MAGNETIC RESONANCE IN CHEMISTRY
Magn. Reson. Chem. 2004; **42**: 000–000
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/mrc.1424

Novel methods of automated structure elucidation based on ^{13}C NMR spectroscopy

Jens Meiler[1] and Matthias Köck[2]*

[1] University of Washington, Box 357350, Seattle, Washington 98195, USA
[2] Alfred-Wegener-Institut für Polar- und Meeresforschung, Am Handelshafen 12, D-27570 Bremerhaven, Germany

Received 20 February 2004; Accepted 16 April 2004

Three new approaches for automated structure elucidations of organic molecules using NMR spectroscopic data were introduced recently. These approaches apply a neural network ^{13}C NMR chemical shift prediction method to rank the results of structure generators by the agreement of the predicted and experimental chemical shifts. These three existing implementations are compared using realistic molecules. The applicability and reliability of such approaches is addressed. Copyright © 2004 John Wiley & Sons, Ltd.

KEYWORDS: NMR; ^{13}C NMR; artificial neural network; automated structure elucidation; chemical shift calculation

INTRODUCTION

The structures of natural products and unknown compounds obtained from organic synthesis are usually elucidated by applying various spectroscopic techniques, such as IR, MS, UV and NMR methods. After the molecular formula has been determined (e.g. from a high-resolution mass spectrum), NMR spectroscopy assumes special importance, as it is the only method out of the four which achieves atomic resolution.

One of the simplest NMR parameters is concurrently one of the most useful: the ^{13}C chemical shift describes with only one number the complex chemical environment of a carbon atom. The chemical shift values for all carbon atoms of an organic compound can be determined easily and yield a characteristic fingerprint of an unknown compound. This fingerprint is unique and theoretically sufficient to elucidate the structure of molecules with 15–20 non-hydrogen atoms, even if an experimental uncertainty of 0.5 ppm is assumed. For a higher number of non-hydrogen atoms, the number of possible constitutions increases faster than the number of possible different ^{13}C NMR spectra (within the experimental uncertainty). Consequently, two substances can yield quasi-identical spectra.

The simplicity of ^{13}C NMR chemical shift experiments paired with the enormous information they provide about the constitutional environment of a carbon atom makes them widely used in structure elucidation. The structural environment of a carbon atom is often represented as a single string of characters, *h*ierarchically *o*rdered *s*pherical description of *e*nvironment (HOSE code).[1] The HOSE codes of many carbon atom environments were stored together with the corresponding chemical shift values in databases.[2–4]

*Correspondence to: Matthias Köck, Alfred-Wegener-Institut für Polar- und Meeresforschung, Am Handelshafen 12, D-27570 Bremerhaven, Germany. E-mail: mkoeck@awi-bremerhaven.de

Using such databases, very accurate predictions of ^{13}C NMR chemical shifts become possible by generating the HOSE code for the carbon atoms of interest and screening the database for similar ones. On the basis of these data, mathematical models were developed that generalize the dependence of the chemical shift value from the molecular constitution. Such models help in understanding the nature of this correlation and can also be applied to predict chemical shifts. The chemical shift calculation using a database has the disadvantage of relatively long search times in the databases and the necessity for access to such large storage systems. To avoid this time-expensive approach, various incremental systems were introduced, which usually rely on multiple linear regression.[5–7] A modern version of such an implementation is used in ChemDraw.[8] Although increment methods usually give less accurate results than database predictions, they are widely used owing to their availability, simplicity and velocity.

In the last 10 years, neural networks[9] were introduced into the field of chemical shift prediction. After being applied to specific groups of organic compounds,[10–13] generic neural networks were introduced that predict chemical shifts for nearly every class of organic molecule.[3,14,15] More recently, they were also applied to protein chemical shift prediction.[16] Their advantage is that they combine the accuracy of database predictions without losing much of the speed of incremental methods. Therefore, they are suitable to be applied to large sets of molecules obtained from structure generators. The details of the neural network used in the following implementations are described elsewhere.[17] The accuracy of the prediction is as good as 1.5 ppm by computing up to 5000 chemical shifts per second.

EXPERIMENTAL

We describe the application of the three structure generators: Molgen, Genius and Cocon (see Table 1). The programs

Table 1. Molecular formulae, computational aspects and results obtained for the model compounds **1**–**3**

AQ1

Structure generator	Compound	Molecular formula	Computational data: No. of possible structures	Computational data: No. of generated structures	Computational data: Overall time[a] (s)	R.m.s.d. (experiment—ANN): Correct (ppm)	R.m.s.d. (experiment—ANN): Best (ppm)	R.m.s.d. (experiment—ANN): Worst (ppm)
MOLGEN	*N*-Allyl-*N*′-ethylthiovren (**1**)	$C_6H_{10}N_2S$	709 259	709 259	8213	0.7	0.7	93.5
GENIUS	Tryptophan (**2**)	$C_{11}H_{12}N_2O_2$	$\sim 6.6 \times 10^{10}$	10 752	400	1.4	1.4	58.7
COCON	Prianosin D acetate (**3**)	$C_{22}H_{19}N_3O_4S$	?	22 572	200	5.9	4.6	23.0

[a] Calculated on a PC equipped with a Pentium III processor (1000 MHz). Calculation time of COCON is not included for **3**.

GENIUS and COCON rely on the direct use of spectroscopic data (see below). MOLGEN is a powerful structure generator that computes all possible structures for a given molecular formula.[18,19] For molecules with up to 12 non-hydrogen atoms, the algorithm computes all possible constitutions in a reasonably short period of time (from a few seconds up to 24 h). The resulting set of structural proposals contains up to a few million members. The ^{13}C NMR spectrum for all structural proposals is then computed using the artificial neural network-based program C_SHIFT.[14,17] The r.m.s.d. of the computed and the experimental chemical shift values over all carbon atoms serves as a quality factor for the similarity of the two spectra.[20c,21] Also, constitutions that yield a r.m.s.d. value below the experimental deviation plus the standard deviation of the prediction method are treated as potentially correct constitution of the unknown compound.[22] Increasing the number of non-hydrogen atoms, the constitutional space soon extends the critical size that can be computed using MOLGEN in a meaningful period of time.

At this point, the search algorithm needs to be modified: instead of considering the complete constitutional space, only a subspace is generated, while ensuring that the correct solution is part of this subspace. GENIUS uses a genetic algorithm to generate this structural subspace.[23] For this purpose, the constitution of each molecule is coded in a string. For these strings, mutation (changing the string) and recombination (combining two strings) operators were defined. With this implementation, a set of constitutions can be treated as a population of individuals that undergoes cycles of recombination and mutation under the influence of a continuous selection pressure:[23] the first step is the generation of a small, randomly chosen part of the constitutional space and the comparison of the predicted with the experimental chemical shift values for all members of this subspace. The resulting r.m.s.d. values are applied as a fitness function and serve as selection criteria for the recombination step. The offspring is now generated by recombining two molecules to form a new one and applying eventually a mutation. The resulting new set of molecules again undergoes the processes of selection, recombination and mutation. This algorithm optimizes the members of the population to meet the experimental NMR spectrum and therefore samples the structural subspace of interest. It suggests the correct solution for all tested examples with up to 15 non-hydrogen atoms and for most examples with up to 20 non-hydrogen atoms.[23]

At the size of about 15–20 non-hydrogen atoms, additional experimental information becomes necessary to solve the structure unambiguously. This information can be lists of necessary (good list) or forbidden (bad list) fragments applied in MOLGEN or GENIUS, which can be known from synthesis, experience or additional experimental data.[21] The COCON algorithm[20,24] is specialized to exploit two-dimensional NMR connectivity information, which drastically reduces the size of the structural space spanned by one molecular formula. Thus, the structural space to be generated is restricted to all structures that meet the experimental connectivity information. The ranking within this subspace is again obtained by computing the ^{13}C NMR chemical shift values[20c] with C_SHIFT.[21] COCON also allows including information from other sources as fixed bonds or forbidden bonds. This corresponds to the good list and bad list philosophy.

Figure 1 illustrates the critical influence of the size of the constitutional space on the choice of the applied algorithm. As long as the constitutional space is small enough to be generated completely, MOLGEN is the tool of choice. The generation of the complete set of possible constitutions guarantees that the correct solution is generated and analyzed. However, MOLGEN becomes too slow as the molecular size increases. In contrast, GENIUS generates only a small part of the constitutional space by incorporating the experimental information into the process of structure generation. It is therefore able to find the correct constitution in much larger constitutional spaces. However, since only part of the structural space is evaluated, no guarantee can be given that the correct solution was generated. Increasing the structural space still further, the ^{13}C NMR spectrum alone is no longer a unique molecular fingerprint, if the uncertainties of the experiment and the shift prediction are taken into account. COCON generates all possible constitutions that meet the connectivity information obtained from two dimensional NMR spectra. In contrast to GENIUS but similar to MOLGEN, the COCON algorithm generates the complete subspace of constitutions that are consistent with the connectivity data; hence the correct constitution will be among the generated structural proposals. However, the similarity among the obtained constitutions and the predicted NMR spectra will be much higher. For that reason, more than one of the generated constitutions can satisfy the experimental NMR spectrum with a small r.m.s.d. value. In that case, no unambiguous solution is possible but the list of possible constitutions can be dramatically reduced, typically by a factor of 0.001–0.002.

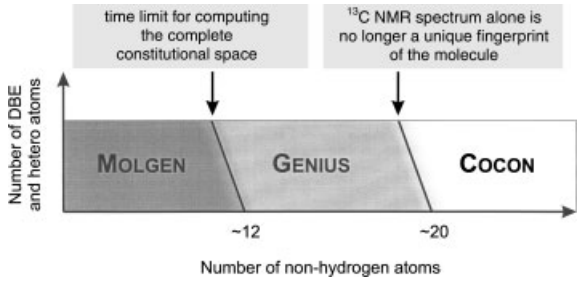

Figure 1. Applicability ranges for the methods of automated structure elucidation. While the total number of constitutions possible for a given molecular formula is smaller than ~10^6, all structures and the respective ^{13}C NMR spectrum may be generated computationally. The structural space is usually sufficiently small to make the agreement of experimental and predicted ^{13}C NMR a unique identifier of the correct constitution. However, if the number of possible constitutions increases above ~10^6 it becomes time-wise inefficient and finally it is impossible to generate the complete structural space. The ^{13}C NMR spectrum remains a sufficient identifier of the correct constitution so that an algorithm that generates the structural space around the correct structure will succeed. If the number of possible constitutions increases above ~10^{12}, the similarity between the experimental and predicted ^{13}C NMR spectra alone is no longer a unique identifier for the correct constitution — other constitutions will gain similarly good agreements. Additional constitutional information from 2D NMR spectra is used to limit the structural space generated.

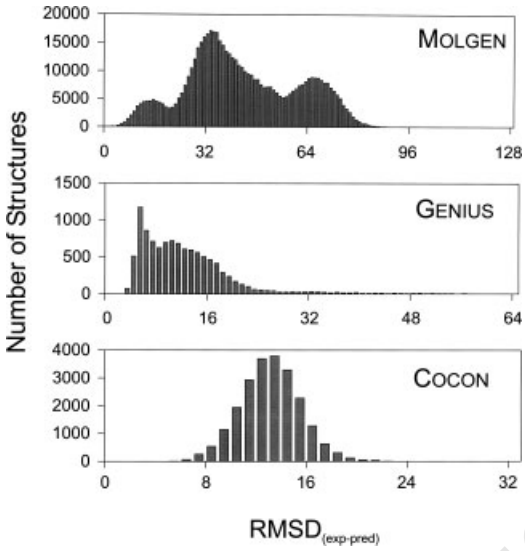

Figure 2. The three histograms visualize the distribution of the computed root mean square deviations (r.m.s.ds) between the experimental and the predicted spectrum over all members of the generated sets of molecules for the three model compounds and the three methods: (a) *N*-allyl-*N*′-ethylthiovren (**1**) solved with MOLGEN in combination with C_SHIFT, (b) tryptophan (**2**) solved with GENIUS and (c) prianosin D acetate (**3**) solved with COCON in combination with C_SHIFT.

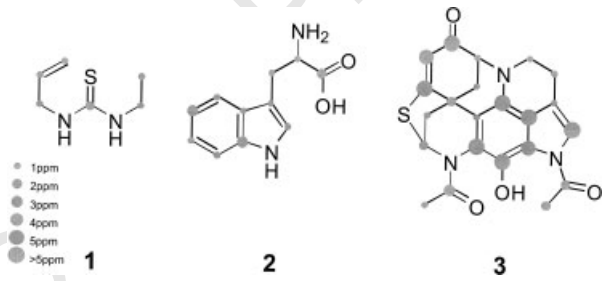

Figure 3. The deviation between the experimental and predicted ^{13}C chemical shifts for **1**,**2** and **3** is illustrated for all carbon atoms by circles of different radii.

RESULTS AND DISCUSSION

The capabilities of these approaches are illustrated with three model compounds. Table 1 summarizes some experimental details and the calculation results. The first example, a thiourea derivative (**1**) with the molecular formula $C_6H_{10}N_2S$, covers a constitutional space of about 700 000 molecular constitutions, which can be generated by MOLGEN. The two double-bond equivalents and three heteroatoms result in a structural space of medium size for a molecule with nine non-hydrogen atoms. The correct constitution is ranked with the lowest r.m.s.d. to the experimental ^{13}C NMR spectrum and is well separated from the second-ranked constitution (0.70 ppm compared with 1.65 ppm). The histogram of the r.m.s.d. value distribution between the experimental ^{13}C NMR spectrum and the computed ^{13}C NMR spectra of all constitutions is shown in Fig. 2. The deviation between the experimental and predicted ^{13}C chemical shift of **1** is illustrated for all carbon atoms in Fig. 3.

The second example, tryptophan (**2**), has 15 non-hydrogen atoms (including four heteroatoms) and eight double-bond equivalents, thus spanning a much larger constitutional space. The complete generation of all structures with MOLGEN would take some days and the prediction/analysis of the NMR spectra would take some years. By contrast, GENIUS generates only about 1[illegible]0 trial constitutions and arrives at the correct solution structure in a few minutes. The correct constitution is ranked with the lowest r.m.s.d. to the experimental ^{13}C NMR spectrum and is well separated from the second-ranked constitution (1.4 ppm compared with 2.8 ppm). The histogram of the r.m.s.d. distribution for the structural proposals of **2** is shown in Fig. 2. The deviation between the experimental and predicted ^{13}C chemical shift of **2** is illustrated for all carbon atoms in Fig. 3.

The third example, prianosin D acetate (**3**),[25] has an unknown number of possible constitutions. The published data set of experimental NMR data was completed with theoretical correlation data [a theoretical data set contains for a given constitution all ^{1}H,^{1}H-COSY correlations which are based on ^3J(H,H) interactions, for ^{1}H,X-HMBC correlations ^2J(X,H) or ^3J(X,H) interactions and for 1,1-ADEQUATE all ^2J(C,H) correlations] for the COCON calculation including all ^{1}H,^{1}H-COSY (6), ^{1}H,^{13}C-HMBC (40), 1,1-ADEQUATE (15) and ^{1}H,^{15}N-HMBC (9) correlations. Even with this almost complete theoretical data set, 22 572 different structures are generated by COCON. The ^{13}C chemical shift deviations were calculated for all structures and the correct structural proposal is ranked as 28th within the first 0.2% with an r.m.s.d. value of 5.9 ppm. The comparable

large r.m.s.d. value obtained for some natural compounds are caused by several reasons: (a) these structures are relatively seldom in databases and therefore underrepresented in the training of the neural networks, (b) they contain many highly substituted carbon atoms and (c) a lot of uncommon structural fragments, which are predicted less accurately; S—C*═C—C═O, for example, is predicted 22.3 ppm too low. However, among the top-ranked 100 structures the 28th is the only one that allows planarity of the conjugated π-electronic system and is therefore the only plausible low-energy solution (supporting information is available from the authors on request). Figure 2 shows the histogram of the r.m.s.d. distribution between calculated and the experimental ^{13}C NMR spectra. The deviation between experimental and predicted ^{13}C chemical shift of **3** is illustrated for all carbon atoms in Fig. 3.

Hence the applicability of the three algorithms correlates strongly with the size of the constitutional space. MOLGEN is limited by the computation time necessary for generating all constitutions and predicting their ^{13}C NMR spectrum. It works reliably for molecules with less than 13 non-hydrogen atoms; for larger molecules the required computation time exceeds 24 h. GENIUS can push this limit to about 20 non-hydrogen atoms, by generating only a dynamically determined part of highly probable structures. This algorithm has the disadvantage that no guarantee can be given that the correct molecule was generated. For larger molecules, the ^{13}C NMR spectrum alone is no longer a unique fingerprint of a molecule, if the uncertainties in the experiment and in the prediction are both taken into account; additional (experimental) information is necessary to obtain an unambiguous solution. Both algorithms, MOLGEN and GENIUS, profit from a 'good list' (fragments that have to be used) and a 'bad list' (fragments that are forbidden). Such fragments can be known from synthesis, experience or other experiments (UV or IR spectroscopy or mass spectrometry) and can restrict the search space dramatically. By applying such lists, the necessary computation times can be reduced and the size of the molecules can be increased. A specialized approach for incorporating additional information is COCON, which uses connectivity information from two-dimensional NMR data to decrease the size of the constitutional space. In combination with the subsequent chemical shift prediction, it is able to reduce the number of possible constitutions even for complex natural products with up to 30 non-hydrogen atoms (or even more) to a small number that can be analyzed by hand. In less complex cases the correct solution is often ranked first.

CONCLUSION

The algorithms presented here give the spectroscopist a variety of tools that help to find the correct solution structure faster and without biasing the search of structural space. For small molecules with less than 21 non-hydrogen atoms, such automated protocols can arrive at the correct solution for the majority of the standard structures in organic chemistry. However, the knowledge and experience of the chemist is required to analyze more complex problems and to evaluate the more exotic structures suggested by a structure generator.

Acknowledgements

The authors thank Martin Will, Reinhard Meusinger, Markus Mehringer, Adalbert Kerber, Thomas Lindel, Walter Maier, Jochen Junker and Erdogan Sanli for their contributions to the individual subprojects in recent years. They also thank Bill Wedemeyer for useful discussions and carefully reading the manuscript. Jens Meiler is supported by the Human Frontier Scientific Program (HFSP).

REFERENCES

1. Bremser W. *Anal. Chim. Acta* 1978; **103**: 355.
2. BASF AG, 2002.
3. Robien W. *Nachr. Chem. Tech. Lab.* 1998; **46**: 74.
4. Advanced Chemistry Development, 1996–2001.
5. Clerc J-T, Sommerauer H. *Anal. Chim. Acta* 1977; 9[illegible]3.
6. Bremser W, Ernst L, Franke B, Gerhards R, Hardt A. *Carbon-13 NMR Spectral Data*. Verlag Chemie: Weinheim, 1981.
7. Fürst A, Pretsch E. *Anal. Chim. Acta* 1990; **229**: 17.
8. Cambridge Soft Corporation, 1985–97.
9. Zupan J, Gasteiger J. *Neural Networks for Chemists*. VCH: Weinheim, 1993.
10. Kvasnicka V, Sklenak S, Pospichal J. *J. Chem. Inf. Comput. Sci.* 1992; **32**: 742.
11. Ivanciuc O. *Rev. Roum. Chim.* 1995; **40**: 1093.
12. Thomas S, Kleinpeter E. *J. Prakt. Chem./Chem.-Ztg.* 1995; **337**: 504.
13. Meiler J, Meusinger R, Will M. *Monatsh. Chem.* 1999; **130**: 1089.
14. Meiler J, Will M, Meusinger R. *J. Chem. Inf. Comput. Sci.* 2000; **40**: 1169.
15. Le Bret C. *SAR QSAR Environ. Res.* 2000; **11**: 211.
16. Meiler J. *J. Biomol. NMR* 2003; **26**: 25.
17. Meiler J, Maier W, Will M, Meusinger R. *J. Magn. Reson.* 2002; **157**: 242.
18. Benecke C, Grund R, Hohberger R, Kerber A, Laue R, Wieland T. *Anal. Chim. Acta* 1995; **314**: 141.
19. Wieland T, Kerber A, Laue R. *J. Chem. Inf. Comput. Sci.* 1996; **36**: 413.
20. (a) Lindel T, Junker J, Köck M. *J. Mol. Model.* 1997; **3**: 364; (b) Lindel T, Junker J, Köck M. *Eur. J. Org. Chem.* 1999; 573; (c) Köck M, Junker J, Maier W, Will M, Lindel T. *Eur. J. Org. Chem.* 1999; 579.
21. Meiler J, Sanli E, Junker J, Meusinger R, Lindel T, Will M, Maier W, Köck M. *J. Chem. Inf. Comput. Sci.* 2002; **42**: 241.
22. Meiler J, Meringer M. *MATCH* 2002; **45**: 85.
23. (a) Meiler J, Will M. *J. Chem. Inf. Comput. Sci.* 2001; **41**: 1535; (b) Meiler J, Will M. *J. Am. Chem. Soc.* 2002; **124**: 1868.
24. (a) Junker J, Maier W, Lindel T, Köck M. *Org. Lett.* 1999; **1**: 737; (b) Köck M, Junker J, Lindel T. *Org. Lett.* 1999; **1**: 2041.
25. (a) Cheng J-F, Ohizumi Y, Wälchli MR, Nakamura H, Hirata Y, Sasaki T, Kobayashi J. *J. Org. Chem.* 1988; **53**: 4621; (b) Kobayashi J, Cheng J-F, Ishibashi M, Nakamura H, Ohizumi Y, Hirata Y, Sasaki T, Lu H, Clardy J. *Tetrahedron. Lett.* 1987; **28**: 4939.

AQ2 AQ3 AQ4

QUERIES TO BE ANSWERED BY AUTHOR

IMPORTANT NOTE: Please mark your corrections and answers to these queries directly onto the proof at the relevant place. Do NOT mark your corrections on this query sheet.

Queries from the Copyeditor:

AQ1 what is ANN
AQ2 please give full reference
AQ3 What is this? Give publisher & location
AQ4 please give full reference