Available online at www.sciencedirect.com
SCIENCE DIRECT®
Phytochemistry 65 (2004) 2711–2717
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

NMRShiftDB – compound identification and structure elucidation support through a free community-built web database

Christoph Steinbeck *, Stefan Kuhn

Max-Planck-Institute of Chemical Ecology, Hans-Knöll-Str. 8, D-07745 Jena, Germany
Cologne University Bioinformatics Center, Zülpicher Str. 47, D-50674, Köln, Germany

Received 5 April 2004; received in revised form 2 August 2004
Available online 22 September 2004

Abstract

Compound identification and support for computer-assisted structure elucidation via a free community-built web database for organic structures and their NMR data is described. The new database NMRShiftDB is available on http://www.nmrshiftdb.org. As the first NMR database, NMRShiftDB allows not only open access to the database but also open and peer reviewed submission of datasets, enabling the natural products community to build its first free repository of assigned ^{1}H and ^{13}C NMR spectra. In addition to the open access, the underlying database software is built solely from free software and is available under an open source license. This allows collaborating laboratories to fully replicate the database and to create a highly available network of NMRShiftDB mirrors. The database contains about 10,000 structures and assigned spectra, with new datasets constantly added. Its functionality includes (sub-) spectra and (sub-) structure searches as well as shift prediction of ^{13}C spectra based on the current database material.
© 2004 Elsevier Ltd. All rights reserved.

Keywords: Computer-assisted structure elucidation; NMR; Structure databases

1. Introduction

Computational tools for computer-assisted structure elucidation (CASE) have increasingly gained access into the work of organic chemists trying to characterize their products and byproducts, as well as of natural products chemists elucidating the structure of a newly isolated compound. In the CASE process (Steinbeck, 2001a), before reaching for one of the more sophisticated but also slow and error-prone ab-initio tools (Munk, 1998; Steinbeck, 2001b), like structure generator based methods, it is advisable to perform a database search using some kind of spectral fingerprint in order to avoid surplus work on data already reported in the literature. This process of identifying known compounds from their physicochemical data is known as dereplication. The 1-dimensional broad-band-decoupled ^{13}C NMR spectrum has proved to be immensely useful for this task, largely because it is not complicated by scalar (through-bond) couplings and can thus be easily converted to a simple list of signal positions. Therefore, databases have been created, in which organic chemical structures are stored together with a subset of their NMR data commonly used for structure identification purposes (for example the one-dimensional ^{1}H and ^{13}C NMR spectra) and an assignment of every atom in the molecule to a signal in the spectrum (Neudert and Penk, 1996; Schutz et al., 1997; Elyashberg et al., 2004). With these databases in hand, a scientist would – upon isolation – first record a one-dimensional proton and carbon NMR spectrum and then search the database using a list

* Corresponding author. Tel.: +49 221 470 7426; fax: +49 221 470 7786.
E-mail address: c.steinbeck@uni-koeln.de (C. Steinbeck).

0031-9422/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2004.08.027

of NMR chemical shifts. As a result, he would be presented with a list of the most similar spectra and, what is most important, with their assigned chemical structures. Our own experience from natural product structure elucidation shows that even in the case of no direct hit, the spectral similarity search has in many instances the potential of pointing the scientist towards the correct class of compounds, just leaving the spectroscopist with the task of sorting out differences in the substitution pattern, for example. During our work with spectroscopic databases we have often wondered why there was no analog of open bioinformatics databases like the Protein Data Bank (PDB) or the large genomic databases for a field as old as the application of nuclear magnetic resonance data in chemistry. The success of the above open-submission databases is supported by journals requiring deposition of data upon submission of a paper in which they are discussed. Thus, inspired by the aspects and ideas discussed above, we decided to exploit the open-source principles (Raymond, 1999) to create an open-access, open-submission, open-source database for organic chemical structures and their NMR data, which we call NMRShiftDB. In order to achieve our goal of broadest possible accessibility, the database has a web interface and can be reached via the URL http://www.nmrshiftdb.org.

2. Results and discussion

When this article went into print, the NMRShiftDB database contained data for about 10,000 chemical structures and their assigned spectra, which are mostly 1D ^{13}C NMR spectra. These spectra where both taken from the latest issues of Phytochemistry and the Journal of Natural Products, as well contributed from in-house databases by third parties.

It is cordially hoped that with the publication of this article, the scientific user community will start contributing to NMRShiftDB, which will be faithfully supported by our group. Since the database schema is prepared for the input of spectra for any type of nucleus, the respective functionality can be enabled on request.

NMRShiftDB is a great chance to revive collections of in-house NMR data by integrating them with a larger data basis, thus enhancing its usefulness and reliability in shift prediction. With this publication, we would like to motivate maintainers of such collections to contribute them to NMRShiftDB. The database is an ideal candidate for such an endeavor, because the valuable data entered by authors will clearly never be endangered to vanish behind the walls of expensive proprietary databases. NMRShiftDB's open-source and open-content license, as well as the network of database mirrors already maintained by various work groups, ensures constant availability and prevents misuse.

In the future, we will focus on development in the following areas: development of a stereochemically enhanced HOSE code representation in order to provide better predictions; improvement of import and export facilities for structural and spectroscopic data and means to export datasets in various journal-compliant formats in order to motivate publishers to require the deposition of NMR datasets published in their journals. In a separate project – based on NMRShiftDB – we are working on a more general natural product information system, extensible to hold information on any desired property for a stored chemical structure.

Last but not least it should be noted that the NMRShiftDB project is also a valuable tool for teaching and learning the use of structure–property databases in chemistry. Contributions and help from, as well as collaboration with, other individuals and groups will be greatly appreciated

2.1. A case study: using NMRShiftDB in a CASE context

Spectral databases are indispensable tools when efficiently embedded into a unified CASE process (Steinbeck, 2003). Fig. 1 shows a schematic embedding of CASE systems based on both deterministic and stochastic structure generators combined with a spectroscopic database, which can serve firstly as a dereplication tool and secondly can provide starting structures for a stochastic CASE process.

In this section, we will demonstrate the use of NMRShiftDB in this context with two examples.

2.1.1. A dereplication

As an example from our own practice, we chose the structure identification of a chromene from the brasilian plant *Calea serrata* (Fig. 2).

About 10 mg of the compound were isolated and subjected to a routine ^{13}C NMR, acquired in a few minutes. After peak picking using the standard software supplied by the instrument vendor the result was copied and pasted into NMRShiftDB's spectrum search field. Performing an identity search yielded a dataset with more than 99% spectral similarity, which clearly indicates a direct hit.

2.1.2. A search for a compound not present in the database

The substance serving as the next example was isolated from the indian plant *Alstonia scholaris*. Again, a routine ^{13}C NMR was measured and an identity search was made. (^{13}C NMR ($CDCl_3$) 190.98, 171.68, 156.10, 146.35, 138.15, 128.19, 123.47, 125.61, 121.02, 119.93, 58.40, 55.46, 55.15, 53.68, 51.98, 51.55, 36.13, 33.62, 32.50, 12.97).

The ranked list of structures with most similar spectra shows the alkaloid rhazimal at position 1 (Fig. 3).

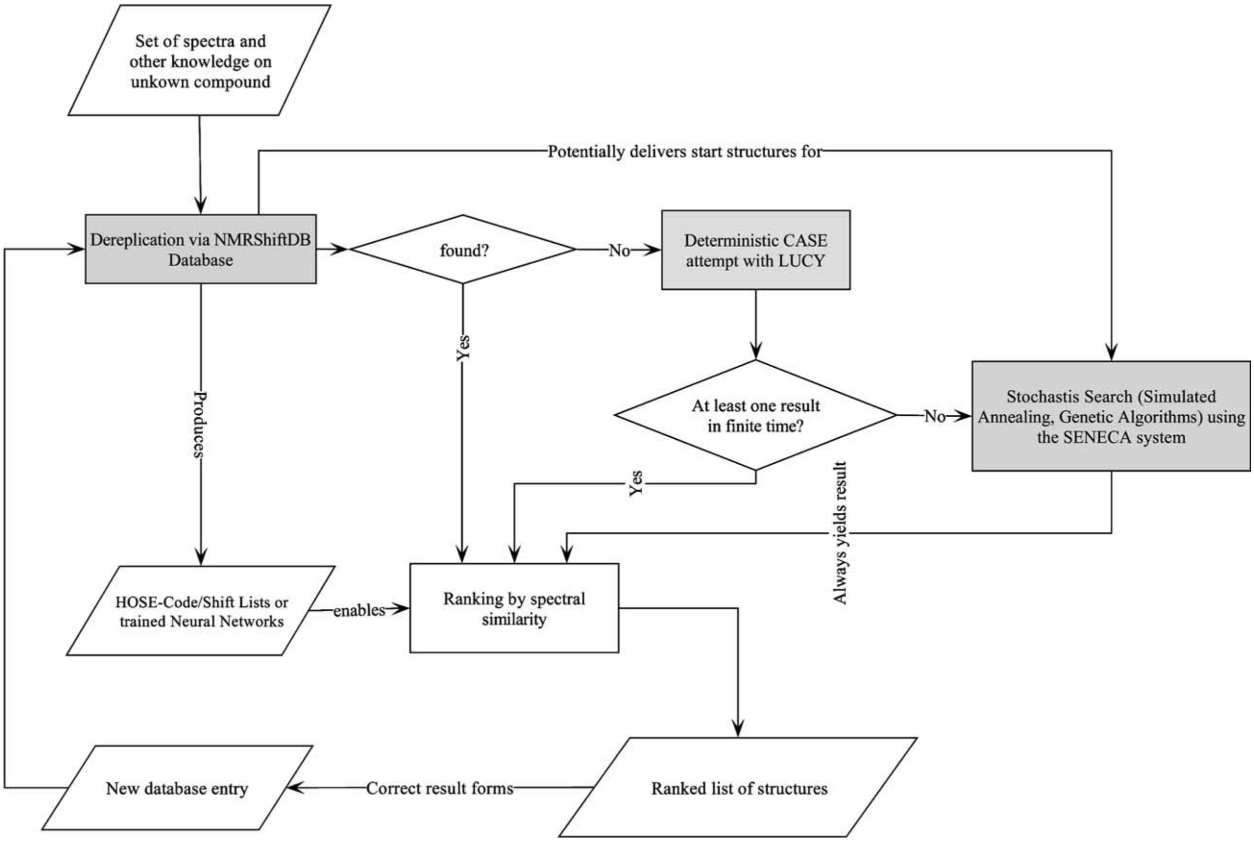

Set of spectra and other knowledge on unkown compound
Potentially delivers start structures for
Dereplication via NMRShiftDB Database
found?
No
Deterministic CASE attempt with LUCY
Yes
Produces
At least one result in finite time?
No
Stochastis Search (Simulated Annealing, Genetic Algorithms) using the SENECA system
Yes
Always yields result
HOSE-Code/Shift Lists or trained Neural Networks
enables
Ranking by spectral similarity
New database entry
Correct result forms
Ranked list of structures

Fig. 1. A unified CASE process. Figure adapted from Steinbeck (2003).

Upon selection of a detailed view of the top result, the user is presented with a best fit of query signal positions to hit signal positions (Figs. 4 and 5).

Together with the given spectral assignment, this allows the user to easily find those positions where the structural hit differs from the unknown. In our case, chemical shift positions with no fit at all comprise $\delta_C = 67.80$ and 194.61 ppm, in addition to which we further find deviations of more than 1 ppm for $\delta_C = 37.73$, 124.56 and 143.3 ppm. Further, no close fit was found for query spectrum shift at $\delta_C = 55.46$ ppm. This draws our attention to atoms 9, 14, 18 and 21 in the hit structure (see Fig. 4).

Carbon spectrum multiplicities from DEPT experiments reveal that the unassigned query carbon atom at $\delta_C = 55.46$ ppm is a CH group. Since the aldehyde carbon in rhazimal is not present in the unknown, we can replace the surplus quarternary carbon at $\delta_C = 67.8$ ppm (atom 18) with this CH group at $\delta_C = 55.46$ ppm, suggesting the compound to be having the structure shown in Fig. 6. A literature search revealed that this compound had been isolated before by Atta-ur-Rahman and Habib-ur-Rehman (1986) and that its reported spectral properties are in agreement with our measured data. This simple exercise can be repeated with similar success based on subsequent hits with lower spectral similarity shown in Fig. 3.

3. Experimental

NMRShiftDB features:

- storage of 2D and 3D chemical structures together with their 1D NMR data,
- linking from structures and spectra to other databases,
- searches for molecules and spectra, as well as in text fields,
- submission of structures and assigned spectra by registered users,
- reviewing of submitted data by registered reviewers,
- a user administration.

The database is compound-centric, i.e., each dataset is grouped around a chemical structure with the following additional data: Unique identifier (NMRShiftDB ID), canonical representation (canonical SMILES), canonical name (generated by ACD/name™ or

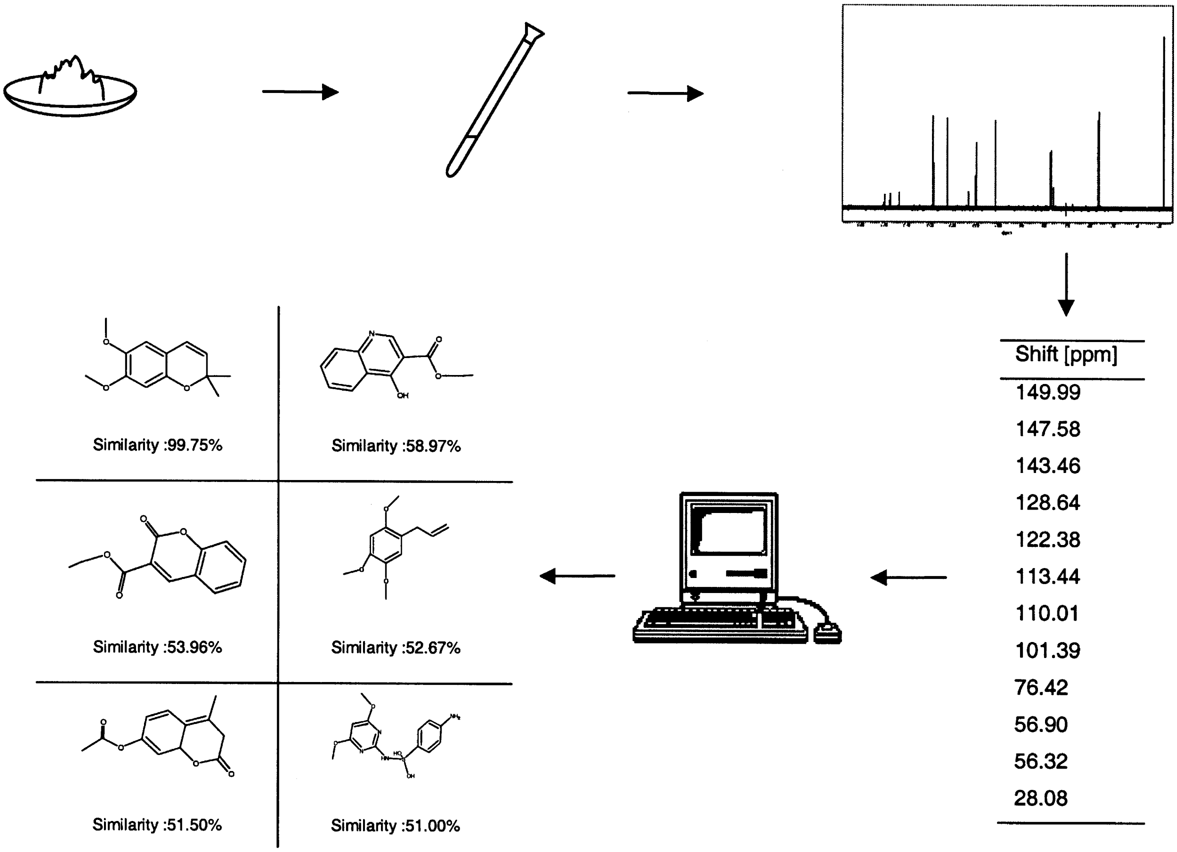


Fig. 2. Dereplication procedure for a chromene isolated from *Calea serrata* (Steinbeck, 2003).

Autonom™), one or more chemical names, an arbitrary number of spectra currently limited to 1D ^{1}H, ^{13}C, ^{15}N and ^{31}P spectra, [1] and their measurement conditions (temperature, spectrometer frequency, solvent and assignment method), compound keywords, spectrum keywords and web links (URLs) to other databases (Fig. 7). Details on the technical realization of NMRShiftDB have been published elsewhere (Steinbeck et al., 2003).

The result of any search is a list of no, one or more datasets fitting the query. Within this list, the dataset is identified by a chemical structure drawing and a chemical name. Eventually, the user will select a dataset from that list and display the entirety of its properties. Besides the pure display on the screen, each dataset can be downloaded as portable document format (PDF) or rich text format (RTF), the latter allowing for easy import into word processor documents. In the following, we describe a variety of possible search strategies leading to potential hit lists.

Chemical Structure Searches can be either performed as identity searches, revealing one or no exact hit, or as substructure searches, yielding a list of structures with a given common substructure, or as similarity searches, yielding a list of structures with similar structural features as a given template structure.

Spectrum Searches are initiated by entering a list of NMR chemical shifts. Based on this list, the collection is searched for similar spectra either using a subspectrum search or a full spectrum search. While the latter yields high similarity values only in cases where the hit spectrum contains exactly the number of signals and signal positions of the query spectrum, the first version tries to find a best match for each of the query spectrum signals and ignores surplus signals in the target spectrum. Spectrum searches can be limited to certain solvents. A more sophisticated limitation to arbitrary combinations of solvents will be added in a future release.

A *Text Search* facility allows for querying the keyword, chemical name and comment fields. Maximum flexibility is given by providing the user with three different search techniques: an exact search yielding a dataset containing the query string exactly as given by the user, a regular expression search for almost unlimited wildcard queries or a (modified) soundex search to locate strings that sound (in the English language) like the query string, thus leading to a certain degree of error tolerance.

[1] The database schema is capable of storing spectral data for any kind of nucleus. The respective feature can be enabled on request upon short notice.

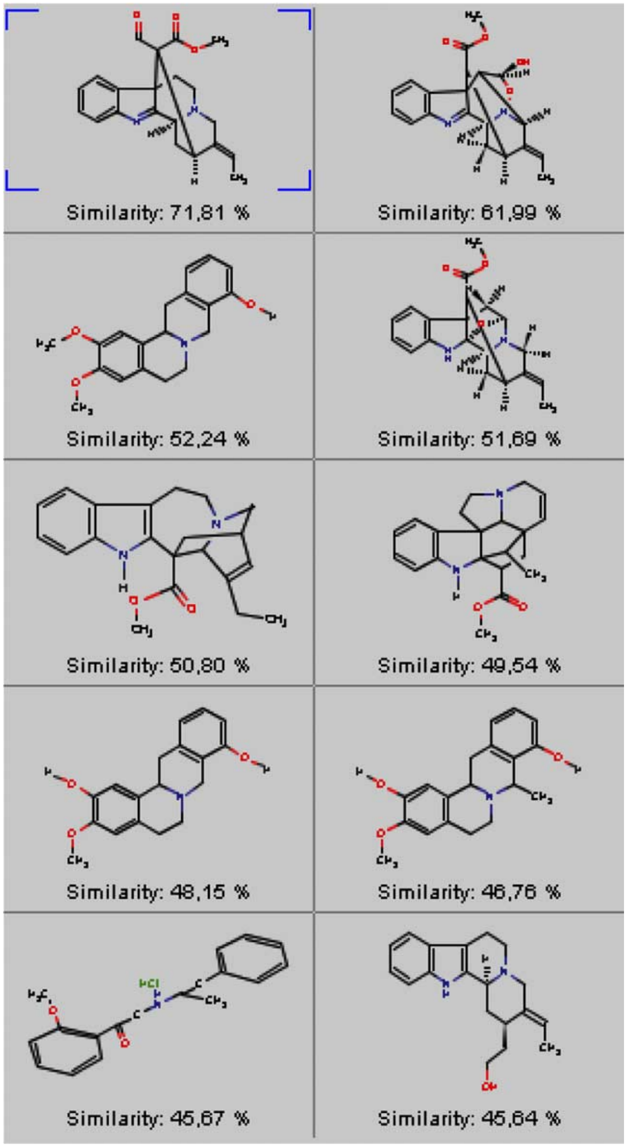


Fig. 3. Screen copy of the result of a spectrum search in NMRShiftDB ranked by spectral similarity.

In NMRShiftDB *Shift Prediction* is performed based on the hierarchically ordered spherical environment (HOSE) codes (Bremser, 1978, 1985). To enable HOSE code predictions for carbon atoms, a canonical linear string description for the spherical environment of every carbon atom in every structure in the database is build, typically upto the size of six spheres. The resulting HOSE codes are stored in look-up tables together with their assigned chemical shift values. This potentially time consuming procedure is done only rarely and when a new structure is added to the database. To perform a shift prediction for a structure not present in the database, HOSE codes are generated for it and searched in the previously generated table. For each of the codes composed for the unknown structure, one will find a number of hits with differing shift values in the look-

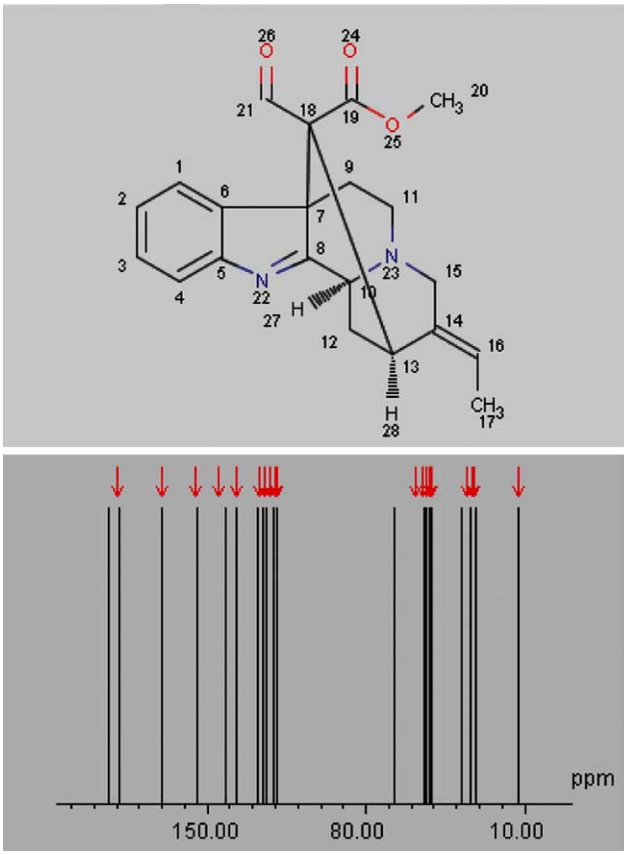
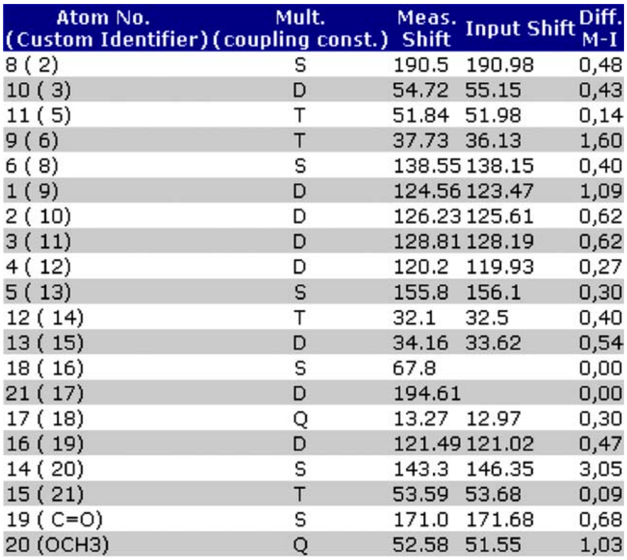


Fig. 4. Graphical display of the most similar spectrum found by a spectrum similarity search together with its assigned structure. Red arrows graphically indicate the position of the query shifts. The associated tabular view is shown in Fig. 5. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



Atom No. (Custom Identifier)	Mult. (coupling const.)	Meas. Shift	Input Shift	Diff. M-I
8 (2)	S	190.5	190.98	0,48
10 (3)	D	54.72	55.15	0,43
11 (5)	T	51.84	51.98	0,14
9 (6)	T	37.73	36.13	1,60
6 (8)	S	138.55	138.15	0,40
1 (9)	D	124.56	123.47	1,09
2 (10)	D	126.23	125.61	0,62
3 (11)	D	128.81	128.19	0,62
4 (12)	D	120.2	119.93	0,27
5 (13)	S	155.8	156.1	0,30
12 (14)	T	32.1	32.5	0,40
13 (15)	D	34.16	33.62	0,54
18 (16)	S	67.8		0,00
21 (17)	D	194.61		0,00
17 (18)	Q	13.27	12.97	0,30
16 (19)	D	121.49	121.02	0,47
14 (20)	S	143.3	146.35	3,05
15 (21)	T	53.59	53.68	0,09
19 (C=O)	S	171.0	171.68	0,68
20 (OCH3)	Q	52.58	51.55	1,03

Fig. 5. Closest fit of a query to a search spectrum with deviations, as graphically depicted in Fig. 4.

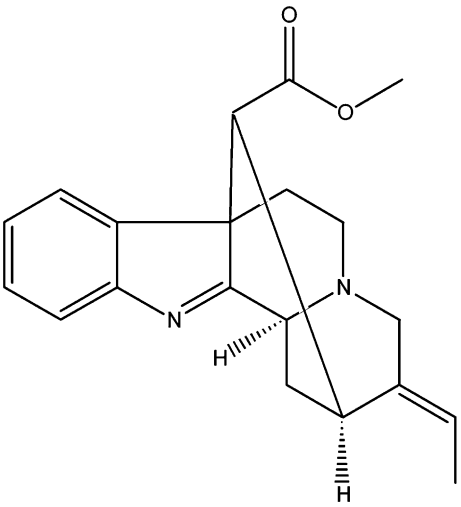


Fig. 6. Structure of candidate compound strictamine.

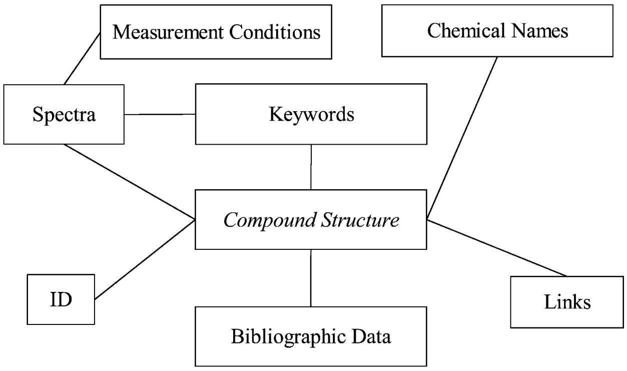


Fig. 7. NMRShiftDB's dataset architecture.

up table. The result is reported as the mean value of the stored database entries, including some statistics like the standard deviation and the highest and lowest shift value for the given HOSE code. If no match is found for a six-sphere query code, the number of spheres is gradually lowered until a hit is found. The quality of such a prediction depends on the number of spheres used for the HOSE code, on the number of hits found for each of the query codes, and of course on the structural diversity of the database. Clearly, a growth in the number of datasets will increase the quality of the shift prediction.

The possibility for dataset submission by the public is the feature distinguishing NMRShiftDB from potentially similar systems. This of course calls for a tight quality management, so that submission of datasets is only possible for registered users upon log-in. During a short one-time *registration procedure* the user enters name, address, affiliation and email address, so that he or she can be contacted in case of problems with a submitted dataset. *Dataset submission* is done via a step-by-step procedure using the regular NMRShiftDB web interface. It comprises: (a) entering or loading the chemical structure, (b) entering or loading a list of NMR chemical shifts with optional intensities, (c) performing an assignment of the shifts entered in step 2 and the atoms of the structure entered in step 1, and (d) entering a number of additional data, like literature reference, measurement conditions and keywords.

Both the assignment and the quality control are semi-automatically supported by applying NMRShiftDB's shift prediction capabilities and presenting the result to reviewers or users as predicted shifts with expectation ranges for the newly entered structure.

Upon submission of a dataset by a contributor, the *review process* is initiated. A notification is sent to an automatically selected reviewer by email, which contains a convenience URL, leading the reviewer to a web page on NMRShiftDB, summarizing the submitted dataset and providing the reviewer with same shift prediction statistics that have already guided the submitter. Since each dataset includes the information on the submitter's identity, the reviewer can easily contact him and clarify questions. Clearly, not anonymity but maximum quality is the more important issue here. Upon inspection, a reviewer has four options: (a) *accept the dataset without changes*, in which case it is marked as reviewed and will become visible in the database, (b) *edit and accept* is the method of choice in case of small errors. Here, the reviewer applies the correction and the dataset is marked as valid, (c) *submit back to the contributor* in case of unclear assignments, unlikely chemical shift values and other larger problems, accompanied with a friendly email summarizing the problems and requesting the contributor to contribute to their resolution, (d) *reject the dataset*. This would be an option in case of serious problems with the dataset which cannot be corrected by procedures (b) and (c). In this case the reviewer is required to explain this measure in an email to the submitter. This option will almost solely be chosen in case of datasets obviously submitted to compromise the database with bad data.

Acknowledgements

The authors thank Dr. Willy von der Lieth, Dr. Wilhelm Hull (German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ)) and Dr. Heinz Kolshorn (University of Mainz, Germany) for generously contributing datasets from in-house databases. We are also grateful to the Max-Planck-Institute of Chemical Ecology, Jena, Germany, for hosting our main server hardware, for providing the technical support, and for constantly adding data to the database. Last not least our gratitude goes to the German Research Council (Deutsche Forschungsgemeinschaft, DFG) for financial support of the NMRShiftDB project.

References

Atta-ur-Rahman, Habib-ur-Rehman, 1986. Isolation and NMR-studies on rhazimal and strictamine. Planta Medica 3, 230–231.

Bremser, W., 1978. HOSE – a novel substructure code. Analytica Chimica Acta 103, 355–365.

Bremser, W., 1985. Expectation ranges of ^{13}C NMR chemical shifts. Magnetic Resonance in Chemistry 23, 271–275.

Elyashberg, M.E., Blinov, K.A., Williams, A.J., Molodtsov, S.G., Martin, G.E., Martirosian, E.R., 2004. Structure elucidator: a versatile expert system for molecular structure elucidation from 1D and 2D NMR data and molecular fragments. Journal of Chemical Information and Computer Sciences 44, 771–792.

Munk, M.E., 1998. Computer-based structure determination: then and now. Journal of Chemical Information and Computer Sciences 38, 997–1009.

Neudert, R., Penk, M., 1996. Enhanced structure elucidation. Journal of Chemical Information and Computer Sciences 36, 244–248.

Raymond, E.S., 1999. The Cathedral and the Bazaar, first ed. O'Reilly and Associates, Sebastopol, CA.

Schutz, V., Purtuc, V., Felsinger, S., Robien, W., 1997. Csearch-stereo – a new generation of NMR database systems allowing three-dimensional spectrum prediction. Fresenius Journal of Analytical Chemistry 359, 33–41.

Steinbeck, C., 2001a. The automation of natural product structure elucidation. Current Opinion in Drug Discovery and Development 4, 338–342.

Steinbeck, C., 2001b. Seneca: a platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. Journal of Chemical Information and Computer Sciences 41, 1500–1507.

Steinbeck, C., 2003. Computer-assisted structure elucidation. In: Gasteiger, J. (Ed.), Handbook on Chemoinformatics, vol. 2. Wiley–VCH, Weinheim, pp. 1378–1406.

Steinbeck, C., Kuhn, S., Krause, S., 2003. NMRShiftDB – constructing a chemical information system with open source components. Journal of Chemical Information and Computer Sciences 43, 1733–1739.