Enhanced Structure Elucidation†

Reinhard Neudert* and Michael Penk

Chemical Concepts GmbH, D-69442 Weinheim, Germany

Received August 25, 1995⊗

Conventional approaches to the structure elucidation of organic compounds are based on the use of spectroscopic data from different sources. The spectroscopist's task is to interpret the spectra and to derive structure proposals. The efficiency of this process depends mainly on his or her knowledge of structure-spectrum correlations, acquired in the course of everyday work. The most time-consuming process, that of assembling structures using the substructure information extracted from the spectra can be performed by computers. The "brain" of the computer is a structure oriented spectroscopic database and the knowledge derived from it. The structure building part is an isomer generator which accepts substructure inferences from the spectroscopist or from interpretation software. A final checking of the generated candidates is performed using spectrum prediction tools. The modules described are part of SpecInfo 3, a new software system for enhanced structure elucidation.

INTRODUCTION

Tens of thousands of new compounds have to be synthesized or extracted from natural sources in order to discover a potential drug. Despite the use of rational drug design techniques, economic success mainly depends on the number of new candidates available for activity tests. Consequently, research groups have begun to introduce new automation procedures such as synthesis robots and combinatorial chemistry. These enhancements on the production side are only one part of the whole task: Additional efforts in the subsequent structure elucidation process are also vital in order to avoid bottlenecks.

Many attempts have been made in recent years to make the structure elucidation process faster and more reliable by using computers. Structure-oriented spectral databases[1−4] can be used very successfully if the quality of the data is high enough. The **first level** of computer supported structure elucidation is library searches on spectra and structures, followed by statistical treatment of the resulting hit lists.[5] Searches for names, formulas, molecular weights, and other parameters are also included here.

If the correlation between spectra and structures in a database is used to predict spectral or structural properties of compounds not contained in the database, we can talk about a **second level** of structure elucidation. The correlation of spectral features with substructures results in powerful prediction tools,[6] especially in ^{13}C-NMR spectroscopy. An increment-based system is the ^{1}H-NMR estimation tool[7] included in SpecTool.[8] The PC-software group I*SEE[9] contains ^{13}C-NMR and ^{1}H-NMR prediction modules using structure oriented databases. Spectrum prediction software has also been developed for mass spectrometry[10] and for infrared spectroscopy.[11,12]

The **third level** of structure elucidation was initiated many years ago[13−15] and includes structure generators and all the modules necessary to use them efficiently. In the years since, some products have become commercially available.[16,17] The most serious problems in the past, those preventing broad application of such programs, were the lack of computing power, storage devices, and networking capabilities in spectroscopy laboratories.

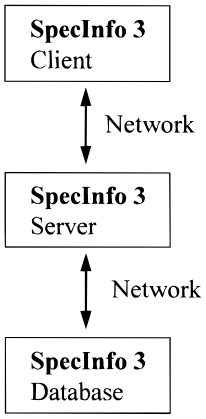

Figure 1. Client/server architecture of SpecInfo 3.

SpecInfo 3[18] is a high-performance system combining software and a database of NMR, IR, and MS spectra with associated structures. It has been completely renewed in the past two years and runs now under the easy-to-use X Window graphical user interface. Mainly depending on the structure oriented database the software contains tools from all the levels of structure elucidation as pointed out above. In the following some features are discussed with emphasis on level two and level three modules.

Many fruitful discussions with Prof. Sasaki and his coworkers and the exchange of programs and ideas influenced the strategy we chose to implement level three modules in SpecInfo.

SPECINFO 3: SOFTWARE ARCHITECTURE

Modern software is constructed on a modular basis, written in a standard programming language, and designed to be portable to other platforms. The new version 3 of SpecInfo has been completely redesigned and now includes X Windows as the graphical user interface and uses a client/server architecture (Figure 1).

The whole system can be divided into three parts:

−the client as user interface with all the input and output features

† Dedicated to Prof. Sasaki on the occasion of his 70th birthday.

⊗ Abstract published in *Advance ACS Abstracts,* February 1, 1996.

0095-2338/96/1636-0244$12.00/0

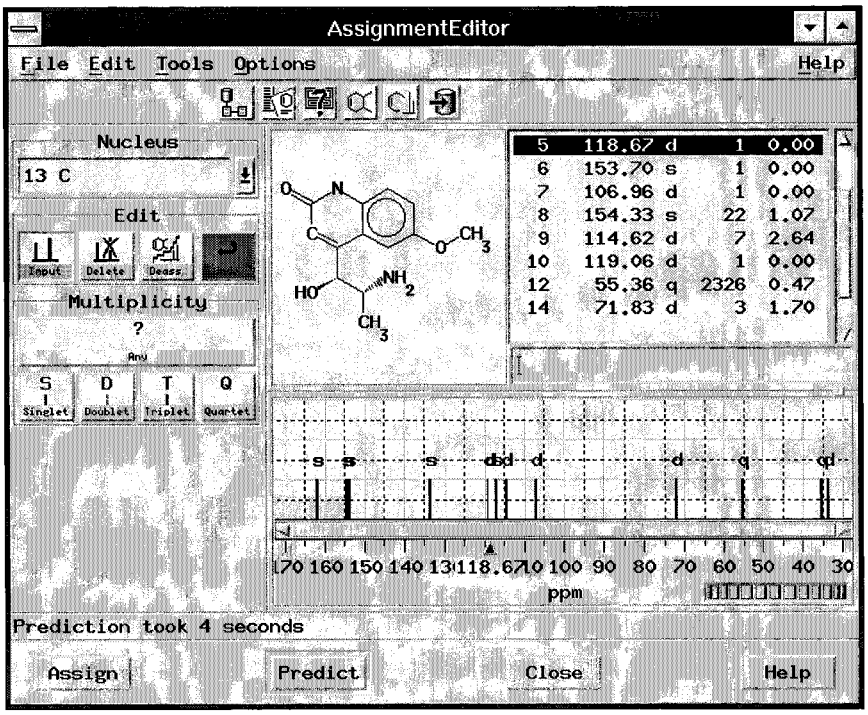

Figure 2. The Assignment Editor allows automatic assignment of experimental chemical shifts to a given structure and the prediction of shifts if no spectrum is available.

−the server, containing the programs for searches, statistics, comparison, spectrum prediction, etc.

−the SYBASE relational database management system, which guarantees a high level of data security

Network protocols are defined between the three parts. In principle, the user interface could run on a PC using an X Window emulator, the server could be installed on a SUN computer, and the database would be on an SGI computer. Available server resources can thus be used in a flexible way. The computers can be in different physical locations, combined by a network, without causing much network traffic, since no graphical information needs to be exchanged. An important advantage of such an open architecture is easy interfacing to other software products supporting the same architecture.

SPECTRUM PREDICTION

Chemical shifts may be generated in a number of different ways. Table 1 summarizes the terms which should be used to avoid misunderstandings.

Table 1. Several Strategies To Generate NMR Chemical Shifts

chemical shifts derived from...	...we call...	computational effort
increment systems (SpecTool etc.)	"estimated"	very small
topology codes applied on experimental data (SpecInfo, WIN-SpecEdit etc.)	"predicted"	small
spin system simulation (LAOCOON, WIN-DAISY etc.)	"simulated"	affordable
quantum mechanical treatment of molecular geometry, "shifts from Ψ" (NMR-CINDO, HyperNMR, etc)	"calculated"	enormous

The principle idea that enables prediction of NMR spectra is an atom-centered topological structure code describing the chemical environment of a given atom.[6] These codes characterize the environment of each individual atom in spheres around the specific center of interest. Obviously, this kind of coding is a very suitable spectrum-structure projection for methods with atom-centered spectral features, such as ^{13}C-NMR. It has been shown that statistical treatment of such codes in mass spectrometry also gives reasonable results.[19]

A detailed description of the limitations of ^{13}C-NMR spectrum prediction has been published.[20] New and unknown environments, data gaps, structure coding conventions, and the length of the substructure code are the most important sources influencing the quality of the results.

Clearly, such powerful tools can be used not only to verify new structures but also to accelerate the time consuming assignment procedure as part of data documentation. As an example, suppose we wanted to introduce an experimental ^{13}C-NMR spectrum into SpecInfo 3. The structure is either drawn using the built-in structure editor or imported as an MDL molfile. Both spectrum and structure are visible in the assignment editor (Figure 2). The command "Assign" predicts a ^{13}C-NMR spectrum from the structure internally and assigns the chemical shifts to the shifts in the experimental spectrum, taking into account the multiplicities of the experimental spectrum and those required by the structure. It is important to note at this point that automatic assignment is only an aid to the spectroscopist, it does not replace his expertise. Leaving this task to the computer alone would lead to a serious loss of quality in the database.

A further application of ^{13}C-NMR prediction is in quality assurance. When a new dataset is introduced into the database, an automatic verification can be applied as an input filter. Predicted and experimental spectra are compared, and chemical shifts are marked as possibly erroneous if the difference between the predicted and experimental values exceeds a predefined value. Finding an appropriate filter value is a difficult task because this determines the number of rejected datasets. A value between 10 and 20 ppm has been found to be generally useful.

As demonstrated in the section "structure generators", spectrum prediction is also a useful tool for the reduction of solution sets produced by such programs.

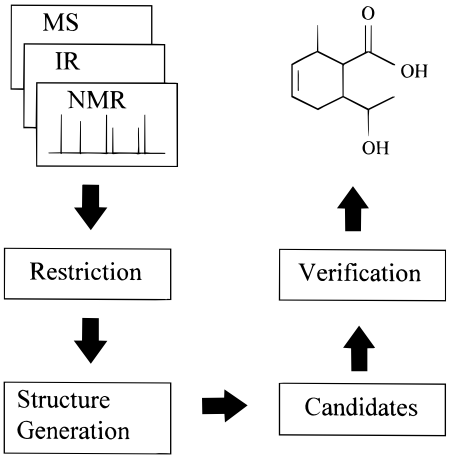

Figure 3. General scheme of automated structure generation: the restriction module extracts structural features from given spectra, the structure generator assembles candidates, and the verification tool checks the candidates for their relevance.

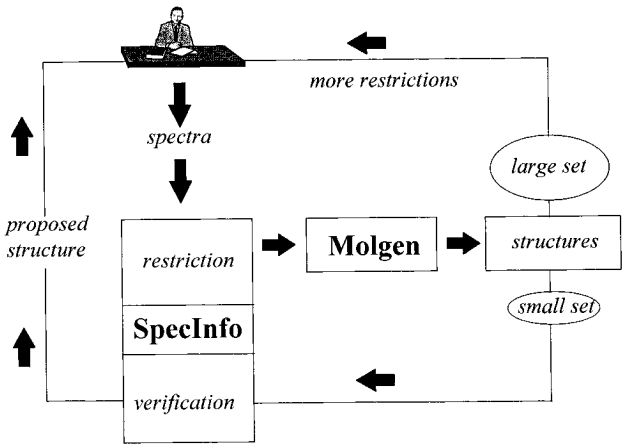

Figure 4. Interactive and combined application of SpecInfo and Molgen.

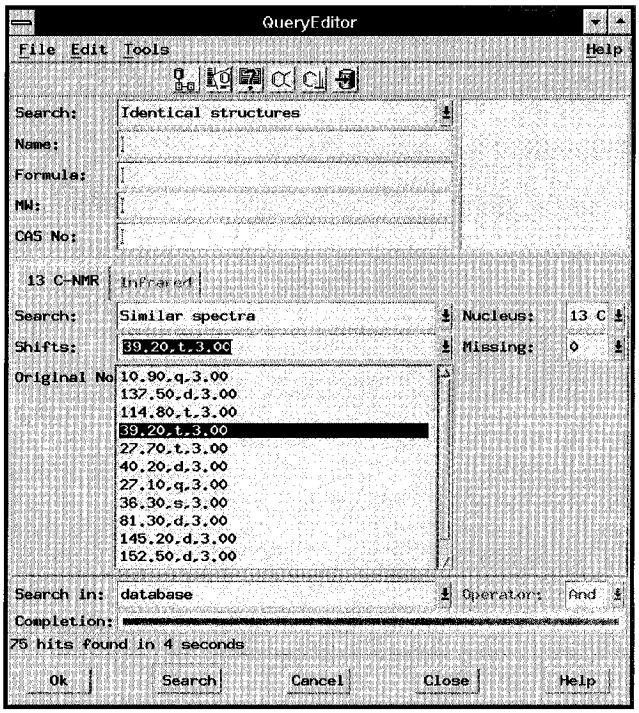

Figure 5. Experimental ^{13}C-NMR spectrum, imported as a JCAMP file from the spectrometer. The last figure in each shift (3.00) is a default value for the allowed deviation between the shifts of query and library spectra.

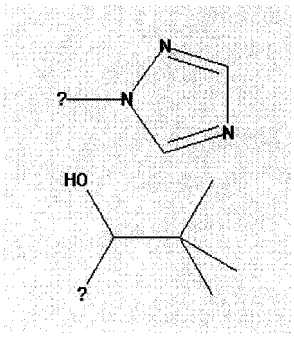

Figure 6. Two significant substructures as a result of a spectrum similarity search

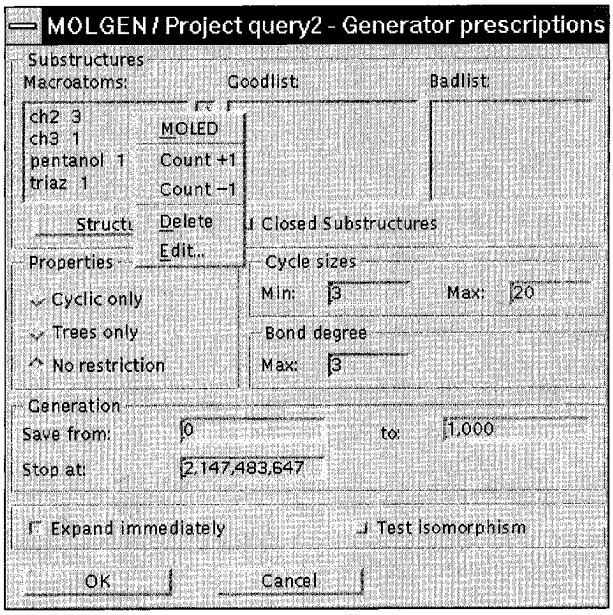

Figure 7. Molgen input window: the triazyl substructure together with 1 CH_3, 3 CH_2, and 3 CH groups are visible in the macrocomponent field. Some other restrictions such as cycle size and maximum bond degree have also been defined. The macrocomponent input expects nonoverlapping substructures, whereas the good list allows overlapping of substructures.

SPECINFO/MOLGEN AS INTERACTIVE STRUCTURE ELUCIDATION TOOL

The use of isomer generators for structure elucidation relies on the basic principles illustrated in Figure 3. The *restriction module* extracts substructure information from the spectral data. Chemics[16] and SESAMI[21] uses not only 1D-NMR spectra but also the 2D-NMR information from common experiments. The novel development SpecSolv,[22] which is an additional part of SpecInfo 3, uses only ^{13}C-NMR spectra as input. All the required restriction and verification steps are completely automated. The interactive system SpecInfo 3 in combination with Molgen[23] does not yet contain restriction modules of that kind. The spectroscopist himself has to introduce restricting criteria.

Two sources can be used to generate structural restrictions: SpecInfo 3 yields substructure information from spectrum similarity searches, and knowledge of the chemical reaction which leads to the unknown structure provides extra information. Parts of the molecule expected not to be involved in the reaction can be defined as "macrostructures".

Whereas the input of restricting substructures is not obligatory, the molecular formula is required. The molecular formula has to be determined from experimental data; elementary analysis and high resolution mass spectrometry

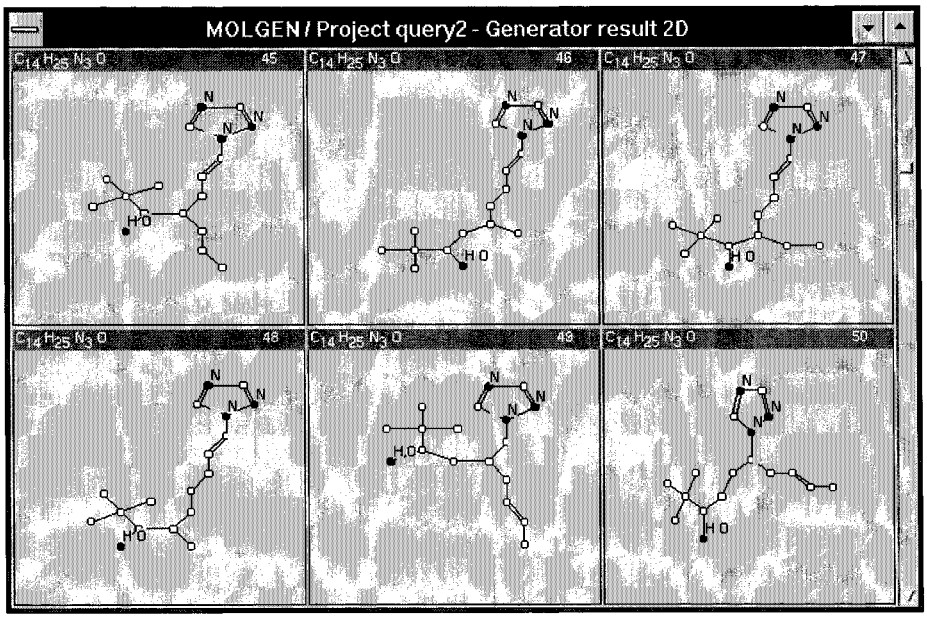

Figure 8. Six of the 239 generated structure proposals. After browsing through the list, the expert can further reduce the solution set by introducing other plausible restrictions.

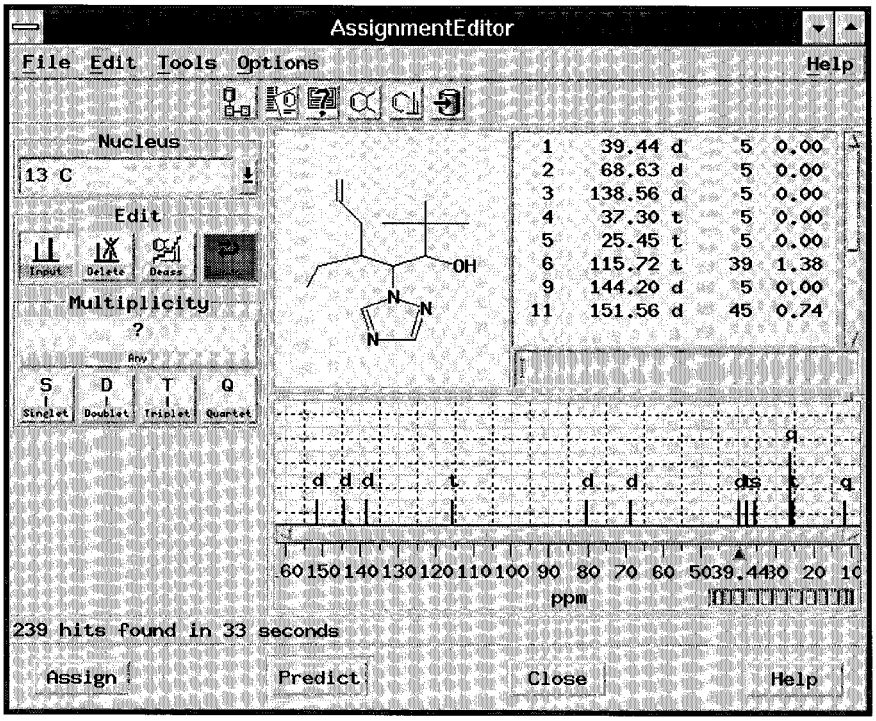

Figure 9. The structure candidate with the best match to the experimental ^{13}C-NMR spectrum is shown. In practice, not only the first candidate but also the next few have to be taken into account.

are appropriate techniques. As demonstrated in the next section, the ^{13}C-NMR spectrum can also be used to calculate the molecular composition.

The structure generator Molgen offers three options for using substructures as restriction criterion:

macrostructure accepts substructures which must be part of the target structure and do not overlap

goodlist contains (possibly) overlapping substructures

badlist defines substructures which cannot be present in the target structure

Certain other restrictions such as ring size can also be defined.

The *verification module* presently depends on ^{13}C-NMR spectrum prediction. Other software[18] uses ^{1}H-NMR estimation at this stage. If a solution set acceptable in size (<5000 structures) is achieved, the ^{13}C-NMR spectra of all the structures are predicted and compared with the experimental spectrum. The best matches are possible candidates.

A partly automated solution for the structure elucidation process can be realized using SpecInfo and Molgen together.

Figure 4 gives an overview of how SpecInfo/Molgen can be used in practice. It clearly reflects the basic idea of an interaction during the structure elucidation process. The spectroscopist does the restriction work at the input of Molgen and is responsible for using the information with due care. An inappropriate input may remove the target structure from the solution set. Structure generator strategies with software restriction tools are very careful not to

eliminate important information at this point.

APPLICATION EXAMPLE

Experimental conditions for the example below are as follows:

Number of ^{13}C-NMR spectra used for predictions: 99 100
Hardware: SUN SPARC classic running Solaris 2.3
Software: SpecInfo 3 in combination with Molgen

We start from a ^{13}C-NMR spectrum including multiplicities and a molecular mass of 251 amu. The multiplicities are 4 CH_3, 3 CH_2, and 6 CH groups. The molecular formula calculation software in SpecTool yields just one possible composition if only C, H, O and N are allowed: C14 H25 N3 O1.

The peaklist of the ^{13}C-NMR spectrum together with multiplicities is transferred to SpecInfo.

The spectrum is loaded into the query editor (Figure 5), and a similarity search in the database is performed. The resulting hit list of similar spectra and corresponding structures is carefully investigated for common structural features. In the example given, two substructures both occur six times in the best ten hits; therefore, these were tentatively assigned to be present (Figure 6).

Next, Molgen is activated, and the molecular formula is entered together with the above substructures (Figure 7). We expect them to be nonoverlapping, and they are therefore stored in the "macroatoms" field. Additional restrictions are obtainable from the ^{13}C-NMR multiplicities: If the substructures from the library search are taken into account, 1 CH_3, 3 CH_2, and 3 CH groups remain.

The generation procedure is started and results in a set of 239 isomers as possible candidates. Six of them are illustrated in Figure 8.

To identify the best candidates, a spectrum prediction and subsequent comparison with the experimental spectrum is carried out. The best candidate appears in first place in the hit list with, in this example, a match factor of 0.7. A match factor of 1.0 represents identity; with a match factor of 0.6, the next hit is not clearly separated and should also be considered as a possible candidate. Additional spectroscopic techniques may help to distinguish between them.

By far the majority of the candidates have match values of about 0.2 and can clearly be rejected.

FUTURE DEVELOPMENTS

Tools of the kind described can help spectroscopists and chemists in their daily work not only in working with pure structure elucidation but also by stimulating creativity. To reduce the tremendous amount of information produced by structure generators, more restriction and verification tools are desirable. ^{1}H-NMR estimation is available but not presently integrated into the SpecInfo 3 system. First results from IR prediction, based on neural networks, seem to indicate its suitability for the task. In the future, software with the necessary performance could be the basis for structure elucidation on the **fourth level**: software making decisions about the kind of experiments necessary to solve a given problem.

REFERENCES AND NOTES

(1) Warr, W. Computer-Assisted Structure Elucidation. *Anal. Chem.* **1993**, *65*, 1045A−1050A. Warr, W. Computer-Assisted Structure Elucidation. *Anal. Chem.* **1993,** *65*, 1087A−1095A.
(2) Bremser, W.; Grzonka M., SpecInfo−A Multidimensional Spectroscopic Interpretation System. *Mikrochim. Acta II* **1991**, 483−91.
(3) 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.
(4) v. d. Lieth, W.; Seil, J.; I. Köhler, I.; Opferkuch, H. J. CNMR Databank Techniques as Analytical Tools *Magn. Res. Chem.* **1985**, *23*, 1048−57.
(5) Bremser, W.; Fachinger, W. Multidimensional Spectroscopy. *Magn. Res. Chem.* **1985**, *23*, 1056−1071.
(6) Bremser, W. HOSE−A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(7) Bürgin Schaller, R.; Pretsch, E. A Computer Program for the Automatic Estimation of 1H NMR Chemical Shifts *Anal. Chim. Acta* **1994**, *290*, 295.
(8) SpecTool; Chemical Concepts GmbH: D-69442 Weinheim, Germany.
(9) I*SEE, Integrated Structure Elucidation Environment; Chemical Concepts GmbH, D-69442 Weinheim, Germany.
(10) Gasteiger, J.; Hanebeck, W.; Schulz, K. P. Prediction of Mass Spectra from Structural Information. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 264−271.
(11) Otto, M.; Gasteiger, J.; Zupan, J.; Herges, R. *Project Interpretation of Infrared Spectra*; 1992−1995, supported by the German Federal Ministry of Research.
(12) Affolter, C.; Clerc, J. T. Prediction of Infrared Spectra from Chemical Structures of Organic Compounds Using Neural Networks. *Chemom. Intell. Lab. Syst.* **1993**, *21*, 151−157.
(13) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Wiley & Sons: New York, 1986.
(14) Funatsu, K.; Miyabayashi, N.; Sasaki, S. I. J. Further Development of Structure Generation in the Automated Structure Elucidation System CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18−28.
(15) Christie, B. D.; Munk, M. E. *J. Chem. Inf. Comput. Sci*. **1988**, *28*, 87.
(16) CHEMICS; Toyohashi University, Japan (see ref 14).
(17) X-Pert; Bruker Company, Germany, 1994.
(18) SpecInfo; Chemical Concepts GmbH, D-69442 Weinheim, Germany.
(19) Neudert, R.; Bremser, W.; Wagner, H. Multidimensional Computer Evaluation of Mass Spectra. *Org. Mass Spectrom.* **1987**, *22*, 321−29.
(20) Grzonka, M. On the Understanding of the Results of NMR Spectra Prediction Using Spectroscopic Databases. *Mikrochim. Acta* **1994**, *116*, 111−122.
(21) 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.
(22) Will, M.; Richert, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221−227.
(23) Grund, R.; Kerber, A.; Laue, R. MOLGEN, a computer algebra system for construction of molecular graphs. *MATCH C.* **1992**, *27*, 87−131.

CI9500997