MCSS: A New Algorithm for Perception of Maximal Common Substructures and Its Application to NMR Spectral Studies. 2. Applications

Lingran Chen† and Wolfgang Robien*

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

Received March 24, 1992

The automatic detection of Maximal Common SubStructures (MCSS) of two given chemical compounds is a central step in the computer handling of chemical structure information. The MCSS algorithm described previously is applied to the problem of spectrum comparison and calculation of chemical shift increments in ^{13}C-NMR spectroscopy. This methodology can be used for automatic error detection in large spectral databases; the implementation of this algorithm into the CSEARCH-NMR database system will be described in detail. The flexibility of the MCSS algorithm itself allows easy programming of very sophisticated tasks, giving a better understanding of substituent effects to the spectroscopist.

INTRODUCTION

The increasing availability of sophisticated NMR instrumentation has developed ^{13}C-NMR spectroscopy into a routine task, even with a few milligrams of a chemical compound. One important aspect of ^{13}C-NMR spectral data is the excellent correlation between structural features and corresponding spectral properties, which possesses enormous practical and theoretical significance. The interpretation of ^{13}C-NMR spectral data is usually based on direct comparison of the spectrum of the unknown with a large reference data collection of known structures and their well-assigned resonance lines. This task can be in principle performed manually, but it is usually very cumbersome and time-consuming. The fast development of computer hardware now allows the handling of fairly large databases on graphics workstations or even at the PC level. The CSEARCH-NMR database system[1] is an excellent tool for handling large spectral data collections in order to support the spectroscopist during the structure elucidation process. This computer program has been extended by the previously described MCSS algorithm[2] for perception of the Maximal Common SubStructures (MCSS) of two given structures. This method can be easily used for direct comparison of chemical shift values allowing automatic increment analysis and error detection within spectral databases. One central purpose of the CSEARCH software is to utilize the excellent graphics capabilities of modern RISC workstations, avoiding excessive output of tables holding numbers.

AUTOMATIC CALCULATION OF CHEMICAL SHIFT INCREMENTS

Detection of structural differences of two given chemical compounds is much more complicated than recognition of spectral differences. The former problem can be transformed into the identification of maximal common substructures, which also gives the information about atom-by-atom correspondences. By having this information available, it is now an easy task to calculate the spectral differences using well-assigned resonance lines. This display of two or more spectra as handled by existing software packages usually consists of a simple graphical representation of the spectral data. The example chosen in Figure 1 displays the schematic ^{13}C-NMR spectra of the four monosubstituted halogenbenzenes. The information the spectroscopist really needs is the change of chemical shift values caused by different substituents. This information can be easily derived by application of our MCSS algorithm to those four structures leading to the display at the bottom of Figure 1.

The methodology described here can be applied to more complicated examples, e.g., to the comparison of a hydroxylated triterpene with its peracetate—a technique which is frequently used in the field of natural product structure elucidation—showing the shift differences induced by acetylation. An example having the same tricyclic parent ring system[3] but bearing a different acid part is shown in Figure 2. The display consists of the two structures with the maximal common substructural fragment drawn by bold lines, the two ^{13}C-NMR spectra and the chemical shift increments marked within the spectrum display and inserted into the structural formula.

AUTOMATIC DETECTION OF DATABASE ERRORS

Since the early 1950s, quite a lot of effort has been made to convert printed spectral data collections into computer-readable ones, and several considerable large computerized databases have been established for IR, MS, UV, and NMR spectra. Some basic functions of the corresponding retrieval software for identification of unknown compounds are now routinely used, and many additional techniques have been implemented in order to support the spectroscopist during the structure elucidation process, making database systems more useful. On the other hand, some important problems[4] are connected with the use of computerized databases; the most

† On leave from the University of Science and Technology of China, Hefei, Anhui 230026, The People's Republic of China.

0095-2338/92/1632-0507$03.00/0

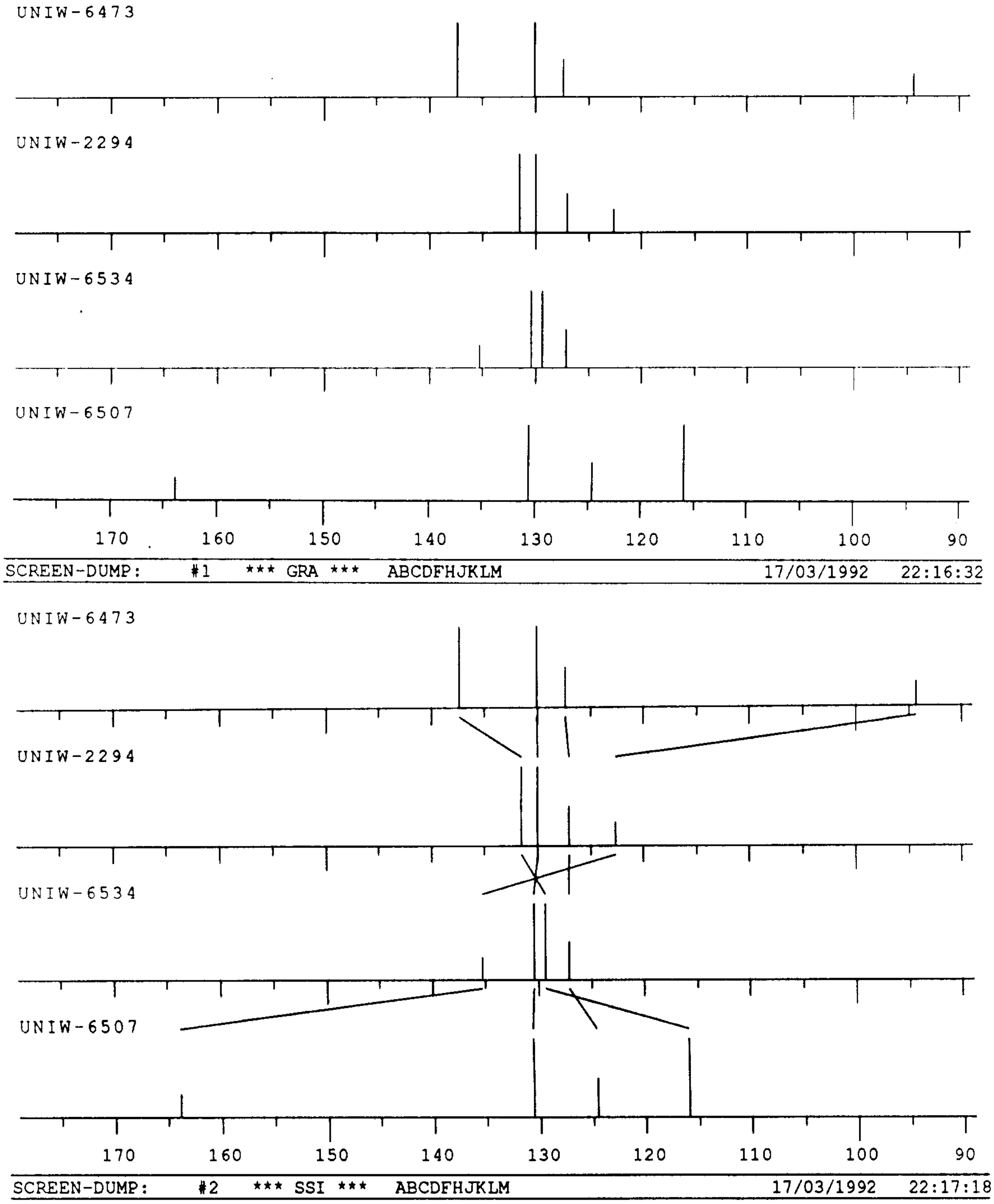

Figure 1. ^{13}C-NMR spectra of Phe-F, -Cl, -Br, and -I (from bottom to top). (Top) Schematic display of spectral data. (Bottom) Interpretation of the spectral data of these four compounds showing the trend of chemical shift variation caused by different substituents.

decisive problem seems to be the validation of the stored reference material. Gray[5] suggested that the development of methods for detecting and eliminating errors within the reference data set is a worthwhile area for research on file-based systems. From the literature, it can be deduced that not much effort has been focused on this topic.[4]

The features of automatic comparison and graphical display of structures and their corresponding spectral data allow the use of our approach described here as a valuable tool for automatic error detection within a reference data collection. The basic idea for using the MCSS algorithm is the consistency check of two identical structures, either during database update with new entries or checking established databases for internal consistency. The principle of this checking procedure can be best understood using the display of two structures and spectra as given in Figure 3. A vertical line between two resonances within the spectrum display corresponds to a small increment near 0 ppm. A larger increment causes a more diagonal line proportional to the size of the increment itself. Assume that the assignment of two resonance lines has been exchanged—e.g., because of an error during data input or a wrong assignment and/or a wrong structure—the two lines representing the increment values will generate a cross. Such a situation, caused not only by errors but also by stereochemical effects or solvent effects (especially during protonation), may induce a similar behavior, but this is at least a very good hint for some inconsistency within the two data sets under investigation. The final decision must be done in any case during the following **manual** verification procedure. It should be pointed out that this method is very time-consuming even when running on high-power graphics workstations, because many thousands of comparisons must be performed. The

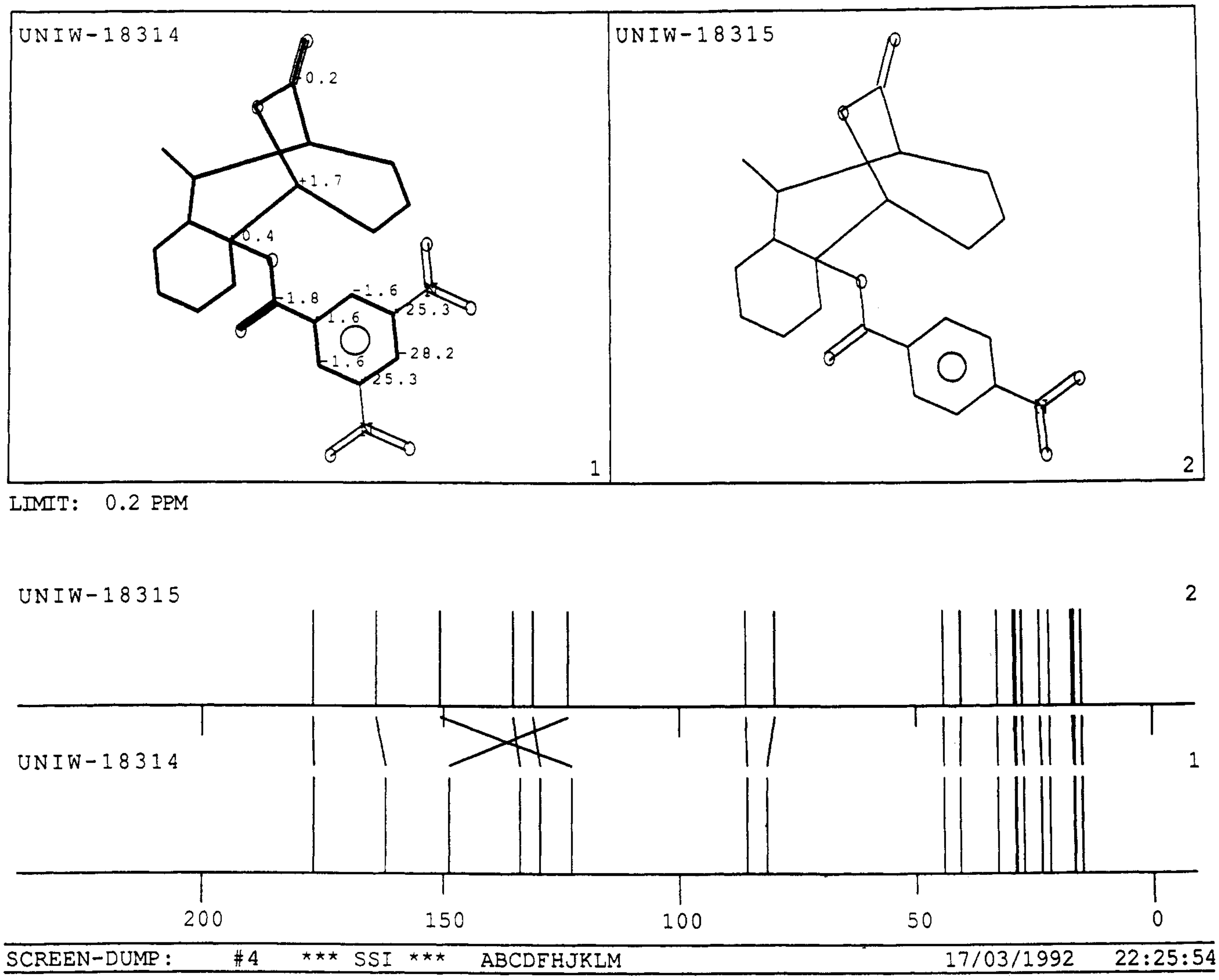

Figure 2. Comparison of a 4-nitrobenzoic acid with a 3,5-dinitrobenzoic acid derivative, showing the shift differences within the parent tricyclic ring system.

algorithm allows flexible selection of the range of entries to be checked, and also the user-defined threshold for increments, which are regarded as "errors", must be selected very carefully. It is strongly recommended to run such checking procedures during each database update in order to ensure high-quality information within spectral databases. A further fact should be clearly stated here: This procedure checks all pairs of identical structures with respect to their two-dimensional topology disregarding therefore stereochemical effects; furthermore, redundancy within the data is essential to apply this criterion. If two entries pass this test without any error message, it cannot be deduced that both of them are definitely correct; the only answer is that the assignment is done in a consistent, but maybe wrong, way.

An alternate method for error detection is based on spectrum prediction[6] and comparison of experimental and estimated chemical shift values using algorithms for automatic resonance line assignment.[1,7] This technique is not restricted to identical structure pairs—a fact inherent to the HOSE-code method—leading to much more reference material, especially at lower bond levels, giving sometimes large expectation ranges. The MCSS-based method here is designed for the exact match of two identical structures with respect to their two-dimensional topology, allowing a very detailed comparison, which is definitely based on redundant data within the database itself. It should be emphasized that both methods are powerful tools for automatic error detection and that both should be applied to existing database systems in order to achieve high-quality reference material for the spectroscopist.

EXPERIMENTAL SECTION

The algorithms described here have been implemented into the CSEARCH–NMR database system. The data collection of ^{13}C-NMR spectra consists of some 80 000 spectra, including the libraries of the University of Vienna, SADTLER Research Laboratories, and the German Cancer Research Center at Heidelberg. The programs have been written in FORTRAN-77 under UNIX operating system running on Silicon Graphics and IBM-RISC/6000 workstations.

CONCLUSION

The methodology described here uses the perception of maximum common substructural fragments for automatic calculation of chemical shift differences. The implementation of this algorithm into the CSEARCH–NMR database system has been used to understand and visualize shift increments during the structure elucidation process. A very decisive application of this approach is the automatic error detection

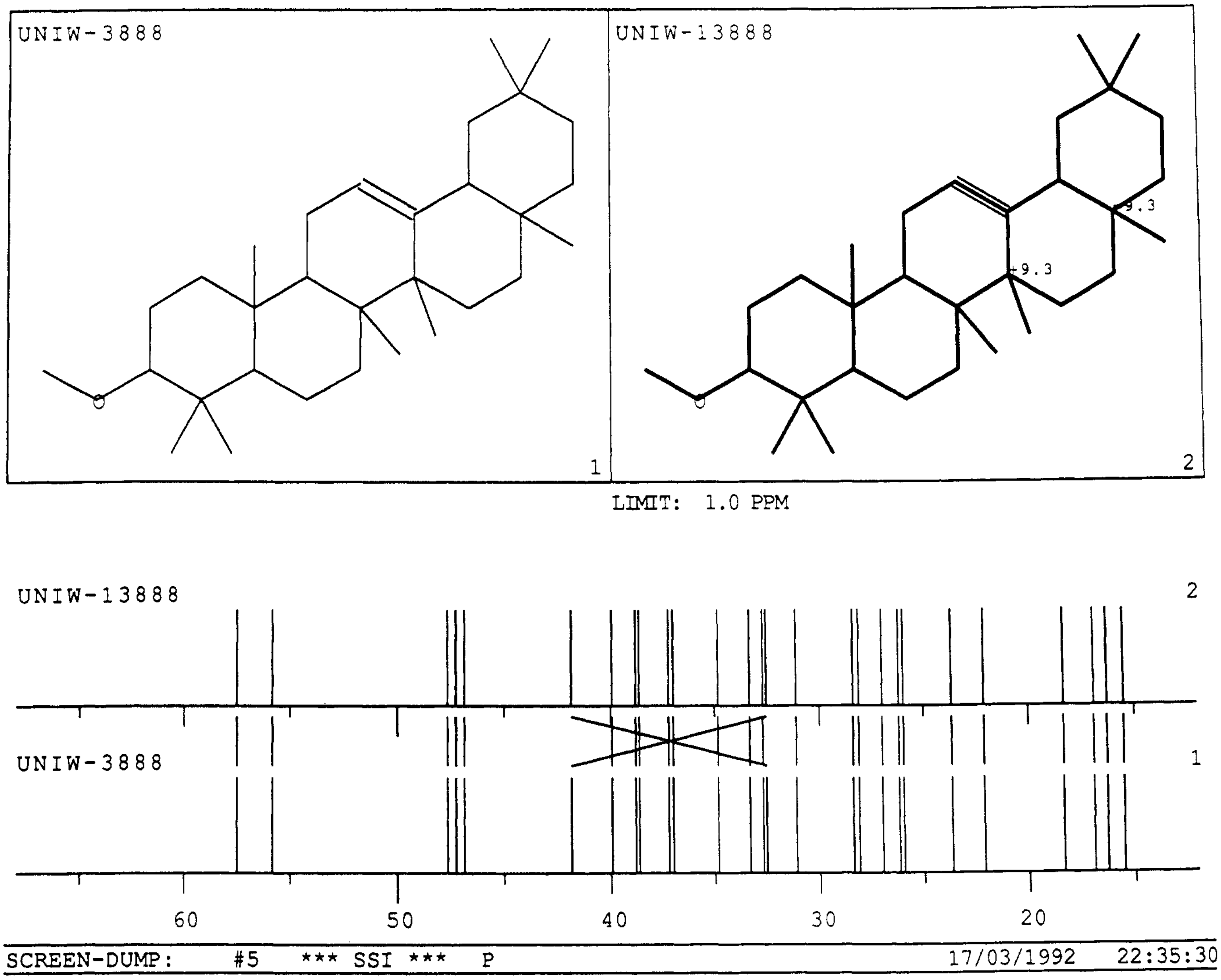

Figure 3. Comparison of two triterpenes showing an (artificially introduced) assignment error.

within a data collection of some 80 000 NMR spectra. Furthermore, the algorithm can be used to aid students in learning and understanding the influence of different functional groups on ^{13}C chemical shift values. The MCSS algorithm described here is a universal approach for comparison of structural and spectral features.

ACKNOWLEDGMENT

L.C. thanks the Austrian Academic Exchange Service for financial support. 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) 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.

(2) Chen, L.; Robien, W., preceding paper in this issue.

(3) Buchbauer, G.; Robien, W.; Sova, A.; Senger, J.; Amesberger, B.; Gerstmayr, G. Synthesis and Reactions of Tricyclotetradecane and Bicyclododecane Derivatives. *Arch. Pharm. (Weinheim)* **1990**, *323*, 127–131.

(4) Zupan, J., Ed. In *Computer-Supported Spectroscopic Databases*; Ellis Horwood Limited: New York, 1986.

(5) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Wiley & Sons: New York, 1986.

(6) Bremser, W. HOSE—A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355–365.

(7) Robien, W. *Computerunterstützte Zuordnung von ^{13}C-NMR Spektren. Mh. Chem.* **1983**, *114*, 365–372.