Monatshefte für Chemie - Chemical Monthly (2019) 150:927–932
https://doi.org/10.1007/s00706-019-02407-5

ORIGINAL PAPER

Computer-assisted peer reviewing of spectral data: the CSEARCH protocol

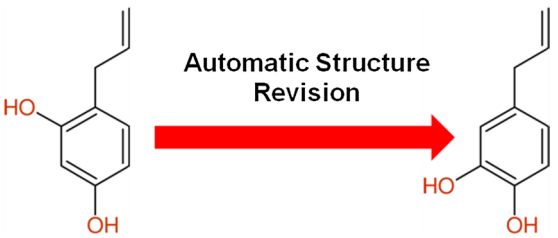
Wolfgang Robien[1]

Received: 24 January 2019 / Accepted: 27 February 2019 / Published online: 29 April 2019

Abstract

Published spectral data used in the process of structure elucidation of organic compounds cover a wide range of quality including many examples of poor characterization of the structure under investigation. The CSEARCH-Robot-Referee has proven its excellent capabilities in detecting inconsistencies between a given structure proposal and the ^{13}C NMR spectral data used for determination. The combination of this automatic structure verification tool with subsequent generation of alternative structure proposals allows fully automatic structure revisions. From the examples given here, the urgent need for a repository of already measured NMR spectra allowing automatic peer reviewing of new conclusions drawn from spectral data is clearly shown.

Graphical abstract



Keywords Computer-assisted structure elucidation · Isomer generation · Spectrum prediction · NMR spectroscopy

Introduction

Structure elucidation of organic compounds is a complex task mainly based on interpretation of spectroscopic data. Among the methods applied, NMR spectroscopy plays an important role because of the large number of sophisticated techniques available allowing deep insight into constitution, configuration, and conformation of an unknown substance. Despite the tremendous development of pulse techniques giving very detailed answers to open questions during the structure elucidation process, the interpretation of the experimental data is the main source of wrong structure proposals in the literature. A large number of review articles deals with structure revisions either arranged by the method of detection [1] or restricted to compound classes [2] or given for a certain period of time [3]. When using X-ray data as proof for determining a structure, it is well established to deposit the experimental data in a database [4] allowing later reinvestigation of them. In NMR spectroscopy, a similar intention has been massively promoted by Pauli and 72 co-authors [5], in parallel the definition of a vendor-independent format named "NMReDATA" holding the raw data and the interpretation

Dedicated to Prof. Dr. Heinz Falk on the occasion of his 80th birthday.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00706-019-02407-5) contains supplementary material, which is available to authorized users.

✉ Wolfgang Robien
Wolfgang.Robien@univie.ac.at

[1] Department of Organic Chemistry, University of Vienna, Währingerstrasse 38, 1090 Vienna, Austria

in terms of a structure proposal together with the signal assignment has been published [6]. The well-known journal "Magnetic Resonance in Chemistry" has announced to request the deposition of the experimental raw data together with the interpretation in the NMReDATA format for publications dealing with structure elucidation [7]. A comprehensive review describes the application of the "CSEARCH Robot Referee" to questionable structure proposals allowing full automatic revision of them [8]. This strategy has already been implemented into the process of submission of manuscripts for two prominent chemistry journals [9, 10]; additionally the quality management of ^{13}C NMR spectra of a database holding "new psychoactive substances" (NPS Data Hub) [11, 12] used by many governmental institutions worldwide, is also based on this concept of computer-assisted peer reviewing to ensure high-quality standards for structural data.

Within this predefined workflow, questionable assignments are identified leading to a list of positions within the proposed structure which need modification to create similar structures, which might better fit the experimental data. It should be mentioned that this process allows also adding/eliminating given substituents to optimize the coincidence between the experimental and the predicted chemical shift values; this concept is therefore not limited to generate isomeric structures only and goes far beyond the scope of traditional isomer generator programs. A simplified implementation using only the comparison of experimental chemical shift values against predicted ones with a subsequent classification was later on extended to ^{1}H NMR data [13, 14].

The workflow of the "CSEARCH-Robot-Referee" as described in [8] has proven its excellent performance initiating the structure revision of Kiusianins C and D [15]. The automatic detection of the revision of 29 structures without citing the wrong papers has been discussed in detail elsewhere [16].

The most important questions during the structure elucidation process are:

- Is the structure proposal compatible with the given experimental data?
- Is there another structure proposal compatible with the given experimental data?
- Are the experimental data already known?

Answering these three questions is possible when we systematically store all published spectral data in a central repository together with the structures derived from them. The implementation of this basic principle allows computer-assisted peer reviewing of structure-oriented publications and has been implemented into the CSEARCH NMR database [8].

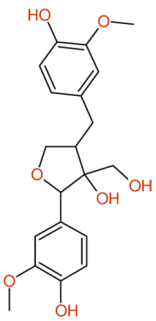
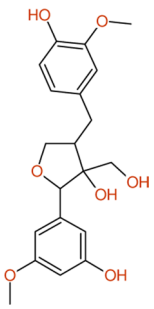
Results and discussion

Applied error propagation from misinterpretation of ^{13}C chemical shift data

The tetrahydrofuranoid lignin (-)-berchemol was isolated as a triacetate from the stems of *Berchemia racemosa* SIEB. *et* ZUCC. [17] and its structure was elucidated using ^{1}H, ^{13}C, HH-COSY, and HC-COSY NMR measurements. The ^{13}C NMR spectral data of berchemol and its triacetate together with the signal assignments are given in Table II of [17], additionally the ^{1}H spectrum as well as the HH-COSY and the HC-COSY of the triacetate are shown in the Figs. 1, 2 and 3 of [17]. Berchemol contains two 1,2-dihydroxylated benzene moieties having four characteristic quaternary carbons resonating at 144.2, 145.8, 146.6, and 146.8 ppm, respectively. This is in full agreement with tabulated shift values taken from NMR textbooks [18], easily distinguishable from 1,3-dihydroxylated benzene derivatives showing resonances typically in the range between 155 and 160 ppm.

In a later paper published by Hu et al. [19], the obviously correct spectral data of berchemol from [17] are used to elucidate the structure of a further lignan derivative isolated from *Saussurea cordifolia* (compound **5** in [19]) having one 1,2-dihydroxylated and one 1,3-dihydroxylated benzene moiety (Table 1). The spectral data for the relevant quaternary carbons are given at 145.9, 147.2, 148.6, and 149.0 ppm—which is in good agreement with the literature data of berchemol—but now used to derive a 1,2- as well as a 1,3-dihydroxylated benzene fragment giving (2*R*,3*S*,4*S*)-4-(4-hydroxy-3-methoxybenzyl)-2-(5-hydroxy-3-methoxyphenyl)-3-(hydroxymethyl)tetrahydrofuran-3-ol

Table 1 Berchemol and (2*R*,3*S*,4*S*)-4-(4-hydroxy-3-methoxybenzyl)-2-(5-hydroxy-3-methoxyphenyl)-3-(hydroxymethyl)tetrahydrofuran-3-ol together with the relevant ^{13}C NMR chemical shift values

	
Compound **1** from [17] Published in 1989—correct CAS-RN: 126882-59-5 ^{13}C: 144.2, 145.8, 146.6, 146.8 ppm	Compound **5** from [19] Published in 2010—wrong CAS-RN: 1227937-39-4 ^{13}C: 145.9, 147.2, 148.6, 149.0 ppm

The known stereochemistry is ignored in both structure diagrams for ease of comparison

(CAS-RN: 1227937-39-4). This wrong structure is probably based on a misleading interpretation of the ^{1}H NMR data as well as the HMBC data. The electronic version of this publication contains no supplementary information prohibiting therefore reinterpretation of the experimental data.

Using the above mentioned, but wrong structure (CAS-RN: 1227937-39-4) as query for searching the CAS-Registry File gives also the "Absolute stereo mirror image" having the CAS-RN 1427038-08-1 showing a 2*S*, 3*R*, 4*R* configuration. This compound appears exactly once in the chemical literature (SCIFinder, Dec 31st, 2018) published in [20] and named deltoignan A (compound 9 in [20]). The spectral data of deltoignan A (^{1}H, ^{13}C, HH-COSY, HC-HMBC) are in good agreement with the data given in [19], therefore the same wrong conclusion with respect to the constitution of the compound was done based on an already wrong structure proposal. The Fitoterapia paper [20] again has no supplementary information allowing the reinterpretation of the spectral data.

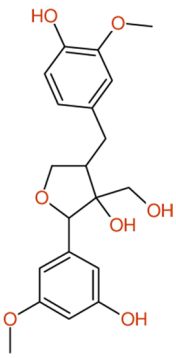
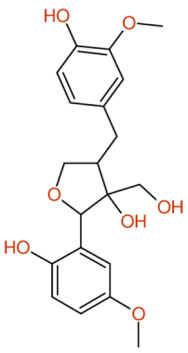
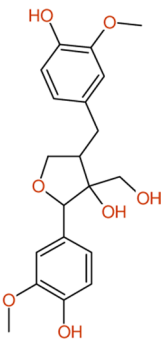
The wrong structure from [19] was further used as reference material to elucidate the structure of compound **2** in [21] creating an additional wrong example of a 1,3-dihydroxylated benzene derivative having ^{13}C NMR chemical shift values at 147.2 and 148.6 ppm for the quaternary carbons. Furthermore, another new compound named viburesinol isolated from the stems of *Viburnum erosum* [21] was elucidated having again a 1,2- as well as a 1,3-dihydroxylated benzene fragment showing the four relevant quaternary carbons at 147.5 (2×), 145.6, and 144.4 ppm incompatible with this structure proposal. The known compounds **4–10** described in [21] contain always 1,2-dihydroxylated benzene moieties with resonance lines within the expected range from 142.9 to 152.0 ppm.

Summarizing the above given detailed analysis leads to the following conclusions:

- In 1989 [17], an obviously correct structure perfectly compatible with the ^{13}C NMR data was determined.
- In 2010 [19], the correct data from 1989 [17] were used as reference material to derive a wrong structure proposal ignoring basic knowledge at textbook level.
- The paper published in 2012 [20] relies on the previous paper [19] introducing again another wrong example making the error statistically more confident.
- The paper published in 2015 [21] also relies on [19] introducing further examples of a substitution pattern incompatible with the given ^{13}C NMR data.
- All three papers successfully passed the peer-reviewing process.
- All publications have no "Supplementary Information".
- In all publications with wrong structure proposals, only already tabulated chemical shift data are given and 2D correlations are visualized in the structural diagrams [19, 20]—there is no possibility to access the raw data to redo processing and/or interpretation starting from scratch.

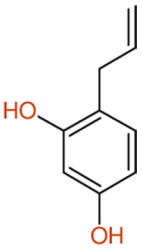
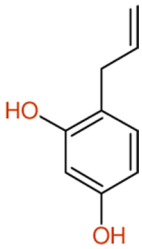
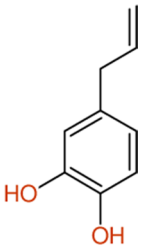
Automatic peer reviewing of the ^{13}C NMR data taken from [20] using the "CSEARCH-Robot-Referee" shows massive deviations between experimentally determined and predicted chemical shift values in the 1,3-dihydroxylated benzene moiety of deltoignan A. Systematic variation of the questionable positions using a structure generator program (Table 2) able to create similar structures leads to 3571 proposals—2 out of these 3571 structures are known compounds, either within the CSEARCH database or within the PUBCHEM collection [22, 23]. The original proposal can be found at position 298 within the sorted hitlist having an average deviation of 3.06 ppm. A real-world alternative is located at position 4 with an average deviation of 1.61 ppm, furthermore a 1,4-dihydroxylated derivative is proposed at position 1 having an average deviation of 1.39 ppm. From this result, it is clearly shown that the originally proposed structure is definitely wrong; the alternatives either have a 1,2- or 1,4-dihydroxylated benzene fragment. The position of the hydroxy and the methoxy group has to be established from the 2D-NMR data making them a necessary bit of information and therefore the raw data should be made available within the supplementary information. It should be mentioned that both wrong structure proposals [19, 20] are contained in the knowledge base used by the "CSEARCH-Robot-Referee" for this evaluation; despite that, the algorithms applied recognize this inconsistency between the given spectral data and the substitution pattern.

Table 2 Automatic structure revision of deltoignan A from [20] using exclusively the ^{13}C NMR chemical shift values; the average deviation together with the position in the sorted hit list is given

		
3.06 ppm Position 298 PUBCHEM: 75214698	1.39 ppm Position 1 PUBCHEM: unknown	1.61 ppm Position 4 PUBCHEM: 14521044

Left: original structure as published. Middle: alternative structure, unknown to PUBCHEM. Right: existing alternative structure

Table 3 Published ^{1}H NMR data together with the structures derived thereof

^{1}H Data in [25] from 2005	^{1}H Data in [24] from 2014	^{1}H Data in [26] from 2011
		
6.62 ppm; J = 8.0/1.9 Hz	6.62 ppm; J = 8.0/2.0 Hz	6.61 ppm; J = 9.0/1.8 Hz
6.71 ppm; J = 1.9 Hz	6.72 ppm; J = 2.0 Hz	6.70 ppm; J = 1.8 Hz
6.78 ppm; J = 8.0 Hz	6.81 ppm; J = 8.0 Hz	6.78 ppm; J = 9.0 Hz

Identical spectral data—different structures determined

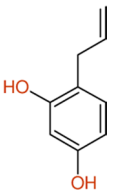
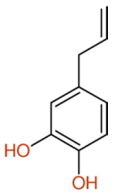
In [24], the isolation and structure elucidation of 4-allylresorcinol by IR, EIMS, ^{1}H, ^{13}C, and 2D-NMR spectra at 400 MHz was described and the interpretation of the spectral data was mainly based on a comparison of the ^{1}H NMR data (Table 3) with the chemical shift values published in [25].

The ^{1}H NMR data in all three publications are nearly identical, whereas the ^{13}C NMR data in [24] fit very well to the chemical shift values given in [26] showing a different structure proposal. Obviously the structure elucidation of 4-allylresorcinol in [25] is based on an incomplete characterization, because the ^{13}C NMR data have not been published. The conclusions given in [24] lead therefore to a wrong structure proposal. Based on the ^{13}C NMR data from [26], the structure of this compound is 4-allylbenzene-1,2-diol instead of the given 4-allylbenzene-1,3-diol (Table 4). When comparing the chemical shift values with reference material from basic NMR textbooks, it is clearly visible that the 1,2-diol gives two signals somewhere in the region of 140–145 ppm, whereas the 1,3-pattern needs two signals in the region of 155–160 ppm. The wrong structure proposed in [24] is based on the incomplete characterization of a compound in [25], whereas the obviously correct data from Ref. [26] are neglected.

The knowledge base of the CSEARCH-Robot-Referee [27] contains both entries from [24] and [26]. Despite the wrong structure proposal being available in the knowledge base, large discrepancies between the experimental and the predicted chemical shift values are found. The carbons within the benzene ring show deviations ranging from 7.7 to 18 ppm (Fig. 1).

The different structure proposal in [26] having nearly identical ^{13}C NMR data is detected via a spectral similarity search and shown as an alternative structure fitting the query spectrum. Subsequent structure generation starting from the wrong proposal [24] creates 425 topologies, 24 of them are real-world structures either known within the CSEARCH and/or the PUBCHEM database [22, 23]. The original structure proposal (1,3-diol pattern) from [24] is found at position 67 with an average

Table 4 Published ^{13}C NMR data together with the structures derived thereof and the result from the evaluation done by the CSEARCH-Robot-Referee using the structure generator

^{13}C-NMR data from [24] 424 alternative structures	^{13}C-NMR data from [26] No alternative structure
	
39.4	39.5
115.3	115.3
115.5	115.6
115.7	115.7
120.9	121.0
133.1	133.2
137.1	137.6
141.7	141.7
143.5	143.5

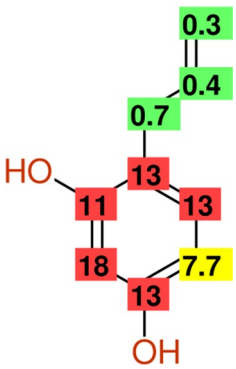


Fig. 1 Differences between experimental [24] and predicted ^{13}C chemical shift values [27] of 4-allylbenzene-1,3-diol showing the necessity of structure revision

deviation of 6.02 ppm, the best fitting real-world alternative (1,2-diol pattern) is at position 1 having a deviation of 1.39 ppm between experimental and predicted chemical shift values. This example shows the immense power of the CSEARCH-Robot-Referee performing a fully automatic structure revision which is in full coincidence with a detailed analysis of the public domain literature.

Conclusion

The examples given here show the urgent necessity to deposit experimental data in an open-access repository to allow computer-assisted peer reviewing of new conclusions drawn from new but similar data. It is usual to base conclusions on already established knowledge, but as it has been shown here that missing quality management of existing interpretation steps of experimental data helps to proliferate errors which create again wrong examples making these errors statistically more confident. Surprisingly many errors done during the structure elucidation process are quite easy to detect, because they are frequently based on ignoring knowledge even at textbook level. The combination of spectrum prediction and subsequent structure generation has been shown to be highly efficient to avoid—at least—the most trivial errors occurring during the structure elucidation process. The "CSEARCH-Robot-Referee" [27] supports the detection of inconsistencies between the experimental ^{13}C NMR data and the given structure proposal derived thereof allowing the generation of similar structures which might better fit the given data. This structure-oriented approach is optionally accompanied by a "Spectral Similarity Search" purely based on the query peaklist using a database of 235 million predicted spectra [28] allowing to retrieve alternative structures having similar spectral data.

Experimental

The "CSEARCH-Robot-Referee" uses approximately 340,000 assigned and curated ^{13}C NMR spectra representing only a small subset (ca. 0.25%) of known organic compounds. Accessing this engine is possible via a web mask [27, 28] or directly from TOPSPIN or the CMC-se program [29]. The examples presented here have been found by selecting structures with a large deviation between published and predicted chemical shift values.

Acknowledgements Open access funding provided by University of Vienna. I am grateful to Prof. Norbert Haider for making his mol2ps program available as part of the CSEARCH-Robot-Referee.

References

1. Nicolaou KC, Snyder SA (2005) Angew Chem Int Ed 44:1012
2. Suyama TL, Gerwick WH, McPhail KL (2011) Bioorg Med Chem 19:6675
3. Maier ME (2009) Nat Prod Rep 26:1105
4. The Cambridge Crystallographic Data Centre (CCDC). https://www.ccdc.cam.ac.uk. Accessed 03 Jan 2019
5. McAlpine JB, Chen SN, Kutateladze A, MacMillan JB, Appendino G, Barison A, Beniddir MA, Biavatti MW, Bluml S, Boufridi A, Butler MS, Capon RJ, Choi YH, Coppage D, Crews P, Crimmins MT, Csete M, Dewapriya P, Egan JM, Garson MJ, Genta-Jouve G, Gerwick WH, Gross H, Harper MK, Hermanto P, Hook JM, Hunter L, Jeannerat D, Ji NY, Johnson TA, Kingston DGI, Koshino H, Lee HW, Lewin G, Li J, Linington RG, Liu M, McPhail KL, Molinski TF, Moore BS, Nam JW, Neupane RP, Niemitz M, Nuzillard JM, Oberlies NH, Ocampos FMM, Pan G, Quinn RJ, Reddy DS, Renault JH, Rivera-Chávez J, Robien W, Saunders CM, Schmidt TJ, Seger C, Shen B, Steinbeck C, Stuppner H, Sturm S, Taglialatela-Scafati O, Tantillo DJ, Verpoorte R, Wang BG, Williams CM, Williams PG, Wist J, Yue JM, Zhang C, Xu Z, Simmler C, Lankin DC, Bisson J, Pauli GF (2019) Nat Prod Rep 36:35
6. Pupier M, Nuzillard JM, Wist J, Schlörer N, Kuhn S, Erdelyi M, Steinbeck C, Williams AJ, Butts C, Claridge TDW, Mikhova B, Robien W, Dashti H, Eghbalnia HR, Farès C, Adam C, Kessler P, Moriaud F, Elyashberg M, Argyropoulos D, Pérez M, Giraudeau P, Gil RR, Trevorrow P, Jeannerat D (2018) Magn Reson Chem 56:703
7. Editorial (2017) Magn Reson Chem 55:1057
8. Robien W (2017) In: Kinghorn AD, Falk H, Gibbons S, Kobayashi JI (eds) Progress in the chemistry of organic natural products, vol 105. Springer, Basel, p 137
9. Ross H (2015) Eur J Org Chem 2015:4
10. https://www.thieme.de/statics/dokumente/thieme/final/de/dokumente/gv/Planta-Guidelines-authors-2018-10.pdf. Accessed 03 Jan 2019
11. Urbas A, Schoenberger T, Corbett C, Lippa K, Rudolphi F, Robien W (2018) Forensic Chem 9:76
12. https://nps-datahub.com. Accessed 03 Jan 2019
13. https://nmrdb.org. Accessed 03 Jan 2019
14. https://nmrshiftdb.org. Accessed 03 Jan 2019
15. Shimada M, Ozawa M, Iwamoto K, Fukuyama Y, Kishida A, Ohsaki A (2018) Chem Pharm Bull 66:771
16. Robien W (2018) Eur J Org Chem 2018:3372
17. Sakurai N, Nagashima S, Kawai K, Inoue T (1989) Chem Pharm Bull 37:3311
18. Pretsch E, Clerc T, Seibl J, Simon W (1989) Tables of spectral data for structure determination of organic compounds, 2nd edn. Springer, Berlin, Heidelberg

19. Li XW, Guo ZT, Zhao Y, Zhao Z, Hu JF (2010) Phytochemistry 71:682
20. Xu JJ, Huang HQ, Zeng GZ, Tan NH (2012) Fitoterapia 83:1125
21. In SJ, Seo KH, Song NY, Lee DS, Kim YC, Baek NI (2015) Arch Pharm Res 38:26
22. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2019) Nucleic Acids Res 47:D1102
23. ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/SDF/. Accessed 03 Jan 2019
24. Salleh WMNHW, Ahmad F, Yen KH (2014) J Appl Pharm Sci 4:87
25. Ghosh K (2005) Molecules 10:798
26. Villegas AM, Catalàn LE, Venegas IM, Garcia JV, Altamirano HC (2011) Molecules 16:4632
27. https://nmrpredict.orc.univie.ac.at/c13robot/robot.php. Accessed 03 Jan 2019
28. https://nmrpredict.orc.univie.ac.at/similar/eval.php. Accessed 03 Jan 2019
29. https://www.bruker.com. Accessed 03 Jan 2019