.* **2004**, *42*, 567–572.
- (13) Peng, C.; Yuan, S.; Zheng, C.; Shi, Z.; Wu, H. Practical Computer-Assisted Structure Elucidation for Complex Natural Products: Efficient Use of Ambiguous 2D NMR Correlation Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 539–546.
- (14) Peng, C.; Bodenhausen, G.; Qiu, S.; Fong, H. H. S.; Farnsworth, N. R.; Yuan, S.; Zheng, C. Computer-Assisted Structure Elucidation: Application of CISOC–SES to the Resonance Assignment and Structure Generation of Betulinic Acid. *Magn. Reson. Chem.* **1998**, *36*, 267–278.
- (15) Almanza, G.; Balderama, L.; Labbe, C.; Lavaud, C.; Massiot, G.; Nuzillard, J.-M.; Connolly, J. D.; Farrugia, L. J.; Rycroft, D. S. Clerodane Diterpenoids And Ursane Triterpenoid From *Salvia Haenkei*. Computer-Assisted Structural Elucidation. *Tetrahedron* **1997**, *53*, 14719–14728.
- (16) Nuzillard, J.-M. Determination Assistee Par Ordinateur De La Structure Des Molecules Organiques. *J. Chim. Phys.* **1998**, *95*, 169–177.
- (17) Mulholland, D.; Randrianarivelojosia, M.; Lavaud, C.; Nuzillard, J.-M.; Schwikard, S. L. Limonoid Derivatives From *Astrotrichilia voamatata*. *Phytochemistry* **2000**, *53*, 115–118.
- (18) Mulholland, D.; Schwikard, S. L.; Sandor, P.; Nuzillard, J.-M. Delevoyin C, a Tetranortriterpenoid from *Entendophragma delevoyi*. *Phytochemistry* **2000**, *53*, 465–468.
- (19) Belofsky, G. N.; Anguera, M.; Jensen, P. R.; Fenical, W.; Köck, M. Oxepinamides A–C and Fumiquinazolines H–I: Bioactive Metabolites from a Marine Isolate of a Fungus of the Genus *Acremonium*. *Eur. J. Org. Chem.* **1999**, 579–586.
- (20) Urban, S.; Blunt, J. W.; Munro, M. H. G. Coproverdine, a Novel, Cytotoxic Marine Alkaloid from a New Zealand Ascidian. *J. Nat. Prod.* **2002**, *65*, 1371–1373.
- (21) Assmann, M.; van Soest, R. W. M.; Köck, M. New Antifeedent Bromopyrrole Alkaloid From The Caribbean Sponge *Stylissa caribica*. *J. Nat. Prod.* **2001**, *64*, 1345–1347.
- (22) Faulon, J.-L. Stochastic Generator of Chemical Structure. 2. Using Simulated Annealing To Search the Space of Constitutional Isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 731–740.
- (23) Steinbeck, C. SENECA: A Platform-Independent, Distributed, and Parallel System for Computer-Assisted Structure Elucidation in Organic Chemistry. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1500–1507.
- (24) Han, Y.; Steinbeck, C. Evolutionary-Algorithm-Based Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 489–498.
- (25) Meiler, J.; Will, M. Genius: A Genetic Algorithm for Automated Structure Elucidation from C-13 NMR Spectra. *J. Am. Chem. Soc.* **2002**, *124*, 1868–1871.
- (26) Steinbeck, C. Recent Developments in Automated Structure Elucidation of Natural Products. *Nat. Prod. Rep.* **2004**, *21*, 512–518.
- (27) Molodtsov, S. G.; Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Martin, G. M.; Lefebvre, B. Structure Elucidation from 2D NMR Spectra Using the *StrucEluc* Expert System: Detection and Removal of Contradictions in the Data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1737–1751.
- (28) Elyashberg, M.; Blinov, K.; Williams, A.; Molodtsov, S.; Martirosian, E. Application of a New Expert System for the Structure Elucidation of Natural Products from 1D and 2D NMR Data. *J. Nat. Prod.* **2002**, *65*, 693–703.
- (29) Elyashberg, M. E.; Blinov, K. A.; Martirosian, E. R.; Molodtsov, S. G.; Williams, A. J.; Martin, G. E. Automated Natural Product Structure Elucidation – The Benefits of a Symbiotic Relationship between the Spectroscopist and the Expert System. *J. Heterocycl. Chem.* **2003**, *40*, 1017–1029.
- (30) Smurnyy, Y. D.; Elyashberg, M. E.; Blinov, K. A.; Lefebvre, B.; Martin, G. E.; Williams, A. J. Computer-Aided Determination of Relative Stereochemistry and 3D Models of Complex Organic Molecules from 2D NMR Spectra. *Tetrahedron* **2005**, *61/42*, 9980–9989.
- (31) Quin, L. D.; Williams, A. J. *Practical Interpretation of P-31 NMR Spectra and Computer Assisted Structure Verification*; Advanced Chemistry Development, Inc.: Toronto, Canada, 2004; pp 101–116.
- (32) ACD/NMR predictors, Advanced Chemistry Development, 110 Yonge Street, 14th floor, Toronto, ON, M5H 3V9, Canada. <http://www.acdlabs.com>: Prediction suite includes ¹H, ¹³C, ¹⁵N, ¹⁹F, and ³¹P NMR prediction.
- (33) *Specinfo*; Chemical Concepts GmbH: Weinheim, Germany.
- (34) Robien, W. *CSEARCH*; Universität Wien: Wien, Austria. http://felix.orc.univie.ac.at/~wr/csearch_server_info.html.
- (35) Meiler, J.; Meusinger, R.; Will, M. Fast Determination of ¹³C NMR Chemical Shifts Using Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169–1176.
- (36) Meiler, J.; Maier, W.; Will, M.; Meusinger, R. Using Neural Networks for ¹³C NMR Chemical Shift Prediction—Comparison with Traditional Methods. *J. Magn. Reson.* **2002**, *157*, 242–252.

- (37) Denisov, A. M. *Elements of the Theory of Inverse Problems (Inverse and Ill-Posed Problems)*; Brill Academic Publishers: Leiden, The Netherlands, 1999.
- (38) 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.
- (39) Schulz, K.-P.; Korytko, A.; Munk, M. E. Applications of a HOUDINI-Based Structure Elucidation System. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1447–1456.
- (40) Peng, C.; Yuan, S.; Zheng, C.; Hui, Y.; Wu, H.; Ma, K.; Han, X. Application of Expert System CISOC–SES to the Structure Elucidation of Complex Natural Products. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 814–819.
- (41) 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.
- (42) Okada, S.; Tonegawa, I.; Matsuda, H.; Murakami, M.; Yamaguchi, K. Botryoxanthin B and a Botryoxanthin A from the Green Microalga *Botryococcus Braunii* Kawaguchi-1. *Phytochem.* **1998**, *47*, 1111–1115.
- (43) Davies-Coleman, M.; Dzeha, T. M.; Gray, C. A.; Hess, S.; Pannell, L. K.; Hendricks, D. T.; Arendse, C. E. Isolation of Homodolastatin 16, a New Cyclic Depsipeptide from a Kenyan Collection of *Lyngbya majuscula*. *J. Nat. Prod.* **2003**, *66*, 712–715.
- (44) Williams, P. G.; Yoshida, W. Y.; Quon, M. K.; Moore, R. E.; Paul, V. J. The Structure of Palau'amide, a Potent Cytotoxin from a Species of the Marine Cyanobacterium *Lyngbya*. *J. Nat. Prod.* **2003**, *66*, 1545–1549.
- (45) 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.
- (46) Junker, J.; Maier, W.; Lindel, T.; Köck, M. Computer-Assisted Constitutional Assignment of Large Molecules: Cocon Analysis of Ascomycin. *Org. Lett.* **1999**, *1*, 737–740.
- (47) Horgen, F. D.; Kazmierski, E. B.; Westenburg, H. E.; Yoshida, W. Y.; Scheuer, P. J. Malevamide D: Isolation and Structure Determination of an Isodolastatin H Analogue from the Marine Cyanobacterium *Symploca hydroides*. *J. Nat. Prod.* **2002**, *65*, 487–492.
- (48) Nogle, L. M.; Gerwick, W. H. Isolation of Four New Cyclic Depsipeptides, Antanapeptins A–D, and Dolastatin 16 from a Madagascar Collection of *Lyngbya majuscula*. *J. Nat. Prod.* **2002**, *65*, 21–24.
- (49) Ley, S. V.; Doherty, K.; Massiot, G.; Nuzillard, J.-M. Connectivist Approach to Organic Structure Determination. LSD-Program Assisted NMR Analysis Of The Insect Antifeedant Azadirachtin. *Tetrahedron* **1994**, *50*, 12267–12280.
- (50) Joshi, B. K.; Gloer, J. B.; Wicklow, D. T. Bioactive Natural Products from a Sclerotium-Colonizing Isolate of *Humicola fuscoatra*. *J. Nat. Prod.* **2002**, *65*, 1734–1737.
- (51) Shi, Q.-W.; Sauriol, F.; Mamer, O.; Zamir, L. O. New Minor Taxane Derivatives from the Needles of *Taxus canadensis*. *J. Nat. Prod.* **2003**, *66*, 1480–1485.
- (52) Gustafson, K. R.; Blunt, J. W.; Munro, M. H. G.; Fuller, R. W.; McKee, T. C.; Cardellina, J., II; McMahon, J. B.; Cragg, G. M.; Boyd, M. R. The Guttiferones, HIV–Inhibitory Benzophenones. *Tetrahedron* **1992**, *48*, 10093–10102.
- (53) Petitjean, M.; Dubois, J.-E. Topological Statistics On Large Structural File. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 332–343.
- (54) Martin, G. E. Small Volume and High Sensitivity NMR Probes. *Annual Reports on NMR Spectroscopy*; Webb, G. A., Ed.; Elsevier: Amsterdam, 2005; pp 1–96.
- (55) Araya-Maturana, R.; Delgado-Castro, T.; Cardona, W.; Weiss-López, B. E. Use of Long-Range C–H ($^nJ_{n>3}$) Heteronuclear Multiple Bond Connectivity in the Assignment of the ^{13}C NMR Spectra of Complex Organic Molecules. *Curr. Org. Chem.* **2001**, *5*, 253–263.
- (56) Elyashberg, M. E.; Blinov, K. A.; Martirosian, E. R. A New Approach to the Computer-Aided Molecular Structure Elucidation: Expert System STRUCTURE ELUCIDATOR. *Lab. Autom. Inf. Manage.* **1999**, *34*, 15–30.
- (57) Index of /~jfaulon/SIGNATURE. <http://www.cs.sandia.gov/~jfaulon/SIGNATURE>.
- (58) Neuhaus, D.; Williamson, M. P. *The Nuclear Overhauser Effect in Structural and Conformational Analysis*, 2nd ed.; Wiley: New York, 2000.

CI050469J