Fully Automated Structure Elucidation—A Spectroscopist's Dream Comes True[†]

Martin Will, Winfried Fachinger, and Joachim R. Richert*

BASF Aktiengesellschaft, Main Laboratory, ZHV/S-B9, D-67056 Ludwigshafen, Germany

Received August 10, 1995[⊗]

To meet the ever increasing demands on quality and efficiency of spectroscopic services an entirely new approach to fully automated structure elucidation has been developed.[1] SpecSolv, as this program was titled, represents a new module for the multidimensional spectroscopic interpretation system SpecInfo. SpecSolv is a self-learning, artificially intelligent system based exclusively on ^{13}C-NMR chemical shift, intensity and multiplicity information which is readily available from ^{13}C-NMR-DEPT spectra. Unlike other structure generators, SpecSolv does not require any additional input from further spectroscopic techniques or even the knowledge of the molecular formula of the unknown. Using a dedicated data base of subspectra-substructure correlations (SSC) and a novel assembly algorithm, SpecSolv is capable of elucidating structures from most organic molecules (C, H, N, O, S, P, X) up to a molecular weight of about 1000 Da in only minutes of computing time.

INTRODUCTION

A highly efficient molecular spectroscopy laboratory represents one of the cornerstones in chemical R&D to maintain a competitive edge on a market with ever increasing requirements of product quality and product safety.

The central factor for success or failure of a product or even an entire company is determined by the time period until an innovation made in the laboratory reaches the market. Success lies with the spectroscopist who plays a key role at most stages of development of a new product. At all times, chemists, engineers, and managers must be able to base their decisions on fast and reliable information from the spectroscopist. Tasks like finding a new lead structure in pharmaceutical research, the development of a technical syntheses, or stringent legal registration processes, quality control of product, and environmental monitoring of production can only be achieved competitively with the immediate availability of spectroscopic data.

With the years acquisition and generation of spectroscopic data has become increasingly facile due to full computer control of the modern spectrometers and the use of automatic sample changers. Also, most larger laboratories employ a more or less sophisticated laboratory information management system (LIMS) to handle raw and processed data and/or to provide some type of accounting and archiving service. *However, the efficiency of a spectroscopy lab is determined by the rate at which it delivers answers to analytical questions and not by the rate of data generation*. Accepting this paradigm, one is able to identify the bottleneck in most spectroscopy labs that focus on structure elucidation: The interpretation of raw data.

Recognizing this limitation, there have been several approaches to develop proficient structure elucidation systems over the last 30 years.[2,3] Some early attempts, mainly those relying mainly on mass spectrometric techniques, like DENDRAL were abandoned in the meantime.[4,5] Being one of the pioneers of the field, Sasaki followed a multidimensional approach, using information from several spectroscopic techniques to efficiently restrict the solution space, *i.e.*, the number of possible hit structures. His CHEMICS program was one of the first attempts to automated structure elucidation.[6,7] Utilizing ^{1}H-, ^{13}C-, 2D- NMR, IR, and mass spectrometry CHEMICS proved to be a viable tool for the spectroscopy expert.[8−11]

At BASF we have been developing the multidimensional spectroscopic interpretation system SpecInfo over a period of about two decades.[12] The latest addition to the program package is the module SpecSolv, a truly innovative tool for fully automated structure elucidation, based solely on ^{13}C−NMR. Contrary to other structure elucidation tools and structure generators,[13−16] SpecSolv does not require any further information from other spectroscopic techniques, expressly not even the knowledge of a molecular formula.

DISCUSSION

Structure elucidation with SpecSolv is divided in three individual steps: (1) acquisition of experimental ^{13}C- and DEPT-NMR spectra and extraction of chemical shift, intensity, and multiplicity information, (2) subspectra search in a dedicated subspectra−substructure correlation (SSC) library and generation of a hit list, and (e) assembly of the substructures using an innovative approach that takes advantage of overlapping substructure information, followed by validation steps for intermediate substructures and the final result. SpecSolv was written in ANSI-C and runs on various platforms, like SUN Solaris and DEC VMS. The core components of the methodology will be described below.

Data Base. The SpecSolv knowledge base consists of a dedicated SSC data base[17] containing more than 400 000 substructures from three-sphere HOSE codes[18] and more than 100 000 from two-sphere HOSE codes. It was derived from our in-house SpecInfo data base with over 200 000 ^{13}C-NMR spectra, including a wide cross-section of literature spectra as well as specific BASF chemistry. High quality and maximum diversity of spectral data are key characteristics of the ideal data base. For all new entries to the SpecInfo data base and the corresponding SSC library both aspects are carefully checked by an expert.

[†] Dedicated to Prof. S. Sasaki on occasion of his retirement.

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

0095-2338/96/1636-0221$12.00/0

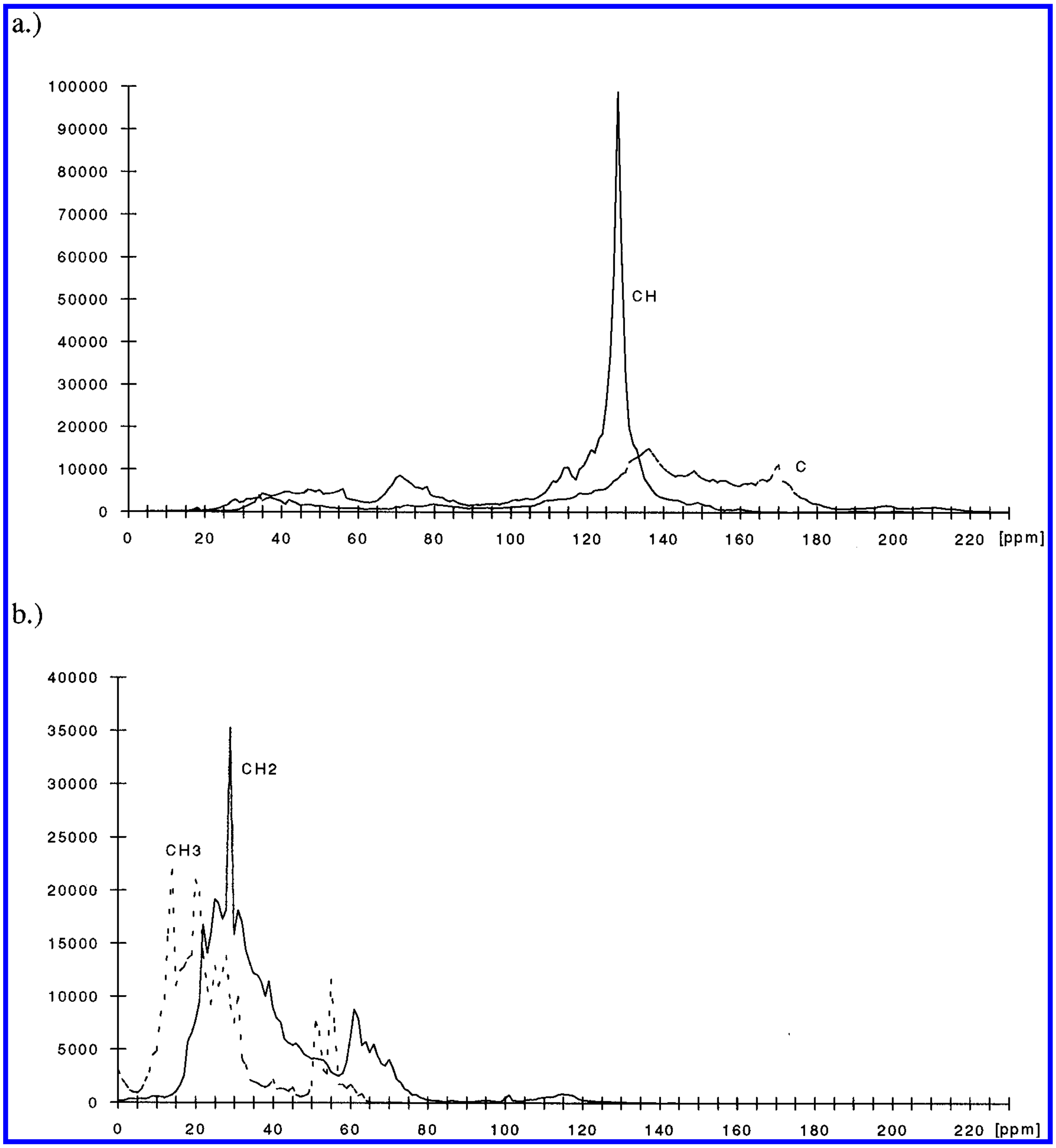


Figure 1. ^{13}C-chemical shift distribution in SpecInfo: (a) shows the distribution of singlets (C) and doublets (CH) and (b) of triplets (CH_2) and quartets (CH_3).

One SSC set contains complete structural information. It consists of subspectrum, connectivity matrix, and the HOSE code for all heavy atoms. ^{13}C-NMR parameters (chemical shift, multiplicity, intensity, and rms-values for chemical shifts) are stored for each carbon atom of every substructure. The rms-values were derived from sets of equivalent substructures in the SpecInfo data base. For unique SSCs no statistical chemical shift deviation is available, so the rms-value for each carbon atom is fixed to a default value depending on the distance from the free valences of the substructure. During subspectra searches these rms-values are utilized as uncertainty parameters within the search window. Besides, each SSC set contains the ^{13}C-multiplicity of the outer (= "open") sphere and technical indices for subspectra search and structure generation process. In contrast to the original SpecInfo-HOSE-code structure encoding, all substructures of the library are completely defined, including ring closures in the outer sphere. During the generation of the SSC library from the SpecInfo data base all chemical shifts were cross-validated, and chemical shifts with significant deviations from expected values were examined by experts.

Subspectra Search. The subspectra search consists of a comparison of all subspectra of the SSC data base with the spectrum of the unknown compound (query spectrum). During a subspectra search only SSCs with chemical shifts, intensities, and multiplicities matching the ones of the query spectrum within a user-defined window are accepted. A typical search yields about 500 substructures (SSCs) for a molecule of about 25 heavy atoms.

SpecSolv offers manual or on-line query input from the NMR-spectrometer via JCAMP files. The subspectra search is spliced into three parts, presearch, main search and fine search.[19] For the presearch all subspectra (SSC) are coded as bit strings encoding the *presence* of chemical shifts with a defined multiplicity within a discrete section of the ^{13}C-NMR-spectrum. The width of these chemical shift sections was determined by the chemical shift distribution of our data base (Figure 1) and chosen in a way that each bit string is coding for a similar number of subspectra. A second bit

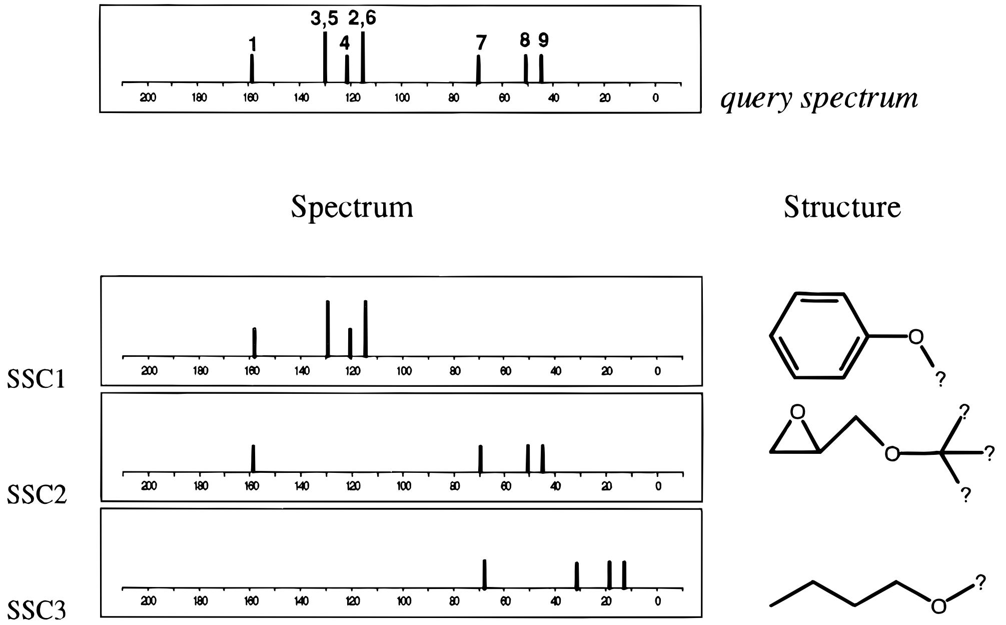


Figure 2. Example for substructure search. Only SSC no. 1 and no. 2 are accepted as hits, showing adequately overlapping chemical shifts, intensities, and multiplicities with the query spectrum. Substructure no. 3 is rejected for lack thereof.

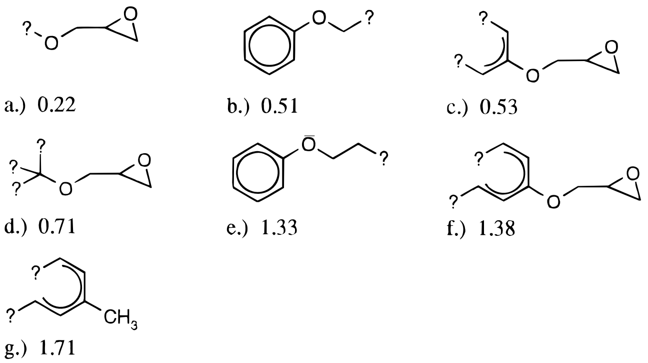


Figure 3. Substructure search. Resulting substructures a−g for unknown by use of SpecSolv standard parameters. The match factor below indicates the average deviation of the chemical shifts between SSC and query spectra in ppm. Question marks indicate open valuences. The number of question marks is identical to the number of open valences.

string contains the *number* of atoms for a given *atom type*. The presearch is done using SQL. During a second step (main search) it is tested whether the number of chemical shifts of a presearch result has a matching *number of lines* within a defined chemical shift window of the query spectrum. The fine search probes for the exact *deviation of the chemical shifts* in both spectra. This final search window typically corresponds to deviations of 2−5 ppm. The surviving substructures are then ranked according their match factors, which are simply defined as the average chemical shift deviation of each atom from the query spectrum in ppm. An example for the substructure search is shown in Figure 2. Spectra 1 and 2 are accepted as hits. All lines of SSC 1 and 2 can be correlated with resonances in the spectrum of the unknown. For example, SSC 3 is rejected for the lack of resonances within the search windows of the query spectrum. Without any restriction of atom types the substructure search yields in total 90 SSCs for the query spectrum.

SpecSolv does not require the knowledge of the atom types in the unknown. However, when the atom types of the unknown molecule are restricted to C, H, and O, the same subspectra search yields only seven SSCs (Figure 3). This example shows the significance of nitrogen containing SSCs. Due to the fact that ^{13}C-NMR is “blind” with respect to all hetero atoms, multivalent hetero atoms lead to a vast number of false hits that possibly require extensive computing resources during the assembly process. As discussed below, a relatively exact estimate of the nitrogen and phosphrous content of the unknown benefits the speed of the automated structure elucidation. Even the short hit list (Figure 3) may—and does—contain “wrong hits” (Figure 3e,g). This is the point where most other structure elucidation systems resign and the spectroscopist has to take over. With a typical number of hits for a molecule with 25 heavy atoms being 500, this shows the main problem with SSCs. How to select the correct substructures? Our answer is a new, innovative structure generator.

Structure Generator. In a novel assembly process the substructures from the SSC search are linked via overlapping common atoms of any two substructures. Several plausibility tests are systematically carried out before a resulting (larger) substructure is accepted and overlaid with the next SSC search result. In this fashion the final structure is being assembled and at the same time cross-validated with a new ^{13}C-NMR shift prediction based on SSCs.

In the HOSE-code description of the SSC each atom of a substructure has an individual HOSE-code, and neighboring atoms have somewhat redundant, *i.e.*, overlapping, descriptions. This redundancy in structural description is the key to the new assembly process.

Figure 4 shows different possibilities of overlapping two substructures. Maximal overlap of two substructures is shown in Figure 4a. Atoms marked with **C** and **C′**, respectively, are the central atom of a particular substructure. If two substructures of adjacent central atoms are matched (Figure 4a), maximal overlap of two substructures via two respectively three atoms to either side of **C** is achieved. Figure 4a leads to a consecutive chain of eight carbon atoms. Figure 4f shows the minimum match of substructures. Only the peripheral atoms of two substructures are overlapped.

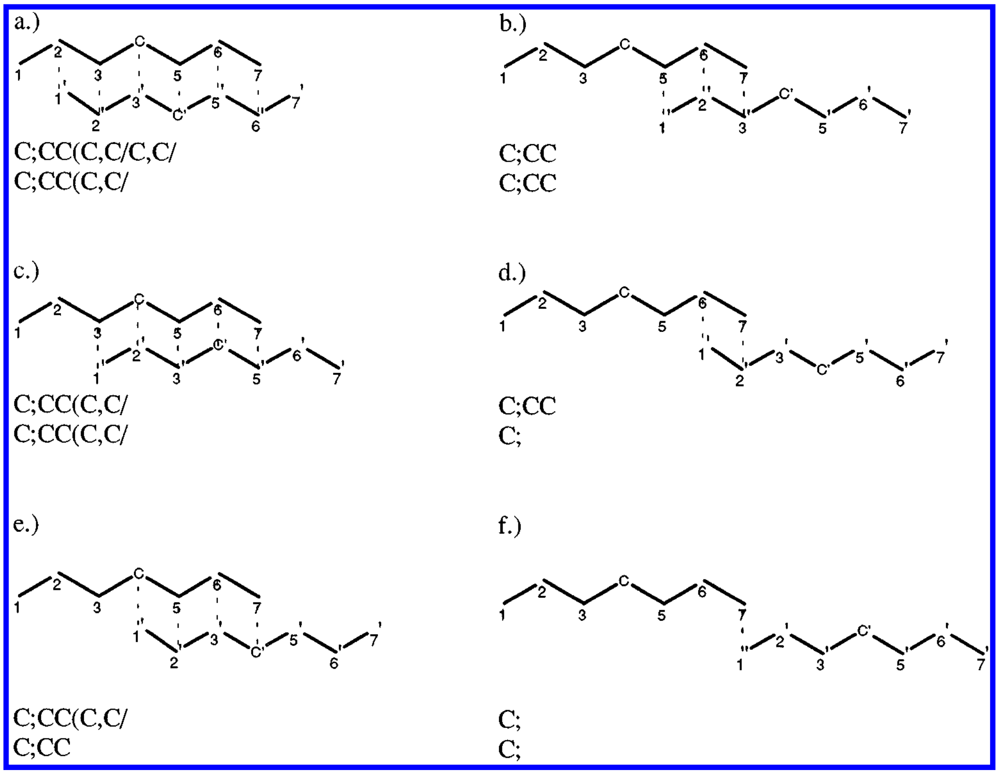


Figure 4. Classes of substructure overlaps.

This link leads to a new chain of 13 continuous carbon atoms.

The examples in Figure 4 shows that there is no need to find all substructures associated with a carbon atom of an unknown molecule. The structure generator can “jump” over up to five missing heavy atoms. Different types of overlap have unlike advantages and difficulties. Figure 4a yields a good chemical shift prediction, but the diversity of accessible molecules is limited, whereas Figure 4f leads to high diversity but only poor spectral prediction. We achieve the best results using the overalp types of Figures 4a−e concurrently. For big molecules (>500 Da) Figure 4f leads to a combinatorial explosion of structures, requiring extensive cross-validation.

Another new feature of the SpecSolv program package is the new spectra prediction method. As mentioned above, a SSC-set contains complete subspectral information. During the assembly process of the individual substructures, all chemical shifts of the resulting (sub)structure are available in the main memory of the computer and a new (sub)-spectrum can be generated by simple combination of the original subspectra. The process is very fast, because access to a data base is not necessary. Moreover, it allows one to check every intermediate substructure as well as the final structure, whether its predicted spectrum is compatible to the query spectrum or not. Assembled (sub)structures with chemical shift deviations outside a defined window are rejected.

This instant validation of intermediate structures is one main difference from other structure generation programs,[2,3] which typically validate structures only after generation of an entire molecule. In these cases spectra of a vast number of wrong structures must be predicted and compared to the target spectrum, rendering automated structure elucidation impossible. Because SpecSolv rejects false hits at a very early stage of the assembly process, the solution space requiring verification is significantly smaller than that for conventional structure generators. In turn, this leads to a crucial reduction of computing time compared to conventional programs.

Using the overlap technique and the new spectra prediction methodology, automated structure elucidation becomes feasible with molecules up to a molecular weight of ca. 1000 Da. Larger molecules require computing time in the range of hours rather than minutes due to the number of combinatorial possibilities to overlap the substructures.

Figure 6 describes a simplified flow chart of the structure generator in SpecSolv. The assembly process starts with the largest substructure and best match factor. A second substructure is picked (step 2) and tested for partial structural identity with the starting substructure (step 3). If there is no redundant (overlapping) atom found, SpecSolv chooses the next substructure (step 2). In cases of a substructural match both subspectra are combined, and the resulting subspectrum is compared to the experimental spectrum (step 4). If the two spectra are not compatible the system goes back to step 2. If they are compatible, the connection table for the bigger substructure is generated (step 5). Now the newly generated substructure is used as starting structure (step 1), and steps 2−5 will be repeated until all substructures are overlapped with the starting substructure and/or all lines of the query spectrum are completely assigned. Usually 10 cycles (10 best substructures) are sufficient for the elucidation of an unknown structure. This is due to the ranking by match factor with the most plausible substructures having the best match factors. Obviously, more cycles with deviating substructures will not lead to a correct final structure.

There have been earlier approaches to use overlapping substructure information to generate structures from spectroscopic information. Two of the programs developed the farthest were GENOA and CONGEN.[20,21] These early attempts did not succeed for a number of reasons, but one central factor was the absence of a stringent verification tool

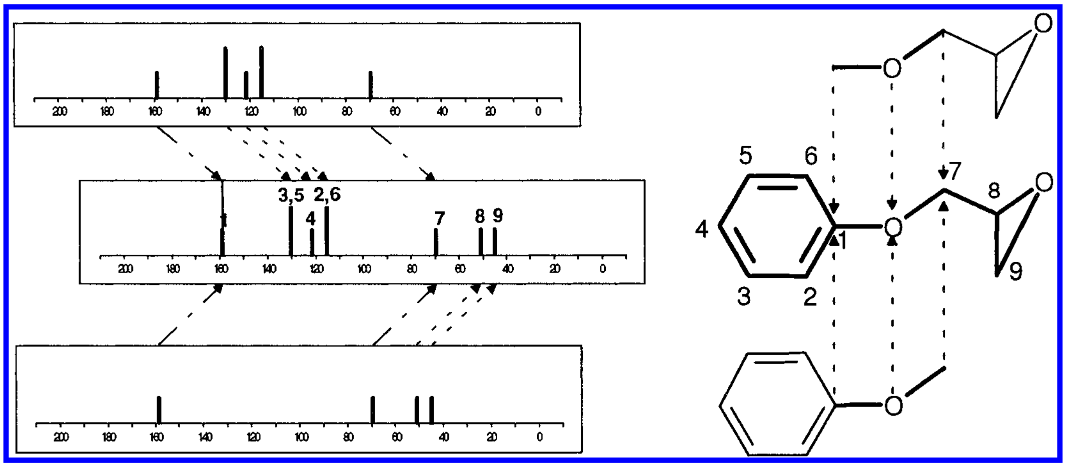


Figure 5. Example of substructure and subspectra match.

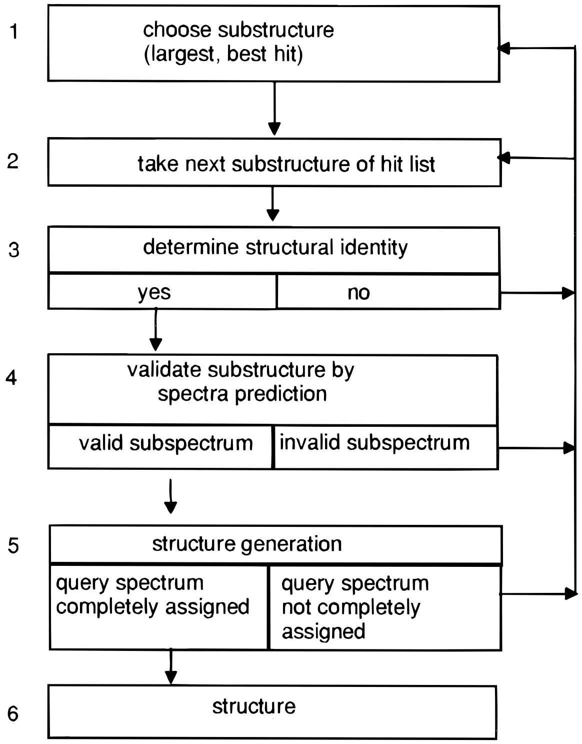

1
choose substructure (largest, best hit)
2
take next substructure of hit list
3
determine structural identity
yes
no
4
validate substructure by spectra prediction
valid subspectrum
invalid subspectrum
5
structure generation
query spectrum completely assigned
query spectrum not completely assigned
6
structure

Figure 6. Flow chart of SpecSolv.

for intermediate substructure proposals. With frequent cross-validations of the intermediate substructure proposals with the query spectrum based on the novel chemical shift prediction, SpecSolv is capable to overcome this problem very efficiently.

The assembly process of SpecSolv can be described as graph (Figure 7). All hits of the subspectra search are starting nodes **SN** of a graph. The overlap of a matching substructure (numbers 1−19) with the starting structure leads to an intermediate substructure, depicted by the nodes of the graph (letters A−S). A particular search path is terminated if the spectrum resulting from the overlap is not completely contained in the query spectrum (open circles). Otherwise, the assembly process continues sequentially until all chemical shifts of the query spectrum are assigned and the resulting structure does not carry any open valences. In the notation of artificial intelligence (AI) this is characterized as a "hill climbing" algorithm.

The number of starting structures can be adjusted to any number. SpecSolv does not stop after arriving at the first final structure suggestion. Because more than one structure can potentially describe the unknown spectrum, SpecSolv does explore the search space defined by the SSC hits completely. Nevertheless, due to the high-accuracy validation steps, generally the assembly process only yields one final structure, with exception of some cases where the resolution of the data base information is not quite sufficient to differentiate between two or three isomers. In rare cases, typically with "exotic" unknowns SpecSolv is not able to assemble a final structure for the lack of SSCs in the data base describing unusual functionalities. In these instances SpecSolv returns the largest assembled substructures with the best match factors.

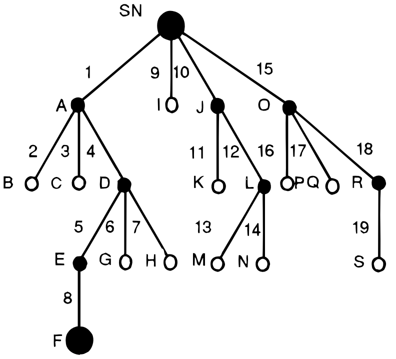


Figure 7. Representation of the SpecSolv assembly process as graph. The solution path of one final hit is 1-A-4-D-5-E-8-F.

SpecSolv does not only contain an artificially intelligent assembly algorithm but also a self-learning data base. Regarding the variety of substructure overlap in Figure 4, it becomes obvious that the system generates new SSCs. The new SSCs are associated with the atoms between two central atoms **C** and **C′** of two overlapped substructures. These new SSCs can be extracted from the assigned experimental data after the structure elucidation process and stored in the SpecSolv knowledge base.

Examples. The new assembly algorithm makes SpecSolv currently the most powerful and fastest structure elucidation tool in the field. For example, using computing time of less than 3 min on a VAX 6610, SpecSolv was able to automatically elucidate the structure of compound **1** that gave rise to the following ^{13}C-NMR spectrum (Figure 8).[22]

Besides the chemical shift information of each carbon atom SpecSolv requires the corresponding multiplicities and intensities of the ^{13}C-NMR spectrum. An upper ceiling for

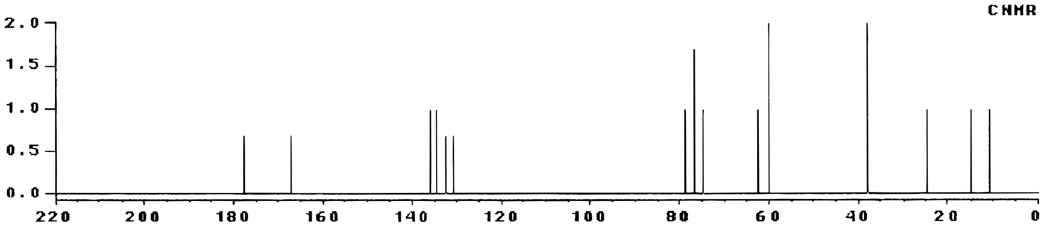


Figure 8. ^{13}C-spectrum of unknown compound **1**.

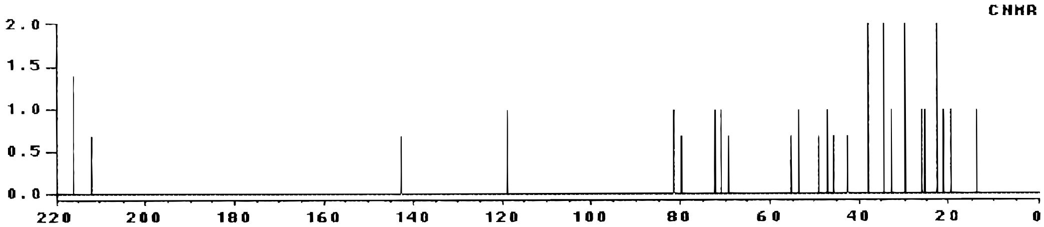


Figure 9. ^{13}C NMR of 23,24-dihydro-15-oxocurcurbitacin F.[24]

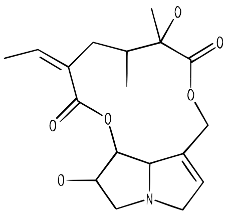
Table 1. Input Table for Unknown **1**[a]

Δ	multiplicity	rel int	Δ	multiplicity	rel int
177.80	S	1	74.60	D	1
166.90	S	1	62.40	T	1
135.90	D	1	60.30	T	1
134.70	D	1	60.00	T	1
132.40	S	1	38.20	D	1
31.00	S	1	37.90	T	1
78.60	D	1	4.80	Q	1
76.70	D	1	4.90	Q	1
76.60	S	1	0.80	Q	1

[a] Maximum molecular formula: $C_{100}H_{100}O_{100}N_{100}$.

a molecular formula should be entered, here $C_{100}H_{100}O_{100}N_{100}$. To minimize computing resources, the number of nitrogens should be estimated relatively close to a realistic value. Without any upper ceiling for the molecular formula or a higher number of nitrogen atoms Spec Solv will still find the correct solution, however, at the expense of increased computing time. The entire input table for this example is shown in Table 1.

The subspectra search yields 150 substructures using SpecSolv's standard parameters. From this set of hits the structure generator builds only one molecule that describes the complete set of ^{13}C-NMR resonances with a match factor of 0.44 ppm. The structure was identified as uspalatine, the molecule described in ref 21.

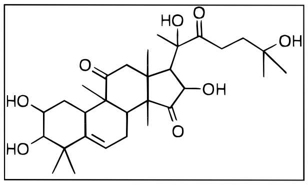


Structure of unknown **1**, Uspalatine

Uspalatine itself was not included in the SpecInfo/SpecSolv knowledge base, indicating that this structure elucidation was a true "interpretation" and not just a "retrieval" process. The completely assigned experimental spectrum is now readily added to the SpecInfo data base as well as the SpecSolv SSC library. An average chemical shift deviation per carbon atom of 0.44 ppm demonstrates the selectivity of the new ^{13}C-shift prediction in SpecSolv. For comparison, the HOSE code shift prediction in SpecInfo yields an chemical shift deviation per carbon atom of 0.55 ppm.

23,24-Dihydro-15-oxocurcurbitacin F, **2**, is a molecule with steroid skeleton and a molecular weight of 534 Da. The ^{13}C-NMR spectrum consists of 30 resonances (Figure 9). Proceeding in the same fashion as described above the substructure search yields 1118 SSCs for the spectrum.

The structure generator yields two candidates in less 3 min computing time. Both structures have good match factors and explain the experimental spectrum; however, the second ranked contains an unusually bridged cyclic system. The best ranked structure corresponds to the correct molecule **2**.



2, 23,24-Dihydro-15-oxocurcurbitacin F,

Limitations. Currently, the structure of ~80% of all pure organic compounds (C, H, N, O, S, P, X with a molecular weight ≤1000 Da) measured routinely in our laboratory can be resolved automatically using SpecSolv. As any library-based system SpecSolv is only capable to elucidate structures that can be described by overlapping SSCs. Thus, it depends crucially on quality and diversity of the SSC data base. Its most sensitive module, the chemical shift prediction, is negatively affected by solvent effects or stereochemical influences on the experimental data, requiring stringent quality control by an expert during update and operation of the data base. Another problem are truncated substructures or HOSE-codes, which have been long-standing issues of discussion with SpecInfo.[12] So far, stereochemistry information has not been introduced to the SSC data base, such that SpecSolv does not describe any configurational effects yet.

CONCLUSION AND OUTLOOK

SpecSolv is a novel automated structure elucidation system based on ^{13}C-NMR and 1D-DEPT spectra that does not require any additional input from other spectroscopic techniques, such as mass spectrometry or IR. Most importantly, SpecSolv does not necessitate the knowledge of the molecular formula of the unknown. It allows unsupervised operation and presents undoubtedly the fastest system currently available. Future developments include the addition of stereochemical information to the SSCs to further improve the chemical shift prediction and an algorithm that allows for the resolution of the individual components of a mixture spectrum automatically. In principle, SpecSolv is by no means restricted to ^{13}C-NMR. Conceivable future projects cover the extension to ^{1}H- and 2D-NMR methods like HSQC, COSY, or TOCSY or even other spectroscopic techniques like IR or MS. With the addition of SpecSolv, the SpecInfo program package matured to a self-learning artifically intelligent structure elucidation system. Using the notation of Neudert et al.,[23] SpecSolv would be a fourth level structure elucidation tool.

REFERENCES AND NOTES

(1) Will, M. In *Proceedings of the 9th Workshop "Computer in Chemistry" Halle, 1994*; Jochum, C., Ed.; in press.

(2) Warr, W. A. *Anal. Chem.* **1993**, *65*, 1045A−1050A.

(3) Warr, W. A. *Anal. Chem.* **1993**, *65*, 1087A−1095A.

(4) Carhart, R. E.; Varkony, T. H.; Smith, D H. In *Computer Assisted Structure Elucidation*; Smith D. H., Ed.; ACS Symposium Series 54, American Chemical Society: Washington, D.C., 1977; p 92.

(5) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. *J. Am. Chem. Soc.* **1969**, *91*, 2973.

(6) Yamasaki, T.; Abe, H.; Kudo, Y.; Sasaki, S. In *Computer Assisted Structure Elucidation*; Smith, D. H., Ed.; ACS Symposium Series 54, American Chemical Society: Washington, D.C., 1977; p 108.

(7) Sasaki, S.; Kudo, Y.; Ochiai, S.; Abe, H. *Microchim. Acta* **1971**, 726.

(8) Sasaki, S.; Kudo, J. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 252−257.

(9) Funatsu, K.; Susuta, Y.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 6−11.

(10) Funatsu, K.; Acharya, B. P.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 735−744.

(11) Funatsu, K.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 745−751.

(12) Bremser, W. *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 247−260.

(13) Christie, B. D.; Munk, M. E. *J. Am. Chem. Soc.* **1991**, *113*, 3750−3757.

(14) Funatsu, K.; Miyabayashi, N.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18−28.

(15) Kerber, A.; Laue, R.; Moser, D. *Anal. Chim. Acta* **1990**, *235*, 221−228.

(16) Elyashberg, M. E.; Serov, V. V.; Martirosyan, E. R.; Zlatina, L. A.; Karasev, Y. Z.; Koldashev, V. N.; Yampolski, Y. Y. *J. Mol. Structure (Theochem)* **1991**, *230*, 191−203.

(17) Utilized software and firmware: Relational data base management system SYBASE SYSTEM 10, implemented on a SUN Sparc20 with Solaris 2.3 operating system.

(18) Bremser, W. *Anal. Chim. Acta* **1978**, *103*, 355.

(19) Will, M., manuscript in preparation.

(20) Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. *J. Org. Chem.* **1981**, *46*, 1708−1718.

(21) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. *J. Am. Chem. Soc.* **1975**, *97*, 5755.

(22) Pestchanker, M. J.; Ascheri, M. S.; Giordano, O. S. *Phytochemistry* **1985**, *24*, 7, 1622−1624.

(23) Neudert, R.; Penk, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 244−248.

(24) Konoshima, T. *Chem. Pharm. Bull.* **1993**, *41*, 1612.

CI950092P