NPR
www.rsc.org/npr

Recent developments in automated structure elucidation of natural products

Christoph Steinbeck *
Cologne University Bioinformatics Center (CUBIC), Zülpicher Str. 47, 50674 Köln, Germany.
E-mail: c.steinbeck@uni-koeln.de; Fax: +49 221 470 7786; Tel: +49 221 470 7426

Received (in Cambridge, UK) 17th June 2004,
First published as an Advance Article on the web 14th July 2004

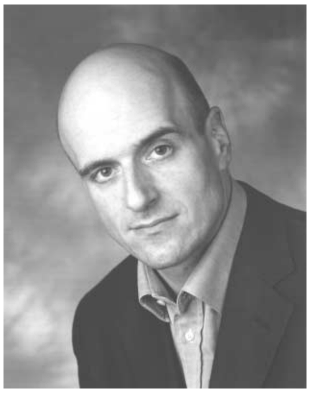
Covering: 1999–2004

Advancements in the field of Computer-Assisted Structure Elucidation (CASE) of Natural Products achieved in the past five years are discussed. This process starts with a dereplication procedure, supported by structure-spectrum databases. Both commercial and free products are available to support the procedure. A number of new programs, as well as advancements in existing ones, are presented. Finally, the option to validate the result by an independent procedure, a high quality *ab initio* quantum mechanical calculation, is discussed.

1 Introduction

A tool for Computer-Assisted Structure Elucidation (CASE) is supposed to provide a chemist, spectroscopist, or anyone else dealing with elucidating the chemical structure of an unknown compound, with suggestions for the compound's molecular skeleton, based on spectroscopic data and other prior knowledge, with the use of a computer program and the least possible amount of user intervention. Despite a good amount of research that has been performed in this field since the late 1970s,[1] usable software has been offered to the practitioner only very recently. The emergence of comprehensive commercial packages, and advancements in existing academic projects for this purpose, has made it desirable to review the current literature starting from where Marcel Jaspars stopped in 1999.[2]

Christoph Steinbeck was born in Neuwied, Germany, in 1966. He studied chemistry at the University of Bonn, where he received his diploma and doctoral degree in the workgroup of Prof. Eberhard Breitmaier at the Institute of Organic Chemistry. The focus of his PhD thesis was the program LUCY for computer assisted structure elucidation. In 1996, he joined the group of Prof. Clemens Richert at Tufts University in Boston, MA, USA, where he worked in the area of biomolecular NMR on the 3D structure elucidation of peptide–nucleic acid conjugates. In 1997 he became head of the Structural Chemo- and Bioinformatics Workgroup at the newly founded Max-Planck-Institute of Chemical Ecology in Jena, Germany. Since fall 2002 he is head of the Independent Research Group For Applied Bioinformatics at the Cologne University Bioinformatics Center (CUBIC). Dr Steinbeck is vice chairman of the Chemistry-Information-Computers (CIC) division of the German Chemical Society.



Christoph Steinbeck

2 On the general strategy for structure elucidation

The details of how a human operator approaches the task of elucidating the structure of an unknown compound, and how this approach might be projected on, or reflected by a CASE process has been discussed in great detail in various articles.[2–4] In addition to what has already been said, the role of comprehensive databases containing chemical structures and (ideally assigned) spectra will be emphasized here. In the form of printed spectral catalogs they have helped and guided the researcher in pre-internet times and before personal computers became abundant on our desktops. Later, commercial database projects (SpecInfo,[5] CSearch[6] and others[7,8]) made searching for spectral and structural patterns of interest much easier and today we become aware that a truly successful CASE system will have to be based on a large database of spectral and associated structural features. Both the integration of such databases in existing CASE systems as well as newly emerged stand-alone systems will be discussed in this article.

With respect to the type of spectroscopic methods useful or needed for successful structure elucidation, one- and two-dimensional NMR methods have become dominant – a trend already reflected in this review's predecessor.[2] A few new NMR methods with practical relevance have been published in recent years and will be mentioned later.

3 Dereplication and structure elucidation support through databases

All parts of the process leading to an elucidated structure have experienced an immense speed-up in the past fifty years. Separation technology, analytical and spectroscopic methods have improved steadily and with good fortune, a chemist might be able to go from a crude extract to a full set of 2D NMR spectra in one day. Of course, the situation can be much worse. Clearly, CASE tools have to keep up with this development. Manually, the elucidation of a complex, hitherto unknown molecular skeleton can take from days up to even months.

DOI: 10.1039/b400678j

In order to avoid the potentially tedious route of *ab initio* structure elucidation – just to discover after some time that the compound was already known – a process called dereplication is used to perform prescreening and to quickly detect known compounds. Here, spectral fingerprint data, which can be mass (MS), infrared (IR) or NMR spectra, are used for spectral similarity searches in either public or in-house structure-spectra databases. Since a close-hit-spectrum will be linked to a chemical structure in such a dereplication database, both avoiding tedious and surplus work as well as identifying very similar compounds can be the merits of such a procedure.

A few early attempts have been made both by large companies as well as academic groups to create databases for all kinds of spectroscopic data. We will focus on NMR databases here, since the use of mass spectral databases is highly dependent on the technique and spectrometer used and IR has lost some ground in this field since the time when people searched printed spectral catalogs for the fingerprint IR they had in hand.

SpecInfo[5] and CSearch,[6] two repositories with a significant amount of data sets, have been on the market for quite a while. SpecInfo is available both online *via* STN International and as a stand-alone product, whereas CSearch was an academic project by Wolfgang Robien in Vienna, later commercialized and distributed by BioRad, and now available freely *via* an email interface, which clearly does not allow for routine and large scale searches. Both products contain tens of thousands, if not hundreds of thousands of structures with their assigned spectra. The assignment is a vital aspect since it is the declared goal of both systems to provide spectrum prediction based on the database content.

In the course of their incredibly dynamic creation of a new commercial structure elucidation system, which will be discussed later, the Canadian firm ACDLabs has, in more recent years, assembled a similarly large database.[7]

Given the long history of NMR and its usage in structure elucidation, it is more than surprising that no publicly assembled database comparable, for example, to the large genomic databases, has arisen in this field. This realization, together with the success of the open source movement in creating free software such as the Unix-like operating system Linux, has guided Steinbeck and coworkers to start the NMRShiftDB project.[8]

NMRShiftDB is an open-access, open-submission database of organic compounds and their NMR data. Also the software driving the system is completely open-source and can be freely copied to replicate the system. NMRShiftDB is available to the public *via* a web-interface on http://www.nmrshiftdb.org (and recently also *via* an alpha-quality stand-alone client). Currently, the project is based on a fail-safe cluster of four mirrored internet servers in three different geographic locations in order to assure high availability for its user base. At present, the database is in its early stages in terms of the amount of data stored. About 9000 compounds with their assigned spectra, mostly carbon NMR, can be used for searching and spectrum prediction.

As has been indicated above, the novel aspect of this project is the possibility for users to enter data and to help the project grow. Submitted datasets are sent to reviewers, also recruited from the user base in order to ensure a good quality level of the database. For this reason, a contributor needs to register an account with the database, to allow both reviewers and database editors to get in touch with him in case of problems with submitted data. Currently, NMRShiftDB's functionality comprises (sub-) spectra and (sub-) structure searches as well as searches in general text fields, keywords, bibliographies and measurement conditions (solvent, temperature, spectrometer frequency). Fig. 1 shows a schematic representation of NMRShiftDB's dataset structure.

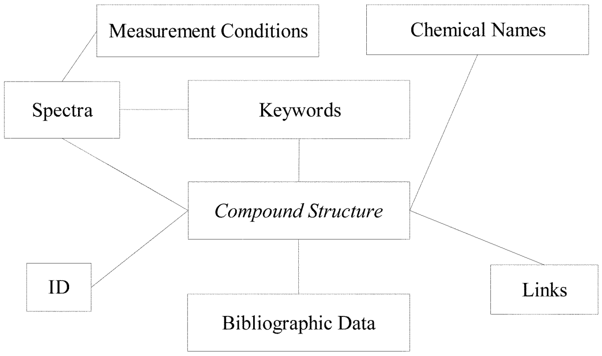


Fig. 1 Simplified structure of an NMRShiftDB dataset. A variety of facts can be queried for each of the components shown. For the compound structure, various query options exist for the molecular formula (gross formula), in addition to the common structure, substructure and similarity searches.

It is intended to expand the system to store other types of spectroscopic data grouped around the central molecular structure. Certainly, an open repository of molecular structures and their assigned spectroscopic properties is a very worthwhile endeavour which deserves the support of both users and publishers, which could require their authors to deposit spectral data published in articles in NMRShiftDB. Since NMRShiftDB provides an on-the-fly quality assessment for entered data, many of the errors still found in the assignment of published data could be prevented.

4 New and updated CASE systems

With a history of more than 25 years now,[1,9] there has long been a set of well documented CASE systems, none of which, however, found any broader application.[3,4] The past five years now have both seen update reports for existing systems as well as the creation of new ones.

One of the most established CASE systems, SESAMI,[3] was developed by the group of Morton E. Munk of Arizona State University at Tempe, Arizona. It integrates two distinct lines of structure elucidation approaches, one based on the principle of *structure assembly*, the other one based on *structure reduction*.

Assemble, a structure assembler, was recently advertised as a stand-alone module, termed Assemble 2.0, to serve as a proposal generator[10] for the practitioner. The authors stress that Assemble 2.0, being a pure structure generator, has only the most basic knowledge about organic chemistry and does not make any attempt to perform spectrum interpretation. Rather, it relies purely on the information explicitly entered by the user.

In order to use Assemble to aid in structure elucidation problems, the user would thus perform the spectrum interpretation himself, and would then feed the extracted structural information into the structure generator. Information digestible by Assemble comprises non-overlapping substructures, minimal or maximal counts for double and triple bonds, number of carbon signals in the molecules used for symmetry perception, number of expected rings, hydrogen counts and hybridization for heavy atoms, and more. As a result of the subsequent structure generation process, Assemble 2.0 will then provide the user with an exhaustive list of non-isomorphic structures fitting the above constraints. This list, if containing more than just a few members, can then be ranked according to the agreement of their predicted proton or carbon NMR shifts with the experimental spectrum of the unknown compound.

Clearly, a system like the above has its merits due to the great transparency of the process for the user, who performs the interpretation of the spectral data himself, generating a list of substructures or fragments, and then uses a tool to get an exhaustive list of structures containing all of the input fragments.

COCOA, a structure generator based on structure reduction, has been incorporated into the SESAMI system in 1988. Recently, Munk and coworkers reported the development of a new structure generator HOUDINI[11] with significantly improved performance over the old COCOA program. HOUDINI is based on two central data structures: firstly, a square matrix of atoms constructed upon input of the molecular formula, on which the new approach relies, as do most other CASE programs. This square matrix is a representation of a hyperstructure, a structure initially containing all possible bonds in agreement with some starting criteria (if no user-defined substructures are present, all connections between any pair of atoms are possible). Secondly, a data structure called substructure representation (SR) is used by HOUDINI, which consists of substructures in the form of atom-centered fragments (ACFs). For each spectroscopic constraint present – say, an HMBC signal – the SR list contains a family of all possible interpretations of this constraint, only one of which can be true. During the structure elucidation process, HOUDINI tries to map ACFs from the SR list onto the hyperstructure, thereby establishing certain bonding constraints. Obviously, many ACFs will mutually exclude each other. Munk and coworkers compare their new approach with the previously used COCOA reductive structure generator.[12]

A set of seven test cases with target structures of between 16 and 76 heavy atoms (non-hydrogen atoms) has been presented. In all cases, 1D proton and carbon NMR, as well as HMQC, HMBC and COSY spectra have been used. The authors emphasize, however, that the system is not restricted to the use of NMR data. All test compounds are molecules with proton-rich carbon skeletons. Herein, heteroatoms are typically non-skeletal atoms like keto-, hydroxyl- or other peripheral groups. For this type of molecule a reduction of the solution space to only a few candidates is usually easy because the molecule's skeleton is uniformly covered by the spectral data, leaving no blackspots. The authors of another new CASE program, COCON, have explicitly addressed this problem, which will be discussed later. HOUDINI is able to solve all of the test cases in less than a minute on a standard desktop PC, if all of the available spectroscopic data is used. The authors put emphasis on the scaling of the computation time with decreasing amounts of spectroscopic constraints, because they noticed earlier that their COCOA-based SESAMI system scaled particularly badly in this respect. Therefore, all of the computation times were measured with and without the use of COSY constraints, which may, in conjunction with HMQC data, unambiguously define large portions of the carbon skeleton, thus greatly reducing the combinatorial space for the structure generator.

Indeed, HOUDINI was able to solve even such constraint-reduced cases in a reasonable time (17 min in the worst case, but usually less than a minute), whereas in three out of four cases, the COCOA-based system did not complete its run after several days of computation time.

Another newcomer, the CASE system COCON[13] was developed by Matthias Köck and coworkers. Being part of the NMR laboratory of Christian Griesinger, this group did not attempt to establish yet another comprehensive CASE system like CISOC-SES, SESAMI or CHEMICS. Rather, the focus was on the efficient use of existing, and the integration of new, NMR experiments to overcome known problems in automated structure elucidation. Besides the common set of experiments – ^{13}C NMR, a set of DEPT spectra, HMQC, HSQC – the emphasis was mainly on the use of two new experiment types, ^{1}H–^{15}N-HMBC and *1,1*-ADQUATE. The latter is one instance out of a series of possible 1,*n*-ADEQUATE[14] pulse experiments showing cross signals in a 2D NMR diagram connecting carbon atoms separated by *n* bonds where at least one of the carbons carries a hydrogen atom. Despite being descendants of the ultimately insensitive but also ultimately useful INADEQUATE experiment, they do not share the most serious shortcomings of their unfavorable ancestor. The 1,*n*-ADEQUATE sequence exploits a sensitivity enhancement of an H,C coherence transfer, coupled with an efficient suppression of undesired signal components using a gradient echo. The latter requires special NMR hardware, which, however, is becoming increasingly common in today's NMR laboratories. A variety of virtues can be attributed to these new experiments. Firstly, the 1,1-ADEQUATE detects only 1-bond carbon–carbon correlations, thus, for the first time, allowing the chemist to distinguish between $^{2}J_{\mathrm{CH}}$ and $^{3}J_{\mathrm{CH}}$ couplings in an HMBC experiment by a simple exclusion procedure.[15] Secondly, connectivities between carbon atoms up to six bonds apart can be detected – crucial information in the CASE of proton poor compounds. Fig. 2 illustrates how the information from 1,*n*-ADEQUATE could be used to elucidate the structure of 5,6-dihydrolamellarin H – an otherwise undetermined CASE problem. The drawback of these experiments is, however, a very long experiment time, compared to the conventional set (COSY, HSQC, HMBC). The recording of the spectrum for a sample of 14 mg of 5,6-dihydrolamellarin H, for example, took 2 days and 17 hours.[14]

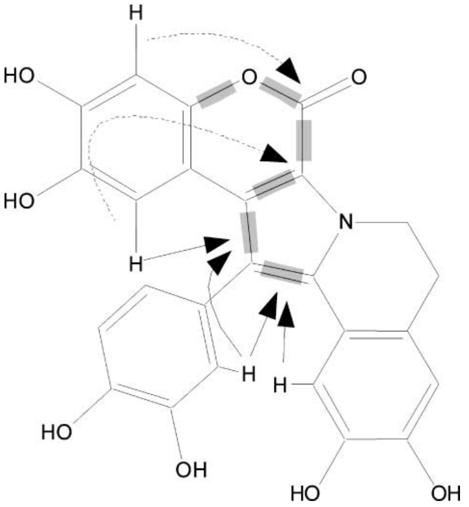


Fig. 2 Constitution of 5,6-dihydrolamellarin H: highlighted bonds indicate those that could not be deduced from HMBC information. Dashed lines indicate the most important correlations obtained from the 1,*n*-ADEQUATE experiment. Solid arrows indicate the protons at which the magnetization for the determination of the central pyrrole bonds was detected. Only those correlations necessary for deducing the constitution are shown, redundant information is omitted.

Especially useful for the automated structure elucidation of alkaloids is the ^{1}H–^{15}N-HMBC, as Köck *et al.* point out,[16] because additional constraints linking ^{15}N nuclei, here at natural abundance, with neighboring protons *via* two or three bonds, will allow the CASE program to further reduce the number of candidate solutions presented to the user.

In addition to solving difficult case problems more efficiently, the desire is to push the size limit for the CASE of larger molecules. COCON has been shown to solve CASE problems for molecules as large as ascomycin ($C_{43}H_{69}NO_{12}$, Fig. 3). It should be noted, however, that the molecular skeleton of the target compound has been extraordinarily well defined due to the use of both 1,1-ADEQUATE as well as ^{1}H–^{15}N-HMBC information. The makers of COCON have made a web version of the program, called WebCocon, available on http://cocon.nmr.de. Interested readers will find a useful example section with predefined input data also for the ascomycin example. The given example shows that almost the entire molecular skeleton is already defined by the available 1,1-ADEQUATE information, a measure that greatly reduces the combinatorial space to be searched by the program. While this nicely exemplifies the benefits to be gained from the new

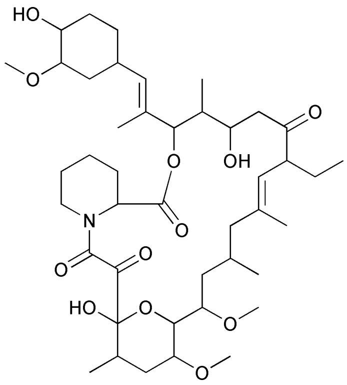


Fig. 3 Constitution of ascomycin ($C_{43}H_{69}NO_{12}$) – a compound used to demonstrate the ability of the program COCON to determine the constitution of large molecules.

experiment types, the reader should be aware that those can not be considered standard experiments in most NMR laboratories, probably both due to the difficult setup as well as the still large amount of compound needed in order to record the experiments in a reasonable time span.

The most promising achievements in terms of practical applicability of a CASE system has been made by a commercial endeavour under the guidance of Mikhail E. Elyashberg and coworkers for the Canadian company ACDLabs. They have recently demonstrated the efficiency of their CASE system *Structure Elucidator* – in short *StrucEluc* – with a large number of examples.[17,18] In reference 18 the compound polycarpol (Fig. 4), which already served as an example for other CASE programs,[19–21] was used to demonstrated the working principles of *StrucEluc.* This dataset comprises 27 HSQC signals, 46 HMBC signals and 19 H,H-COSY signals. In 1996, when the data were first used, the CASE program LUCY took about two hours on a Pentium 100 PC, whereas *StrucEluc* needed just six seconds to find the identical set of six solutions in agreement with the spectroscopic data. This impressive speed is, among other improvements, achieved by taking advantage of a large library of 215000 molecular structures with their associated ^{13}C NMR spectra. This library has been decomposed into a fragment library containing more than one million fragments and the corresponding ^{13}C NMR spectra. In what they call the "standard" mode of operation, using a technique first described by Will, Fachinger and Richert,[22] the ^{13}C NMR of the unknown compound is used as the primary source of information. In a first run, an identical spectrum is searched in the spectral library. If nothing can be found, the library of fragment spectra is searched in order to assemble a list of fragments with corresponding subspectra not contradicting the experimental spectrum. Even without the traditionally required molecular formula (MF), this fragment list, whose size can be reduced by removing impossible fragments based on ^{1}H NMR or IR data, is then used to combinatorially generate all molecular constitutions.

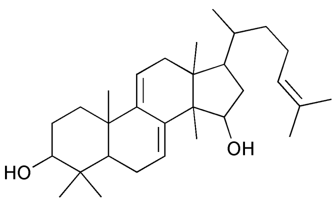


Fig. 4 Structure of the triterpene polycarpol ($C_{30}H_{48}O_2$) used as a test case by both Steinbeck[20,21] and Elyashberg *et al.*[18]

Only if no chemically correct structure can be found, the program falls back to the so called "classical" mode, now requiring the input of a molecular formula in order to supplement missing information. The classical mode has features and functionality typical for classical CASE expert systems, as there are GOODLIST (substructures that must be present in the unknown compound), BADLIST (substructures that must be absent therein), minimum and maximum ring cycle sizes, multiplicities of bonds, *etc.* Upon generation of a list of structures in agreement with the fragment set generated from the input data, the solution set can be ranked by applying a carbon shift prediction and ranking by agreement of predicted and experimental spectra – a function used throughout the system for a final decision on the right structure in all available modes. In the case of large molecules or novel skeletons, both the classical and the standard mode are likely to fail and the program provides the chemist with a third, the so-called "common" mode, which is capable of using 2D NMR information, including, but not limited to: H,H COSY, HSQC/HMQC, HMBC, ROESY, NOESY and INADEQUATE. The *StrucEluc* system allows for the import of original vendor spectral data (Varian, Bruker, JEOL) and is reportedly capable of performing peak pickings on these data. Most of the parameters of the subsequent CASE run are determined automatically. However, those parameters, like implicit H count, allowed hybridizations or likely heteroatom attachments, can be edited by the user. For a workflow scheme covering dependencies and links between the modes of *StrucEluc* see Fig. 5.

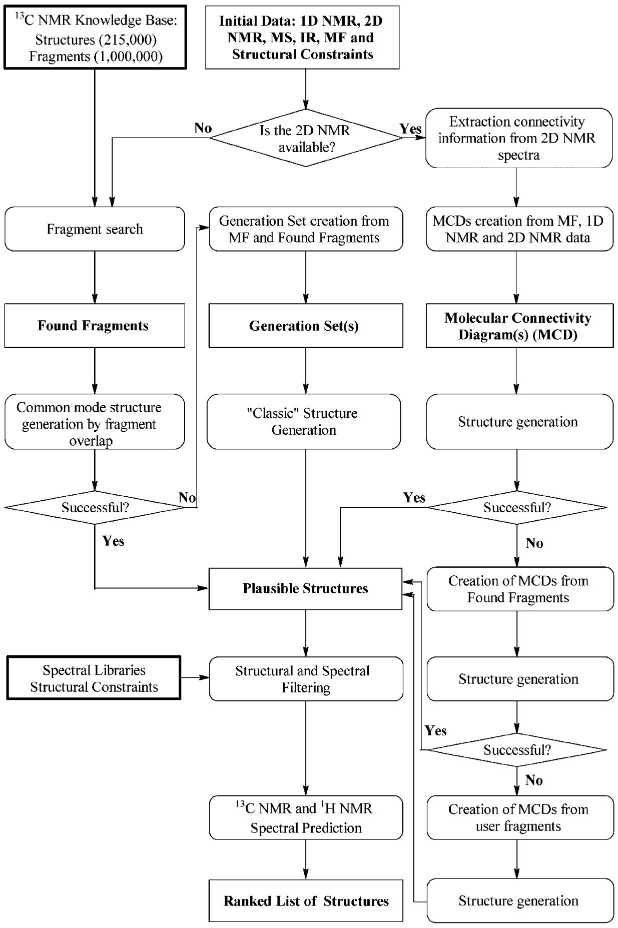

13C NMR Knowledge Base: Structures (215,000) Fragments (1,000,000)
Initial Data: 1D NMR, 2D NMR, MS, IR, MF and Structural Constraints
Is the 2D NMR available?
No
Yes
Extraction connectivity information from 2D NMR spectra
Fragment search
Generation Set creation from MF and Found Fragments
MCDs creation from MF, 1D NMR and 2D NMR data
Found Fragments
Generation Set(s)
Molecular Connectivity Diagram(s) (MCD)
Common mode structure generation by fragment overlap
"Classic" Structure Generation
Structure generation
Successful?
No
Yes
Yes
Successful?
No
Plausible Structures
Creation of MCDs from Found Fragments
Spectral Libraries Structural Constraints
Structural and Spectral Filtering
Structure generation
Yes
Successful?
No
13C NMR and 1H NMR Spectral Prediction
Creation of MCDs from user fragments
Ranked List of Structures
Structure generation

Fig. 5 General workflow scheme of the StrucEluc system. Reproduced with kind permission from reference 23. Copyright (2004) American Chemical Society.

In reference 23 the *StrucEluc* authors present an extensive study with more than 150 CASE problems that were published in recent years in *Journal of Natural Products* or where raw data was provided by cooperating laboratories. More than 60 problems were related to structures with more than 30 heavy atoms. A few problems' underlying constitutions were between 60 and 90 skeletal atoms. In all the cases, where data from other CASE systems were published and allowed a comparison, the *StrucEluc* system performed considerably faster than the competitor.

In the course of documenting use cases for the *StrucEluc* system, Martin *et al.* have presented an application of the *StrucEluc* system for identifying degradants of a complex alkaloid using NMR cryoprobe technology.[24] In this pharmaceutical application, a 2.5 mg sample of the nonacyclic alkaloid cryptospirolepine which had been stored in a sealed NMR tube in DMSO-d6 for more than 10 years (!), was studied in order to identify the products formed by its degradation during the long storage time. An LC/MS run showed that the original substance had been completely degraded to form 26 products, most of them in a 2–3% range based on the LC peak area. The two major components DP-1 and DP-2 with 35 and 16% were targeted for identification. For those, NMR samples of just 0.5 mg and about 100 µg were prepared by HPLC chromatography. Gradient COSY and HSQC spectra were recorded. Based on these spectra and mass spectrometric information, *StrucEluc* identified the compound DP-2 to be cryptolepinone by first generating 208 output structures and then performing a ranking based on carbon shift prediction (Fig. 6).

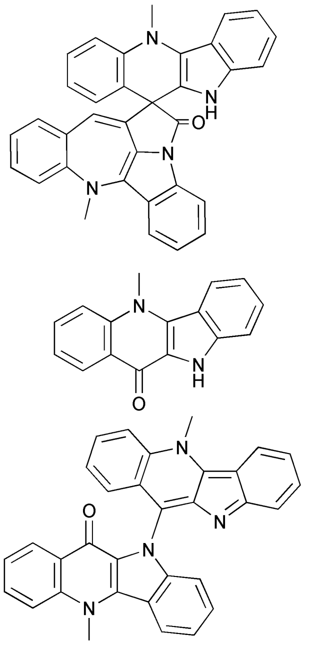


Fig. 6 Cryptospirolepine and its two main degradation products, identified using Cryoprobe Technology and ACD/*Structure Elucidator*.

Obviously, spectroscopically well-defined CASE problems of medium size (20–40 skeletal atoms) are the domain in which expert systems perform best. A large number of, for example, HMBC constraints allows a computer program do what it's good at: *Combinatorics*. And this is also the domain where such an expert system clearly outperforms a human expert when it comes to covering the whole range of possible solutions implied by the spectroscopic data. Based on the cryptolepine problem discussed above, Blinov and coauthors have investigated a more challenging application of the *StrucEluc* system to cases where only limited 2D NMR data are available.[25] In the course of this article, the authors discuss and criticize various other CASE approaches, including the work of Munk and coworkers, COCON and the SENECA program, which will be discussed below. It is pointed out, that, to date, no group has ever attempted to show the general applicability of their system to cover the diverse structural space. It is further stated that none of the various algorithms allow for the consideration of user defined fragments, thereby taking advantage of a human expert's capability to hypothesize about structural moieties in the unknown. Finally, a missing ability to detect contradictions in spectral data is noted. The *StrucEluc* system, as one may expect, is claimed to address all of these shortcomings in the course of the article. Firstly, the system is reported to be tested with more than one hundred test compounds. No attempt is made, however, to address the question of structural diversity of this test suite and its superset of all naturally existing organic compounds, be they discovered or not. With tens of millions of known organic compounds and hundreds or thousands of compound classes, neither ten nor one hundred test cases can be considered to be representative. Clearly, a database of CASE problems, accessible with open standards, would be a great help here.

Secondly, user-defined GOODLISTs have reportedly been part of other CASE systems. CHEMICS,[26] for example, supports GOODLIST fragments, admittedly only with one free valence. Faulon's SIGNATURE system makes extensive use of substructural fragments for the CASE process.[27] Regardless of the limitations of one or the other system, a strong statement like "The ability to use substructural fragments as input [. . .] does not exist" is certainly not justified.

Thirdly, the *StrucEluc* system incorporates a heuristic algorithm to search for skeletal atoms displaying connectivities of non-standard length.[17,18] It tries to overcome its problems with signal interpretation by increasing the number of bonds to be represented by a particular cross peak. For example, the range of an HMBC signal may be expanded from $^2J_{\mathrm{CH}}$ and $^3J_{\mathrm{CH}}$ to longer range CH couplings *via* four or even more bonds. Clearly, the term "contradictions" in spectral data is not justified here, since the spectral data are perfectly consistent here – as opposed to data containing artifacts – but the assumptions used to interpret them are wrong. Further, at least one CASE system, SENECA,[20,21] has been shown to deal naturally with the various long range couplings. This is done by measuring the bond path length by which a certain 2D NMR cross signal can be explained in a candidate structure. Standard path lengths ($^2\mathrm{J}_{\mathrm{CH}}$ and $^3\mathrm{J}_{\mathrm{CH}}$ in an HMBC, for example) are rated higher than the rarer longer range paths. Still, structures in which certain signals can only be explained by, say, a five-bond CH-coupling in an HMBC spectrum, will be ranked higher than those where the signal can not be established at all. Typically, cut-offs are still chosen (default: coupling paths longer than 6).

Classical CASE systems use deterministic structure generators in order to generate all possible solutions from a given set of spectroscopic and structural input data. Since most of the CASE systems discussed in the literature did not have access to large spectroscopic databases like that assembled by ACD/Labs, they were and are mostly relying on relatively small structural fragments deduced from, for example, NMR DEPT data (list of CH_n groups) or from infra red (IR) spectroscopy (functional groups). Using small-sized substructural fragments or just the skeletal atoms derived from a molecular formula, deterministic generators quickly reach a computational limit with respect to the number of fragments or skeletal atoms, for which they can still generate all possible combinations to form chemically correct molecular structures. This phenomenon is often called a combinatorial explosion. As has been shown above, deducing large fragments from spectral databases prior to the combinatorial step is one solution to overcoming this problem. Another one is to use non-exhaustive algorithms which have shown to tackle large search spaces like the space of constitutional isomers of a compound of, say, 60 skeletal atoms. Typical representatives of this algorithm class are Genetic or Evolutionary Algorithms (GA or EA) or Simulated Annealing (SA).

In recent years, some groups attempted to use structure generators based on these stochastic algorithms in order to overcome the above mentioned problem. Steinbeck and coworkers presented a CASE system, SENECA, in which they implemented structure generators both based on Simulated

Annealing[20] as well as on an Evolutionary Algorithm.[21] The system requires the knowledge of a molecular formula and a set of NMR spectral data (^{13}C NMR, DEPT, HSQC, HMBC, HH COSY) in order to perform its operation. It starts with a randomly generated candidate structure (or a whole population of those in the case of EA), *i.e.* a randomly generated isomer of the given gross formula. It then mutates this candidate (or the members of the population) over multiple generations, while monitoring the improvement with a fitness function. This fitness function is composed of various contributions, each accounting for the fitness of a candidate with respect to a particular kind of spectroscopic constraint, like 2D HH-COSY or HMBC or 1D carbon NMR spectra. The ultimate goal of each structure elucidation run is to maximize the fitness of the candidate structure(s). Fitness is measured in two ways: a) the comparison between calculated and experimental spectra – a measure used in the case of 1D carbon spectra –, or b) a structure-spectrum compatibility test – as used for the 2D NMR spectra. For the latter, a signal from the HMBC spectrum, for instance, is taken and it is investigated whether this signal can be found to be in agreement with the given structure, *i.e.* whether the two atoms involved are separated by an acceptable path length (for the HMBC two or three bonds between a given C–H pair, with a lower occurrence also longer paths). It must be noted, that this scheme allows for easy accounting for rarer spectral features, like the aforementioned long-range couplings in HMBC by simply assigning higher scores to common interpretations and lower scores to rare interpretations. A structure, where all HMBC cross signals can be established by two or three bond couplings will have a higher score than those where some cross signals can only be explained with four bond paths.

The SENECA system has not been tested with a large suite of CASE problems. Rather, a small set of terpenes with growing molecular size has been chosen to evaluate the growth of computation time with a growing number of skeletal atoms. The result is promising – calculation times scale much smoother than with a combinatorial deterministic algorithm – but the system has yet to prove its practical applicability with test cases as large as those tackled by, for example, the *StrucEluc* system.

Another attempt to use stochastic algorithms for CASE was made by Meiler and collaborators. The goal of this project, called GENIUS,[28] was to find the chemical structure with the least amount of data for which structure elucidation success can reasonably be assumed, which usually comes down to the carbon-13 NMR spectrum.[29] Meiler *et al.* used a genetic algorithm and the agreement between predicted and experimental spectra as a fitness function. For this, a neural network was trained, based on the SpecInfo database, to perform fast carbon-13 shift predictions.[30] Prediction speed is of enormous importance here, because the GA has to search a large space of potential candidates, for each of them assessing the fitness by the above procedure. The GENIUS program was able to elucidate the structure of compounds with up to 20 skeleton atoms. Above this limit, the approach taken here is likely to fail, no matter how optimized the program code or how fast the available desktop PC may be, simply due to the exponential growth of the search space and the lack of information further reducing it (like long-range 2D NMR correlations).

The neural network shift prediction developed by Meiler has also been promoted as a post-pocessing tool for COCON, to rank its output structure according to the fit between predicted and experimental spectra. This has been discussed above.

5 Validation

If the user faces the lucky situation of being presented with a top-scoring solution to his or her CASE problem, there might still be the desire for an independent validation of this result. One has to keep in mind that all CASE systems presented above rely on databases for ranking, which endangers the user to encounter false-positive results in the case of structural features or molecular skeletons not properly represented by the underlying material. A validation using first principles would thus be advantageous.

The enormous increase in computation speed of today's desktop computers, paired with algorithmic improvements in quantum chemical programs, have made *ab initio* computations of medium-sized organic molecules[31] accessible to the regular bench chemist with moderate computation times. Within hours or a few days, depending on the type and size of the computation, a high level geometry optimization, followed by a chemical shift calculation, can today be performed on commodity-type computers.[32] This method was recently evaluated by Barone and coworkers for its potential to validate structure proposals gained from either a manual or an automated elucidation process.[33] The group performed GIAO NMR calculations on a number of natural products. It could be shown that in most cases the calculated values were in excellent agreement with the experimental data, with *R* values for least squares fits greater than 0.995. Polar compounds are problematic, however, because here chemical shifts are influenced by the polar solvent in which these compounds are necessarily dissolved. For a large number of carbon skeletons with a relatively low number of polarizing groups, the calculation of *ab initio* chemical shifts presents an excellent way, superior to all knowledge based methods for shift prediction,[34] for an independent validation of structural proposals from CASE programs or manual elucidation.

6 Conclusion

A refreshing number of new and ambitious projects and programs for computer assisted structure elucidation have entered the field in the past five years. Most notably, the first commercial system with general applicability has been introduced. *ACDLabs Structure Elucidator* combines both flexible algorithms for *ab initio* CASE as well as a large database for a fast dereplication procedure. The lucky few who can afford it will certainly be happy with it. For the rest, a number of free or cheaper academic systems, COCON, HOUDINI, SENECA and others, are available, which will, however, have a steeper learning curve.

An open NMR database with organic structures and assigned spectra has been instantiated, which is called NMRShiftDB and is available to the public *via* http://www.nmrshiftdb.org.

Despite the promising developments presented in this article, however, the well-working, ubiquitously used CASE application has yet to arrive.

7 References

1 B. G. Buchanan and E. A. Feigenbaum, *Artif. Intell.*, 1978, **11**, 5.
2 M. Jaspars, *Nat. Prod. Rep.*, 1999, **16**, 241.
3 M. E. Munk, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 997.
4 C. Steinbeck, 'Computer-Assisted Structure Elucidation', in *Handbook of Chemoinformatics*, ed. J. Gasteiger, Wiley-VCH, Weinheim, 2003.
5 R. Neudert and M. Penk, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 244.
6 V. Schutz, V. Purtuc, S. Felsinger and W. Robien, *Fresenius' J. Anal. Chem.*, 1997, **359**, 33.
7 ACD/CNMR DB Add-on, ACDLabs, Inc., Toronto, 2002.
8 C. Steinbeck, S. Kuhn and S. Krause, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1733.
9 B. G. Buchanan, D. H. Smith, W. C. White, R. J. Gritter, E. A. Feigenbaum, J. Lederberg and C. Djerassi, *J. Am. Chem. Soc.*, 1976, **98**, 6168.
10 M. Badertscher, A. Korytko, K. P. Schulz, M. Madison, M. E. Munk, P. Portmann, M. Junghans, P. Fontana and E. Pretsch, *Chemom. Intell. Lab. Syst.*, 2000, **51**, 73.
11 A. Korytko, K. P. Schulz, M. S. Madison and M. E. Munk, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1434.

12 K. P. Schulz, A. Korytko and M. E. Munk, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1447.
13 T. Lindel, J. Junker and M. Köck, *J. Mol. Model.*, 1997, **3**, 364.
14 B. Reif, M. Köck, R. Kerssebaum, H. Kang, W. Fenical and C. Griesinger, *J. Magn. Reson., Ser. A*, 1996, **118**, 282.
15 M. Köck, B. Reif, W. Fenical and C. Griesinger, *Tetrahedron Lett.*, 1996, **37**, 363.
16 M. Köck, J. Junker and T. Lindel, *Org. Lett.*, 1999, **1**, 2041.
17 K. A. Blinov, M. E. Elyashberg, S. G. Molodtsov, A. J. Williams and E. R. Martirosian, *Fresenius' J. Anal. Chem.*, 2001, **369**, 709.
18 M. E. Elyashberg, K. A. Blinov, A. J. Williams, E. R. Martirosian and S. G. Molodtsov, *J. Nat. Prod.*, 2002, **65**, 693.
19 C. Steinbeck, *Angew. Chem., Int. Ed.*, 1996, **35**, 1984.
20 C. Steinbeck, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1500.
21 Y. Han and C. Steinbeck, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 489.
22 M. Will, W. Fachinger and J. R. Richert, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 221.
23 M. E. Elyashberg, K. A. Blinov, A. J. Williams, S. G. Molodtsov, G. E. Martin and E. R. Martirosian, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 771.
24 G. E. Martin, C. E. Hadden, D. J. Russell, B. D. Kaluzny, J. E. Guido, W. K. Duholke, B. A. Stiemsma, T. J. Thamann, R. C. Crouch, K. Blinov, M. Elyashberg, E. R. Martirosian, S. G. Molodtsov, A. J. Williams and P. L. Schiff, *J. Heterocycl. Chem.*, 2002, **39**, 1241.
25 K. A. Blinov, D. Carlson, M. E. Elyashberg, G. E. Martin, E. R. Martirosian, S. Molodtsov and A. J. Williams, *Magn. Reson. Chem.*, 2003, **41**, 359.
26 K. Funatsu and S.-i. Sasaki, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 190.
27 J. L. Faulon, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1204.
28 J. Meiler and M. Will, *J. Am. Chem. Soc.*, 2002, **124**, 1868.
29 J. Meiler and M. Will, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1535.
30 J. Meiler, R. Meusinger and M. Will, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1169.
31 J. C. Facelli, *Concepts Magn. Reson.*, 2004, **20A**, 42.
32 M. C. Nicklaus, R. W. Williams, B. Bienfait, E. S. Billings and M. Hodoscek, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 893.
33 G. Barone, L. Gomez-Paloma, D. Duca, A. Silvestri, R. Riccio and G. Bifulco, *Chem. Eur. J.*, 2002, **8**, 3233.
34 C. Steinbeck, in 'Correlations between Chemical Structures and NMR Data', in *Handbook of Chemoinformatics*, ed. J. Gasteiger, Wiley-VCH, Weinheim, 2003.