Application of the Maximal Common Substructure Algorithm to Automatic Interpretation of ^{13}C-NMR Spectra

Lingran Chen and Wolfgang Robien*

Department of Organic Chemistry, University of Vienna, Währinger Strasse 38, A-1090 Vienna, Austria

Received January 3, 1994®

A novel program has been developed for deducing automatically the most common structural fragments from a set of similar structures by means of the maximal common substructure (MCSS) algorithm. These reference structures are obtained by using spectral search techniques on our NMR spectral database containing about 90 000 ^{13}C-NMR spectra, with the ^{13}C-NMR spectrum of the compound under investigation as the query spectrum. The molecular formula of the target structure can be further used to reduce the number of structures within the final hit list to about 100 so that they can be processed in a few minutes of CPU time on a graphics workstation. In order to improve the results, the structures are preprocessed by removing those carbon atoms which are outside a user-defined range (usually 2–4 ppm) of each query shift value. The investigations shows that the MCSSs thus obtained usually contain main structural features of the target compound and can be directly used as structural constraints in the structure generation process.

INTRODUCTION

The modern electronic computer applications in chemistry has brought about the development of "intelligent" software systems for aiding chemists in structure elucidation of organic compounds. These systems usually work in three steps. First, the system analyzes the experimental spectroscopic data of the compound under investigation to extract the information about present/absent structural fragments. Second, the structure generator of the system assembles the obtained structural fragments into all the possible structure isomers. Finally, spectral properties are automatically estimated and compared with the corresponding experimental data for each isomer, and the coincidence of the estimated and experimental data is then used as the measure to rank the structural candidates.

The structure generation is in principle a combinatorial problem. The efficiency of this process greatly depends upon the quantities and quality of the structural constraints obtained in the spectral interpretation step. If enough structural fragments have been deduced in the first step, the structure generator can generate only a few isomers. However, if the structural information from step 1 is limited, a lot of isomers can be generated, which will not only take a long time to compute but also make the selection of the correct structure more difficult. Thus the efficient of the third step is strongly affected by the spectral interpretation step.

Since the 1960s, quite a lot of effort has been made on the automatic interpretation of the spectra of organic compounds. The DENDRAL system[1] is an early example for assisting chemists in the structure determination of organic compounds from mass spectra. The PAIRS program[2] was designed to extract structural information from IR spectra. Owing to the good correlation between chemical shift values and the carbon atoms of organic compounds, ^{13}C-NMR spectroscopy plays a very important part in organic structure elucidation. A variety of computer programs have been developed to extract automatically the structural information from ^{13}C-NMR spectra. Most of these programs are based on substructure/shift correlation libraries. Although many correlation libraries designed for purposes such as chemical shift estimation usually consist of quite large substructures which can be as large as four,[3,4] even five,[5] sphere fragments, the libraries designed for extracting fragments which can be directly used as the structural constraints in the structure generation process consist of usually much smaller substructures. For example, in Bremser's method,[6] the one- and two-sphere fragment library is used in the spectral interpretation instead of the detailed four-sphere description of the correlation tables. The statistical analysis of HOSE-code-derived substructures allows a probability-based generation of constraints independent of the spectroscopic method. In Munk's system,[7] the one-concentrically-layered, atom-centered fragments which are deduced from ^{13}C-NMR spectra and other spectroscopic properties are used as the structural building units for its structure generation COCOA. It is apparent that in order to improve the efficiency of the structure generation process, larger structural constraints are indispensable.

In the current contributions, we describe a new approach for deducing automatically larger structural fragments based on our maximal common substructure (MCSS) algorithm.[8]

EXPERIMENTAL SECTION

The database accessed consists of about 90 000 ^{13}C-NMR spectra, including the libraries of the University of Vienna, SADTLER Research Laboratories, and the German Cancer Research Center at Heidelberg. The algorithm was implemented in FORTRAN 77 under the UNIX operating system on an IRIS-Indigo workstation. The program consists of some 3000 lines of source code, excluding the MCSS subroutines. The procedures of preparing the reference structures are as follows: First, a spectral similarity search is performed on the database with the ^{13}C-NMR spectrum of the compound under investigation as the query spectrum; then, the molecular formula range search is performed on the obtained hit list to reduce the number of entries to about 100. The average CPU time to handle 100 structures is about 20 min on an IRIS-Indigo workstation.

MAIN PROCEDURES

The method presented here is based on the following well-accepted concept: similar ^{13}C-NMR spectral features usually

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

0095-2338/94/1634-0934$04.50/0

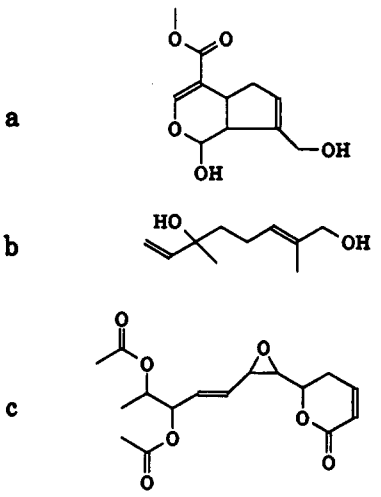


Figure 1. Structures of the three compounds under investigation together with their ^{13}C-NMR spectral data. (a) Genipin: 168.2/S, 152.7/D, 142.6/S, 129.6/D, 110.6/S, 96.3/D, 61.0/T, 51.1/Q, 47.8/D, 38.9/T, 36.4/D. (b) 1-Hydroxylinalool: 144.9/D, 134.9/S, 125.7/D, 111.8/T, 73.3/S, 68.3/T, 41.8/T, 27.6/Q, 22.4/T, 13.7/Q. (c) Spicigera lactone: 170.3/S, 169.9/S, 162.5/S, 144.4/D, 130.4/D, 127.3/D, 121.6/D, 74.2/D, 73.7/D, 70.2/D, 57.9/D, 56.0/D, 27.2/T, 21.1/Q, 21.0/Q, 15.2/Q.

imply similar structural features. Therefore, through the analysis of the reference structures corresponding to the most similar spectra retrieved from the database using the spectrum of the compound under investigation as the query spectrum, we can deduce some common structural fragments which most probably occur in the target structure. These fragments can be used as structural constraints during the structure generation process.

The procedures of the approach can be described as follows:

1. A spectral similarity search is performed on the reference collection using the ^{13}C-NMR spectrum of the target compound as query spectrum.

2. The number of entries retrieved is reduced to a reasonable scope (about 100) using molecular formula range search. Details of the ranges used for present elements are given in the figure captions of the examples described.

3. The reference structures obtained in step 2 are preprocessed by eliminating unreasonable atoms.

4. Each structure is compared to all other structures in the hit list using the MCSS algorithm.

5. The MCSSs generated from step 4 are collected and analyzed.

The detailed description of each step is given below.

The spectral similarity search method applied here is based on the SAHO (spectral appearance in hierarchical order) code.[9] The reasonable number of structures which can be handled in acceptable CPU time is about 100. The method we use to reduce the number of structures obtained by the SAHO technique is to perform a molecular formula range search on the above hit list. The range of the molecular formula is adjusted in such a way that a total number of about 100 entries is obtained.

However, it must be pointed out that the SAHO method uses a relatively high tolerance of the chemical shifts. Therefore, the common fragments deduced by direct comparison of the reference structures obtained by the SAHO approach are usually quite unreliable. A reasonable way to improve the result is to cut-off those carbon atoms from the reference structures whose shift values are quite different from any query shift values before they are handled by the MCSS program. However, one main problem is to define shift ranges; too small a tolerance of shift values will exclude too many carbon atoms, resulting in structural fragments that are too small, while too a large tolerance will generate too many unreasonable fragments. The investigation indicates that the query shift of ±1–2 ppm is a practical choice. Nevertheless, the program allows the user to define the threshold value of the shift ranges. The carbon atoms whose shift values are outside any of the user-defined shift ranges will be automatically removed from each reference structure. Besides, all heteroatoms which do not connect to the remaining carbon atoms are also removed.

Through the above manipulation, the structures become smaller and some of them may have been divided into two or more disconnected parts. Instead of comparison of two structures, each connected part of a structure is isolated and then compared with each connected part of all other structures if both parts have at least one shift value within the same user-defined shift range.

Besides the improvement of the reliability of the results, removing the unreasonable atoms from the reference structures also achieves a significant gain of CPU time, because it can greatly decrease the complexity of the structures to be processed by the MCSS algorithm. However, for flexibility, this preprocessing step of the structures is designed as an option. The program can directly analyze unpreprocessed structures, as requested. The generated MCSSs are sorted according to the number of occurrences; furthermore, the MCSS list can be compressed by eliminating smaller MCSSs which are substructures of other MCSSs.

RESULTS AND DISCUSSION

Genipin (Figure 1a) has been taken as the first example. A total number of 27 MCSSs were deduced with a shift deviation of 2.5 ppm used for preprocessing the reference structures retrieved from the database, as shown in Figure 2. The number of occurrences of each MCSS and the total number of comparisons of the reference structures handled by the MCSS algorithm are displayed at the lower-right corner of each square. The MCSSs were sorted according to the descending order of their occurrences. Among those 27 MCSSs, there are only two invalid MCSSs, which are not substructures of the target compound. The atoms of the invalid MCSSs which cannot be matched to those of the target structure have been marked with asterisks.

From Figure 2 it can be seen that two invalid MCSSs (no. 22 and no. 27) have very low occurrences: 2 and 1, respectively. From the statistical point of view, we can select the MCSSs with high occurrences as our solution. However, when we try to define a criterion to exclude the invalid MCSSs, some problems appear. The number of occurrences decreases dramatically; the sizes of the MCSSs increase with the decrease of the occurrences in a complex way. The MCSSs with high occurrences (>100) are too small to be useful to the problem—the largest one consisting of only four non-hydrogen atoms. Even the MCSSs having as low as 20 occurrences show the same situation. Most of the meaningful MCSSs have very low occurrences (<9). This makes it difficult to define a reliable criterion to distinguish the valid and invalid MCSSs. On the other hand, eliminating all the MCSSs with occurrences of less than 5 may exclude some very useful MCSSs, such as MCSS 24.

As has been mentioned previously, the reliability of the results depends also strongly upon the shift deviation used for preprocessing the reference structures. The results obtained with the different deviation values for genipin are summarized in Table 1.

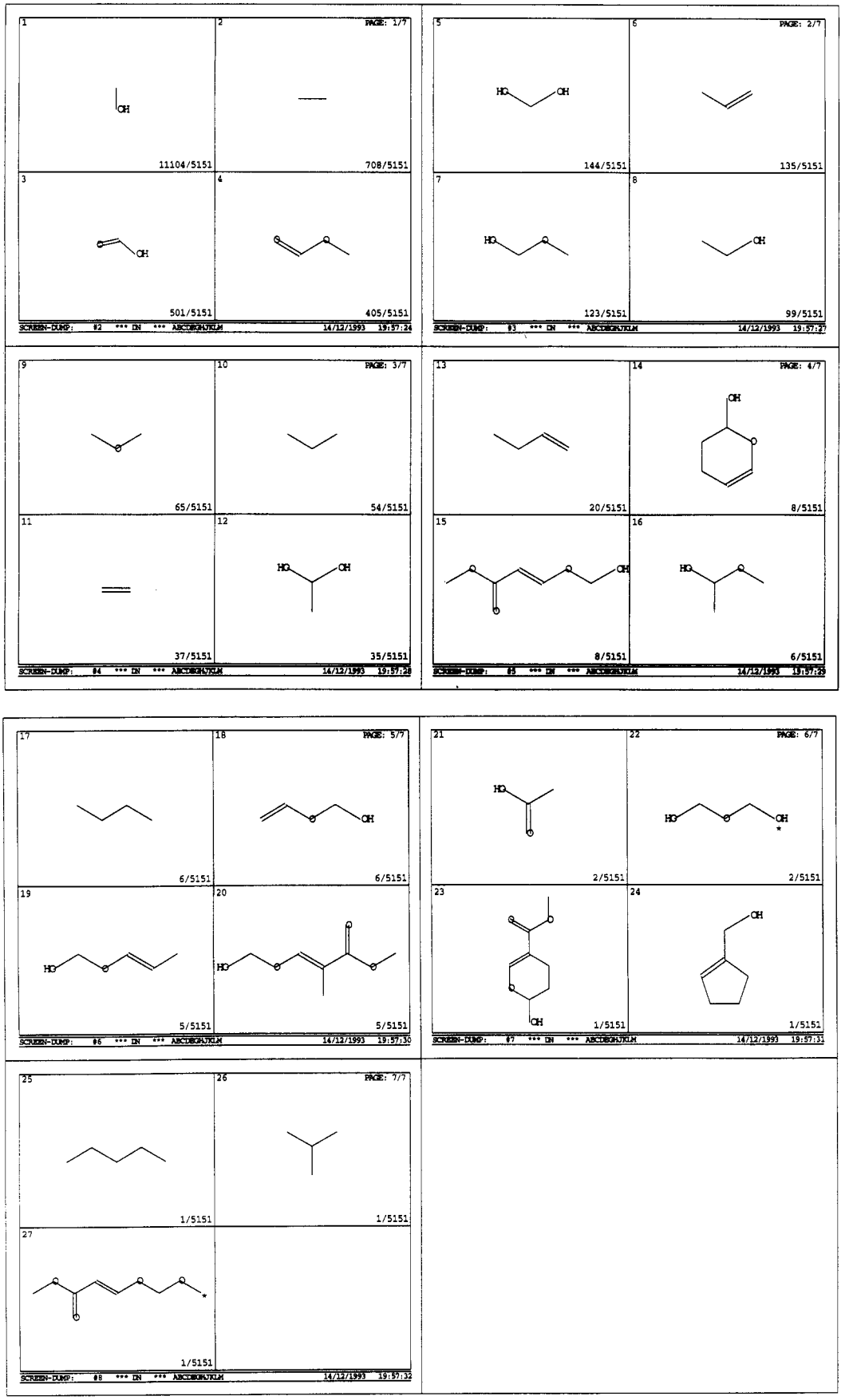


Figure 2. Total number of 27 MCSSs generated for genipin with a shift deviation of 2.5 ppm used for preprocessing the reference structures. Two of them are invalid MCSSs. The number of occurrences of each MCSS and the total number of comparisons of the reference structures using the MCSS program are displayed at the lower-right corner of each square. The MCSSs were sorted according to the descending order of their occurrences. The atoms of the invalid MCSSs which cannot be matched to those of genipin are marked with an asterisk.

From Table 1 it can be seen that, with the increase of the shift deviation, the number of invalid MCSSs increases faster than that of valid MCSSs, the number of occurrences of the first invalid MCSS in the sorted MCSS list also increases

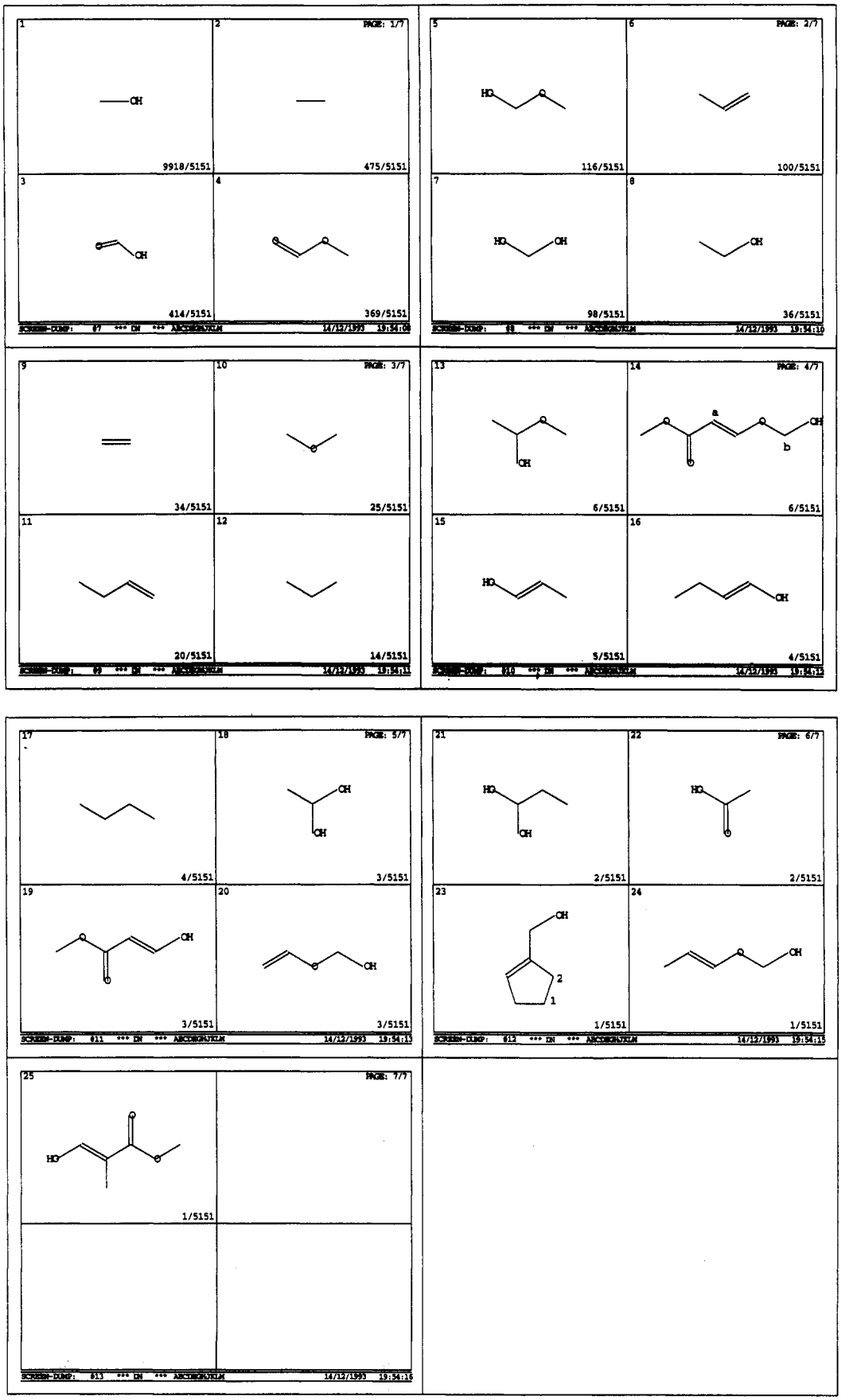


Figure 3. Total number of 25 MCSSs generated for genipin with a shift deviation of 2.0 ppm used for preprocessing the reference structures. All these MCSSs are valid. The MCSSs were sorted according to the descending order of their occurrences.

quickly. The CPU time increases with the increase of deviation values. It seems that the shifts deviation of 2.0 ppm could be used as a good criterion because the 25 MCSSs thus deduced are all valid ones; furthermore, some very useful MCSSs, such as MCSSs 14 and 23, have been obtained, as shown in Figure 3.

From a comparison of MCSSs 14 and 23 with the structure of genipin in Figure 1a, it can be seen that the complete

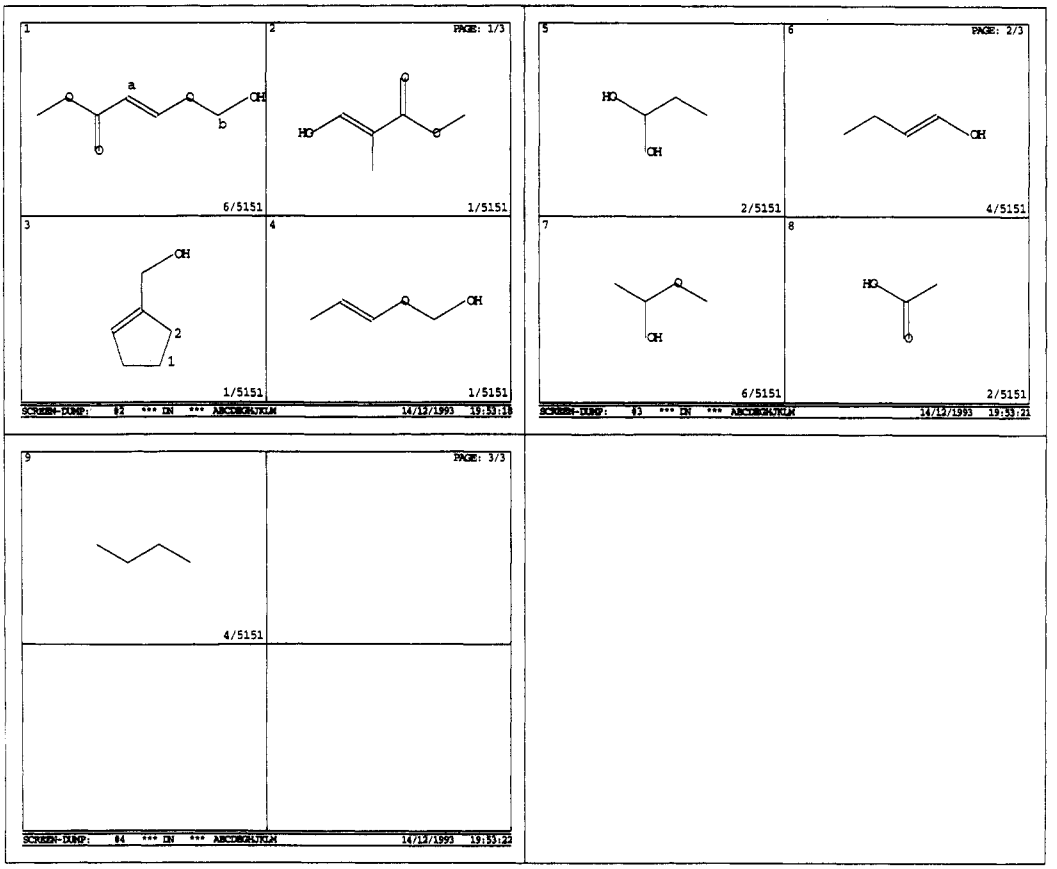


Figure 4. Compressed MCSS list generated from Figure 3. The MCSSs were sorted according to the descending order of their size (=the number of non-hydrogen atoms).

Table 1. Results for Genipin ($C_{11}H_{14}O_5$) Using Different Shift Deviations for Preprocessing Reference Structures[a]

deviation (ppm)	1.0	1.5	2.0	2.5	3.0	4.0	5.0	6.0	7.0
CPU time (s)	456	884	1220	1444	1806	1994	3431	5949	7420
MCSSs	7	16	25	27	39	58	72	80	110
number of total occurrences of MCSSs	2515	5380	11 660	13 483	17 189	21 394	40 684	65 027	78 364
Before Compression									
number of valid MCSSs	7	16	25	25	32	39	46	54	73
number of invalid MCSSs				2	7	19	26	26	37
position of the 1st invalid MCSS				22	25	19	20	22	32
occurrences of it				2	4	18	34	36	41
After Compression									
number of valid MCSSs	4	6	9	6	3	3	5	6	10
number of invalid MCSSs				2	4	3	7	8	13

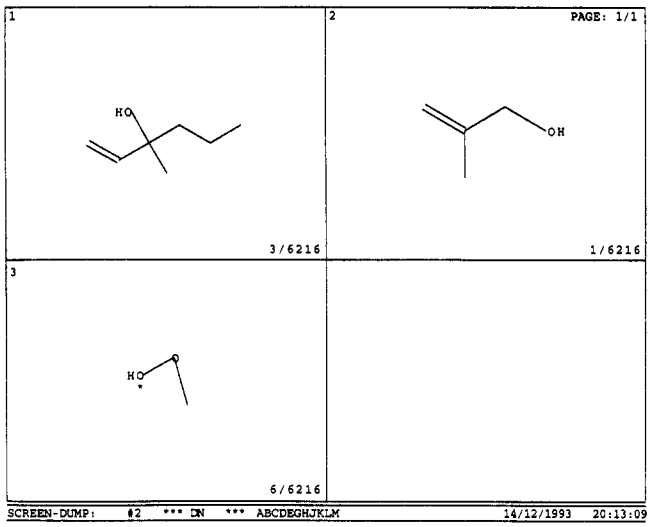
[a] A set of 102 structures was selected by performing a molecular formula range (C_5C_8–$C_{17}H_{36}O_{11}$) search on the hit list of a total of 708 entries which were retrieved from the database by means of a spectral similarity search.

structure of genipin can be assembled by simply connecting atoms a and b of MCSS 14 to atoms 1 and 2 of MCSS 23, respectively. The other MCSSs can be used as further constraints to support these two connections.

It is interesting to note that genipin was chosen by Lindley et al.[10] and Bremser et al.[6] to demonstrate the abilities of their structure generation programs GENOA and ACCESS, respectively. In the former case, besides the ^{13}C-NMR spectrum and the molecular formula of genipin, ^{1}H-NMR and 2D-NMR data had to be used to deduce enough structural fragments. In the latter case, the correct solution was obtained by using only the ^{13}C-NMR spectrum and the molecular formula. Our result further confirms that these two kinds of information are indeed enough to solve the genipin problem.

The reader may have noted that the MCSS lists shown in Figures 2 and 3 contain somewhat duplicate information, because many smaller MCSSs are merely the substructures of other MCSSs. Thus the MCSS list can be compressed by removing all those MCSSs which are substructures of the remaining ones without losing significant information. The 25 MCSSs shown in Figure 3, for instance, can be compressed into 9, as shown in Figure 4. This new MCSS list is sorted according to the number of non-hydrogen atoms within each MCSSs. In later discussions, the MCSS lists are referred to as compressed lists.

It must be pointed out that the current method to represent structures does not give explicitly the information about atom hybridization. During the comparison of structures, this information is calculated from other parameters of the original structures. Thus the hybridization information of some atoms of the MCSSs generated may have been lost if these atoms connect to other atoms with only single bonds in the MCSS. Because of this ambiguity, MCSSs 7 and 8 in Figure 4 are treated as valid MCSSs. MCSS 8 is in fact a substructure of MCSS 1.



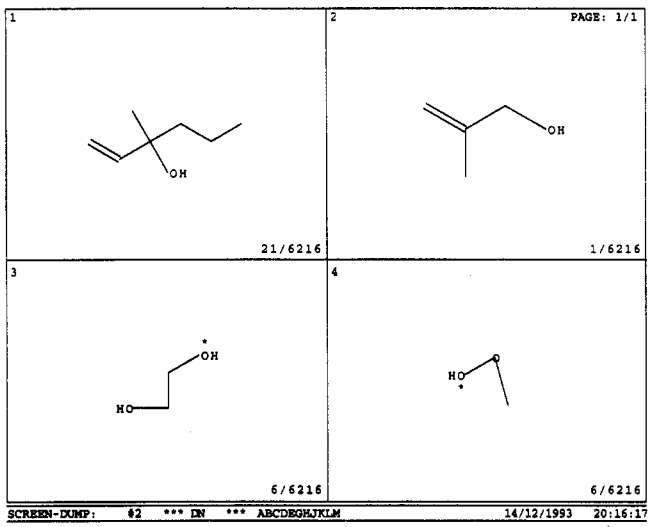


Figure 5. Compressed MCSS lists generated with different shift deviations for 1-hydroxylinalool ($C_{10}H_{18}O_2$). A set of 112 structures was selected by performing a molecular formula range ($C_{10}H_{14}O$–$C_{11}H_{19}O_3$) search on the hit list of a total of 1441 entries which were retrieved from the database by means of a spectral similarity search. The atoms of the invalid MCSSs which cannot be matched to those of 1-hydroxylinalool are marked by an asterisk. Three occurrences of 1-hydroxylinalool were found in the database and have been excluded from the final hit list. (a) Shift deviation: 1.0 ppm. CPU time: 261 s. (b) Shift deviation: 1.5 ppm. CPU time: 384 s.

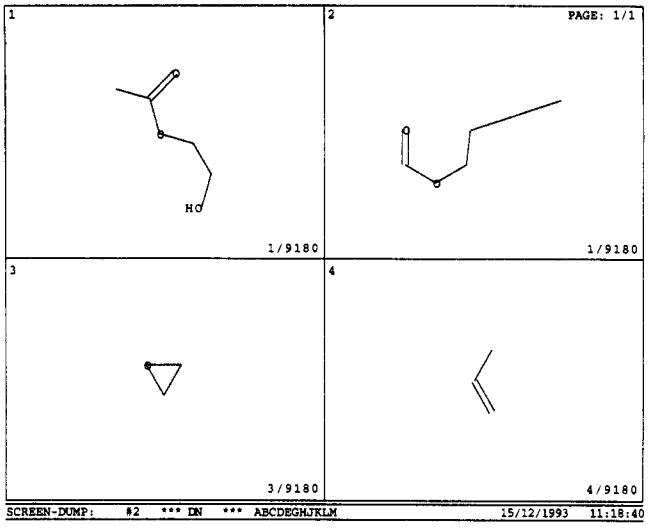
It should be noted that using 2.0 ppm or lower shift deviations cannot always lead to all valid MCSSs. As an example, consider 1-hydroxylinalool,[11] as shown in Figure 1b. The results are given in Figure 5.

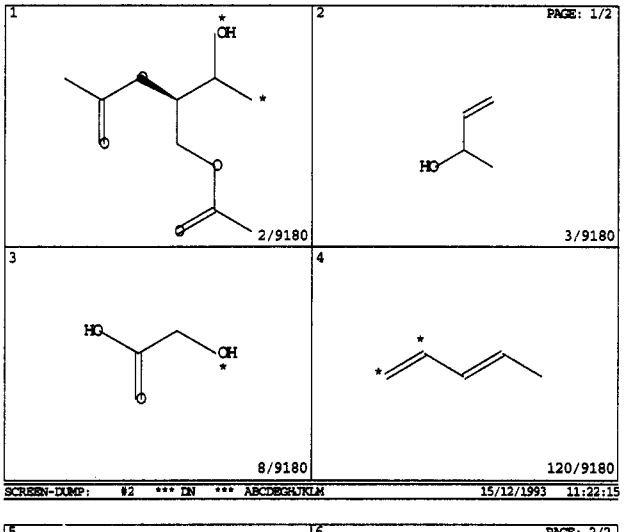
When deviations of 1.5 and 2.0 ppm were used, identical results were obtained, as displayed in Figure 5b, except that there are three occurrences at the 2 ppm level of MCSS 2 instead of one at the 1.5 ppm level.

Both compressed lists contain four MCSSs, but two of them are valid. With the deviation of 1.0 ppm, the similar result was obtained but with only one invalid MCSS. It can be seen from Figure 5a that the two valid MCSSs are so large that the complete target structure can be generated by simply connecting atom a of MCSS 1 to atom b of MCSS 2.

Since the MCSSs are planned to be used as alternative structural constraints in the structure generation process, the existence of some invalid MCSSs presents no big problem to such a structure generator which can prospectively eliminate most structures containing invalid fragments.

The above example gives us a hint that we should run the program beginning with a lower deviation, e.g., 1.0 ppm. The large MCSSs, such as MCSSs 1 and 2 in Figure 5a, lead to the final solution. Otherwise, the same procedure is repeated with 2.0 ppm deviation.





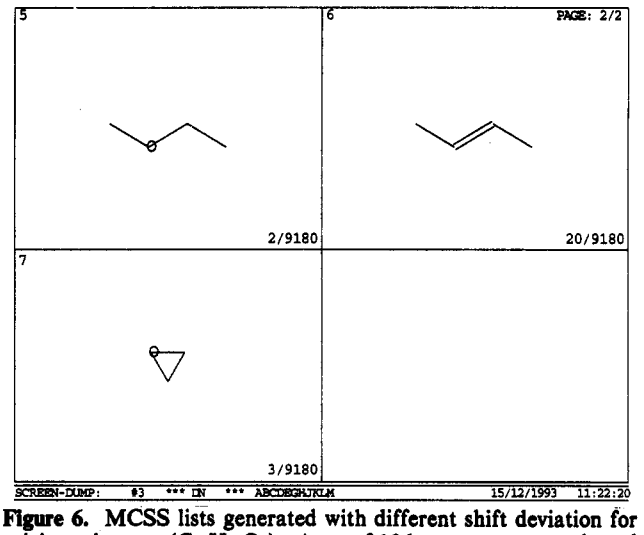


Figure 6. MCSS lists generated with different shift deviation for spicigera lactone ($C_{16}H_{20}O_7$). A set of 136 structures was selected by performing a molecular formula range ($C_{12}H_{10}O_2$–$C_{19}H_{45}O_{10}$) search on the hit list of a total of 1679 entries which were retrieved from the database by means of a spectral similarity search. The atoms of the invalid MCSSs which cannot be matched to those of spicigera lactone are marked with an asterisk. No identical structures of spicigera lactone were found in the database. (a, top four panels) Compressed MCSS list when using a shift deviation of 1.0 ppm. CPU time: 675 s. (b, bottom eight panels) Compressed MCSS list when using a shift deviation of 2.0 ppm. CPU time: 1925 s.

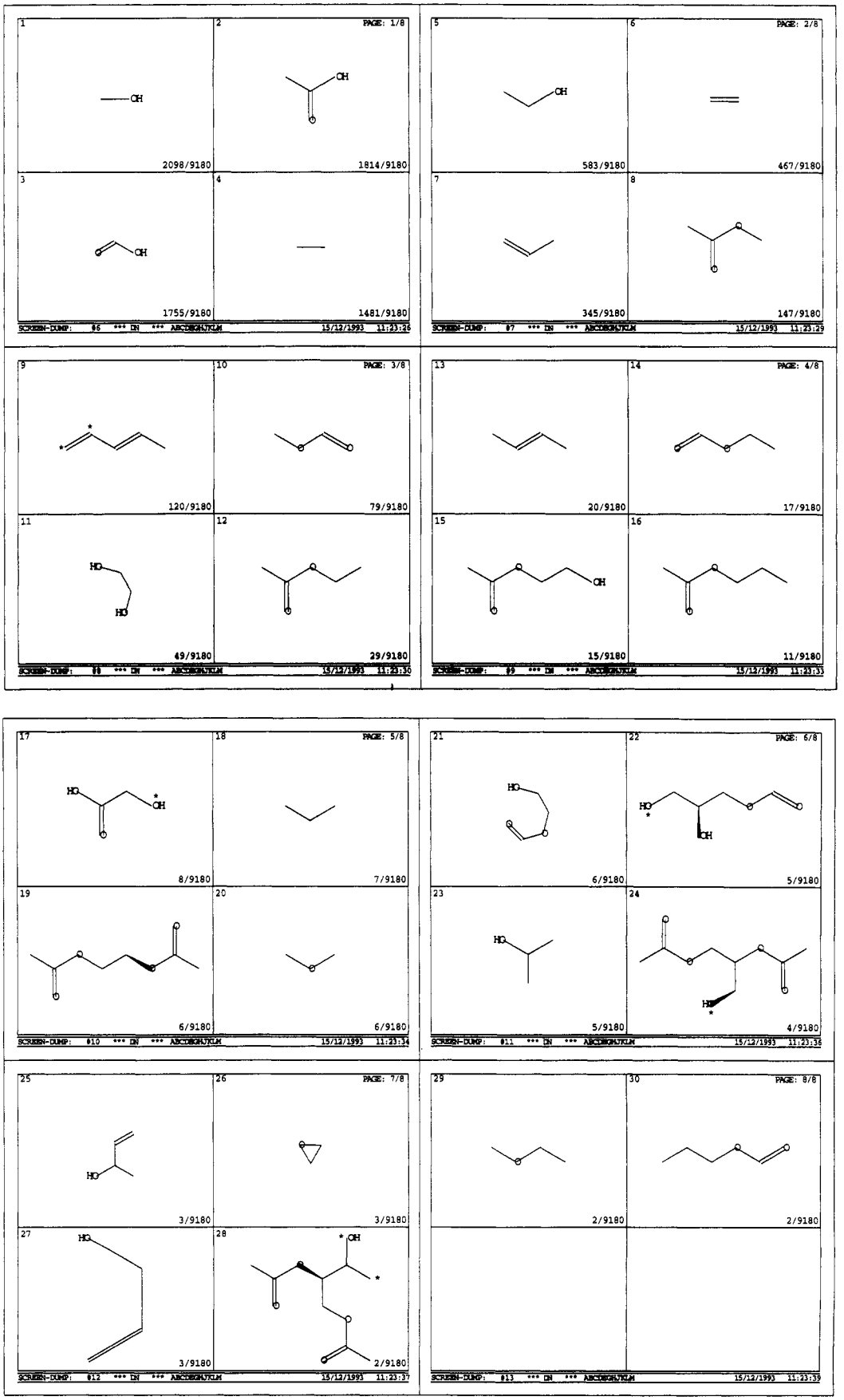


Figure 7. Original MCSS list when using a shift deviation of 2.0 ppm, generated for spicigera lactone ($C_{16}H_{20}O_7$). Conditions are the same as those given for Figure 6.

The third compound under investigations, spicigera lactone as shown in Figure 1c, is a recently reported new natural-product, isolated from the petroleum extract of the inflorescences.[12] The complete ^{13}C-NMR assignment was achieved using two-dimensional NMR experiments. The MCSSs deduced with different shift deviations are given in Figure 6 and Figure 7. A 1.0 ppm deviation leads to a compressed list containing four valid MCSSs, as shown in Figure 6a. It can

be easily seen that these MCSSs represent almost all the important structural features of the target structure. Therefore they can be used as the final solution. The compressed list generated with the shift deviation of 2.0 ppm is given in Figure 6b. This list should be especially noted. It contains three invalid MCSSs and four small valid MCSSs. The quite large valid MCSS (no. 1) in Figure 6a has now disappeared here. By inspection of the original MCSS list shown in Figure 7, we can find that some large valid MCSSs, such MCSSs 15 and 19, were in fact deduced, but they were then eliminated in the compressing process because they are substructures of MCSS 28.

From the above discussion, some conclusions can be drawn: (a) We should be satisfied with medium-sized (6–10 non-hydrogen atoms) MCSSs showing several informative structural features; trying to get very large MCSSs containing more than 10 non-hydrogen atoms by increasing shift deviations is dangerous. (b) The compressed MCSS list outlines the main structural features, while the original list offers detailed information. The MCSSs in both the compressed and the original MCSS lists should be checked before a final decision is made to use them. If the compressed list contains only a few very large MCSSs with very low occurrences, some already-deleted quite large MCSSs with higher occurrences should be taken back from the original list, such as MCSS 19 in Figure 7. (c) Using the occurrence criterion may exclude some very useful MCSSs, such as MCSS 26 in Figure 7 because of its low occurrence; it should be kept in mind that including more invalid MCSSs into the selected MCSS list is better than excluding very useful valid MCSSs from it.

CONCLUSION

The program described in this paper is a useful tool to deduce automatically relatively large structural fragments according to ^{13}C-NMR spectral information and optionally to the molecular formula of the unknown compound. These fragments can be directly used as structural constraints; optional user interaction allows the deletion of undesired fragments in order to accelerate the isomer generation process.

ACKNOWLEDGMENT

L.C. thanks the Austrian Academic Exchange Service for a research fellowship. We thank the staff of the University Computing Center for helpful discussions during program development. This project was supported within the European Academic Supercomputing Initiative (EASI).

REFERENCES AND NOTES

(1) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Wiley & Sons: New York, 1986.
(2) Woodruff, H. B.; Smith, G. M. Computer Program for the Analysis of Infrared Spectra. *Anal. Chem.* **1980**, *52*, 2321–2327.
(3) Bremser, W. HOSE-A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355–365.
(4) Gray, N.A.B.; Crandell, C. W.; Nourse, J. G.; Smith, D. H.; Dageforde, M. L.; Djerassi, C. Computer-Assisted Structural Interpretation of Carbon-13 Spectral Data. *J. Org. Chem.* **1981**, *46*, 703–715.
(5) Kalchhauser, H.; Robien, W. CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 103–108.
(6) Bremser, W.; Fachinger, W. Multidimensional Spectroscopy. *Magn. Reson. Chem.* **1985**, *23*, 1056–1071.
(7) Christie, B. D.; Munk, M. E. The Role of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Enhanced Structure Elucidation. *J. Am. Chem. Soc.* **1991**, *113*, 3750–3757.
(8) Chen, L.; Robien, W. MCSS: A New Algorithm for Perception of Maximal Common Substructures and Its Application to NMR Spectral Studies. 1. The Algorithm. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 501–506.
(9) Bremser, W.; Wagner, H.; Franke, B. Fast Searching for Identical ^{13}C NMR Spectra via Inverted Files. *Org. Magn. Reson.* **1981**, *15*, 178–187.
(10) Lindley, M. R.; Shoolery, J. N.; Smith, D. H.; Djerassi, C. Application of the Computer Program GENOA and Two-Dimensional NMR Spectroscopy to Structure Elucidation. *Org. Magn. Reson.* **1983**, *21*, 405–411.
(11) Carabedian, M.; Dagane, L.; Dubois, J.-E. Elucidation by Progressive Intersection of Ordered Substructures from Carbon-13 Nuclear Magnetic Resonance. *Anal. Chem.* **1988**, *60*, 2186–2192.
(12) Aycard, J.-P.; Kini, F.; Kam, B.; Gaydou, E. M.; Faure, R. Isolation and Identification of Spicigera Lactone: Complete ^{1}H and ^{13}C Assignments Using Two-Dimensional NMR Experiments. *J. Nat. Prod.* **1993**, *56*, 1171–1173.