Structure Identification of Novel Compounds using simple IR, 1H and 13C NMR spectroscopy and Computational Tools

Cho, Kwnag-Hwi; Hansen, Poul Erik

Published in:
Bulletin of the Korean Chemical Society

DOI:
10.1002/bkcs.11924

Publication date:
2020

Document Version
Peer reviewed version

Citation for published version (APA):
Cho, K.-H., & Hansen, P. E. (2020). Structure Identification of Novel Compounds using simple IR, 1H and 13C NMR spectroscopy and Computational Tools. *Bulletin of the Korean Chemical Society*, *41*(1), 78-83. https://doi.org/10.1002/bkcs.11924

Structure Identification of Novel Compounds using simple IR, ^{1}H and ^{13}C NMR spectroscopy and Computational Tools

Kwang-Hwi Cho[†,*] and Poul Erik Hansen[‡,*]

[†]School of Systems Biomedical Science, Soongsil University, Seoul 06978, Republic of Korea
**E-mail: chokh@ssu.ac.kr*

[‡]Department of Science and Environment, Roskilde University, DK-4000 Roskilde, Denmark
**E-mail: poulerik@ruc.dk*

Abstract :

A simple method is suggested to identify the structure of novel compounds with basic IR, ^{1}H and ^{13}C NMR spectroscopy and computational tools. With the molecular formula obtained from high resolution mass spectrometry, all possible isomers are calculated. The absence or presence of particular functional groups which were inferred from IR, ^{1}H and ^{13}C NMR and DEPT spectra is used to sort all the possible isomers using SMARTS code and RDKit. Then, the final structure(s) are identified by comparing the correlation between Quantum mechanically calculated ^{13}C NMR chemical shielding constants and experimental ^{13}C chemical shifts of molecules in consideration. We have applied this protocol to five natural compounds, the number of heavy atoms ranging from 11 to 15, and correctly identified the structures of all test cases. The limitations and further consideration are discussed.

Keywords: generation of isomers, spectroscopic filters, DFT calculations of nuclear shieldings

Introduction

Structure elucidation of organic molecules is typically done by a combination of 1D NMR and 2D NMR techniques such as COSY, NOESY, HSQC and HMBC spectra, mass spectrometry (MS) and may be Infra Red (IR) data. However, with the use of suitable computer programs the 2D NMR step can be avoided. This can be done in a three step process: i) calculations of all possible isomers ii) use of spectroscopic filters iii) determination of relevant NMR data.

Meiler and Meringer [1] used neural networks to predict ^{13}C chemical shifts of structures generated by MOLGEN [2-3] which can generate all possible isomers of a given molecular formula. Since the number of possible isomers growth rapidly with number of non-hydrogen atoms in a molecule, they applied their methods to the molecules containing up to 12 non-hydrogen atoms. Later Meiler and Köck [4] extended this study using either MOLGEN, GENIUS [5-6] or COCON [7-11] to increase the size of molecules that could be studied. Instead of searching for a whole space of given molecular formula as done by MOLGEN, GENIUS uses a genetic algorithm (GA) to find the possible candidates to extend the size of molecules to apply. The difference between the experimental chemical shifts and the chemical shift calculated by neural network is used as a fitness function and serve as selection criteria for the recombination step. In the paper, they claim that GENIUS can predict the structure of a molecule up to 20 non-hydrogen atoms. As well known, GA is very efficient algorithm for finding optimum solution, however, GA does not always guarantee for finding the optimum solution especially for a huge search space. They also suggested COCON to extend the applicability of automated structure elucidation to the molecules with more than 20 non-hydrogen atoms. A drawback of the COCON algorithm is that one is not sure that the correct molecule is generated.

In present study MOLGEN is used to generate possible isomers and make sure that the right answer is among the isomers generated. However, we extended the limitation by applying spectroscopic filters to reduce the number of candidates during the isomer generation step with MOLGEN or after the step, then the ^{13}C chemical shifts were calculated for the candidates using high level of quantum mechanical (QM) calculation. Lastly, the calculated chemical shielding constant were compared with experimental chemical shift to pinpoint the real molecule. In contrast to the previous works, ^{13}C chemical shifts is not the only information to find the right answer, but also ^{1}H NMR, IR and DEPT spectra are used as spectroscopic answers. The spectroscopy filters are converted into SMARTS [12] to use for a delicate filtering procedure. In the process SMARTS has been essential. Quantum mechanical calculations of nuclear shielding are used to obtained ^{13}C NMR chemical shielding constant. The resulting structures are selected by a comparison of the calculated ^{13}C chemical shielding constant with the experimental chemical shielding. The use of calculated chemical shifts rather than data from a data base, ensures that new structural elements can be treated properly.

The present approach covers compounds containing the following atoms, C, H, O and N and formulas with less than 16 heavy atoms. It has been tested on a series of structures, some of which like e.g. cytisine, have unusual structural elements.

Methods

Generation of spectroscopic filter

A high resolution MS data is leading to a molecular formula. Having obtained a molecular formula, a check can of course be made knowing symmetry from the ^{1}H spectra, counting the number of carbons from ^{13}C NMR spectra and the number of protons attached to carbon from the DEPT spectra.

The spectroscopic filters are generated in the following way following simple rules:

i) From the molecular formula the number of double bond equivalents (also referred to as degree of unsaturation) can be calculated. This can be helpful in deciding whether an aromatic moiety is present or not. From the Infra Red spectra typical functional groups like: OH, NH, C=O, COOH, C≡C, C≡N, N=C=O, NO_2 are identified.

ii) From ^{1}H NMR spectra integrals are used, presence of singlets and simple coupling patterns, which can be used to determine the substitution patterns of benzene rings (and symmetry) and to determine vicinities, are determined. For double bonds, it is possible to determine cis-trans geometries, ^{1}H chemical shifts can tell about aldehydes, carboxylic acid and presence of aromatic rings. A CH_2 group having two different chemical shifts may indicate the presence of a center of chirality. Unusual low ^{1}H chemical shifts could indicate a three-membered ring

iii) DEPT spectra are recorded to classify carbons into, C, CH, CH_2 or CH_3 types. ^{13}C chemicals shifts are primarily used to sort carbonyl containing groups into aldehydes, ketones; esters, amides and carboxylic acids. Setting up these filters it is important that also "negative" information is included, e.g. no presence of a C≡N triple bond etc.

Molecular formula and the spectroscopic filters as well as ^{13}C NMR chemical shifts are presented in Table 1 and Table 2.

Table 1. Used Mass data and Spectroscopy filters

Formula	High Res. MS	Filters	Ref
$C_{10}H_{16}O$	170.1528[a]	1 C=O, 1 4°C (CH_0), 3 CH, 2 CH_2, 2 CH_3, CH_3-CH, $(CH_3)_2$-CH, 1 3-membered ring No C=C, No C≡C	13
$C_{10}H_{14}N_2$	163.1233[b]	Aliphatic : 1 CH_3 singlet, 3 CH_2, 1CH Aromatic : 4 CH (one H is isolated), 1 4° C No NH	14
$C_{12}H_{14}O$	174.1051	1 phenyl (C_6H_5), 1 OH, 1 CH_3CH_2, 3 Aliphatic 4° C, 1 CH_3 singlet	15
$C_{11}H_{14}N_2O$	191.1179[b]	1 NH, 1 C=ONR_2 (probably conjugated), 4 CH_2, 2 4°C (one in C=ONR_2), 2 aliphatic C, 3 olefinic CH (they are neighbors), at least 1 chiral center No : C≡N, NO, C=O, CHO, OH, Ar-Ring, N=N , -N=C=O	16
$C_8H_{12}N_4O_3$	213.0988	Aliphatic : 1 CH_2 (not next to a CH_n), 1 CH_2-CH Aromatic : 2 CH (they at least 4 bonds apart), 1 4°C 1 C(=O)OH, 1 C(=O)ONH, 1 chiral center	17

[a] M+NH_4^+ [b] M+H^+

Table 2. Used experimental ^{13}C Chemical shifts data.

Formula	Types[a] and chemical shifts	Ref.
$C_{10}H_{16}O$ [b]	C=O (221.4), CH_1 (25.5, 32.9, 47.4), CH_2 (18.7, 39.7), CH_3 (18.2, 19.7, 20.0), C_q (29.7)	18
$C_{10}H_{14}N_2$	CH_1 (68.9), CH_2 (22.6, 35.3, 57.0), CH_3 (40.4), CH_a (123.6, 134.8, 148.7, 149.6), C_{aq} (138.8)	18
$C_{12}H_{14}O$	CH_2(36), CH_3 (9, 28), C_q(68, 83, 93), CH_a(128.5, 128.5, 128.5, 128.5, 132), C_{aq}(123)	19
$C_{11}H_{14}N_2O$	C=O(163.7), CH_1(27.6, 35.4, 105.1, 116.8, 138.8), CH_2(26.2, 49.7, 52.7, 53.7), C_q(150.8)	18
$C_8H_{12}N_4O_3$	C=O(172,177.9), CH_1(54), CH_2(44), CH_3(30), CH_a(118.5, 136.5), C_{aq}(130.2)	17

[a] C=O : carbonyl carbon, C_q : aliphatic 4° carbon, CH_1 : aliphatic CH, CH_2 : aliphatic CH_2, CH_3 : aliphatic CH_3, C_{aq} : aromatic 4° carbon, CH_a : aromatic CH, [b] Data for Thujone from sage

Structure Enumeration

MOLGEN is a program to generate all constitutional isomers that correspond to a given molecular formula, with optional further restrictions, e.g. presence or absence of particular substructures. The information on presence or absence of particular substructure are inferred from IR, ^1H and ^{13}C NMR and DEPT spectra as illustrated in Table 1. For structure enumeration with MOLGEN, no restrictions were applied for the cases of $C_{10}H_{16}O$, $C_{10}H_{14}N_2$, and $C_{12}H_{14}O$ in order to see the exact number of possible constitutional isomers, CPU time and the amount of storage required. For other cases, all connectivity isomers cannot be generated due to the disk space of the computer we used (shown in the second column of Table 3). Each isomer generation steps with MOLGEN require less than 24 hours with an INTEL i7 CPU machine. For the practical reason, we only tested moderate size of molecules in this study. Currently, MOLGEN uses SDF [20] format to store the outputs. There are several ways of saving structure with line notation. If MOLGEN could use SMILES [21-23], InChI [24], or yaInChI [25] as output format, it could reduce the size of storage required dramatically.

Table 3. The number of generated isomers with MOLGEN before and after final filtering

Formula	# of isomers[a]	# of isomers[b]	Δ[c]
$C_{10}H_{16}O$	452,458	113	0.00075
$C_{10}H_{14}N_2$	138,809,165	32	0.00051
$C_{12}H_{14}O$	272,917,140	13	0.03453
$C_{11}H_{14}N_2O$	N/A[d]	19	0.00008
$C_8H_{12}N_4O_3$	N/A[d]	54	0.00092

[a] all possible isomers generated with MOLGEN, [b] After filtering with SMARTS, [c] Correlation coefficient difference between ranks 1 and 2. [d] Not available because filters were used during isomer generation step.

Filtering with SMARTS code

Some of the information (presence or absence of particular substructures) were difficult to apply in the constitutional isomer generation step with MOLGEN. SMARTS and RDKIT [26] are used to filter out unnecessary isomers from the output from MOLGEN. The numbers of the final isomers are presented in the third column of Table 3. At this stage, we tried to reduce the final number of candidates as much as possible in order to reduce the number of molecule to be calculated by QM calculation. Geometry optimization was performed on the final candidates with MMFF94 [27] force field implemented in RDKIT.

Computation of ^{13}C NMR Chemical Shielding Constant

QM calculations were performed with the final candidates to get the ^{13}C NMR chemical shielding constant of the molecules. QM calculations have been done in two steps. First, Geometry optimization with PM6 followed by B3LYP/6-311+G(2d,p) level of theory has been done to reduce computation time. Then, ^{13}C NMR chemical shielding constant is calculated at mPW1PW91/6-311+G(2d,p) level of theory with the SCRF (solvent=chloroform, smd) option. The level of theory for QM calculations was suggested by Lodewyk et. al. [28] All QM calculations were done with the Gaussian 16 (G16) software package. [29]

Assigning an Experimental Chemical Shift data to a corresponding structure

G16 gives only isotropic shielding constants (which can be converted to chemical shifts) of different atoms, and no information about the real intensities. As QM NMR calculations are done with a (non-dynamic) single point structure, you will e.g. usually get different chemical shifts of protons of a methyl group, which are obviously averaged to single peak with relative intensity=3 in your observed spectrum. This may of course also be true for a number of other structures like methyl groups in isopropyl or t-butyl groups or symmetrical benzene rings. In addition, one has to take tautomrism into account. To assign an experimental data to a proper structure, correlations between experimental chemical shift and QM chemical shielding are calculated with linear regression module in scikit-learn [30] implemented in Anaconda Python [31]. To get the correlation coefficients, carbon types and values (experimental chemical shift and calculated chemical shielding) are considered at the same time. The used carbon types are listed in the footnote a of Table 2. Since the same type of carbon could have different values of chemical shields or chemical shift in a molecule, all possible combinations are considered during correlation coefficient calculation.

Results and discussion

$C_{10}H_{16}O$

$C_{10}H_{10}O$ is Thujone which is a natural compound isolated from sage. Without any filtering of particular substructure restriction, 452,458 constitutional isomers have been generated with MOLGEN. Among them, only 113 isomers are considered for QM calculation after filtering with SMARTS. As can be seen all compounds in the Figure 1 contain more than one chiral center. After considering chiral isomers of each compounds, the chiral isomers with highest correlation coefficient for each compounds were selected and the results are shown in Figure 1. The numbers below the depictions represent rank, ID, and correlation coefficient between experimental chemical shifts and computed chemical shielding constants, respectively.

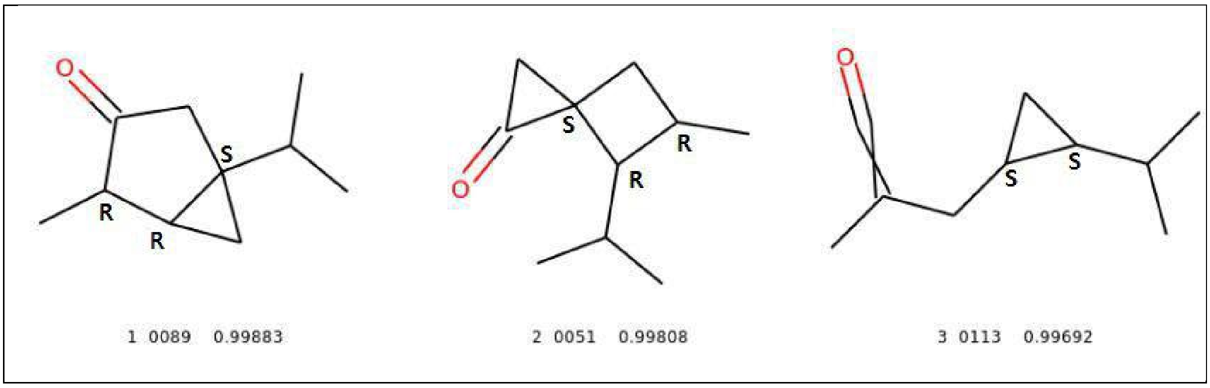

Figure 1. Three highest ranked isomers of is $C_{10}H_{16}O$. The numbers below the depictions represent rank, id, and correlation coefficient, respectively.

$C_{10}H_{14}N_2$

$C_{10}H_{14}N_2$ is Nicotine. By using only molecular formula, 138,809,165 constitutional isomers have been generated with MOLGEN. Among them, only 32 compounds remained after filtering with SMARTS. The selected molecules with highest correlation coefficient are shown in Figure 2.

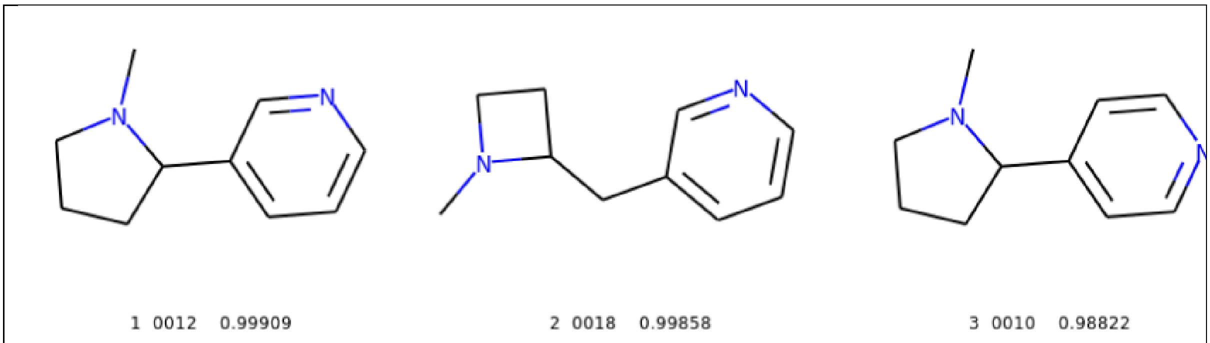

Figure 2. Three highest ranked isomers of $C_{10}H_{14}N_2$. The numbers below the depictions represent rank, id, and correlation coefficient, respectively.

$C_{12}H_{14}O$

$C_{12}H_{14}O$ is 3-methyl-1-phenylpent-1-yn-3-ol. Without any restriction of substructures, 272,917,140 constitutional isomers have been generated with MOLGEN. Among them only 13 compounds remained after filtering with SMARTS. After considering chiral isomers of each compounds, the selected molecules with the highest correlation coefficient are shown in Figure 3.

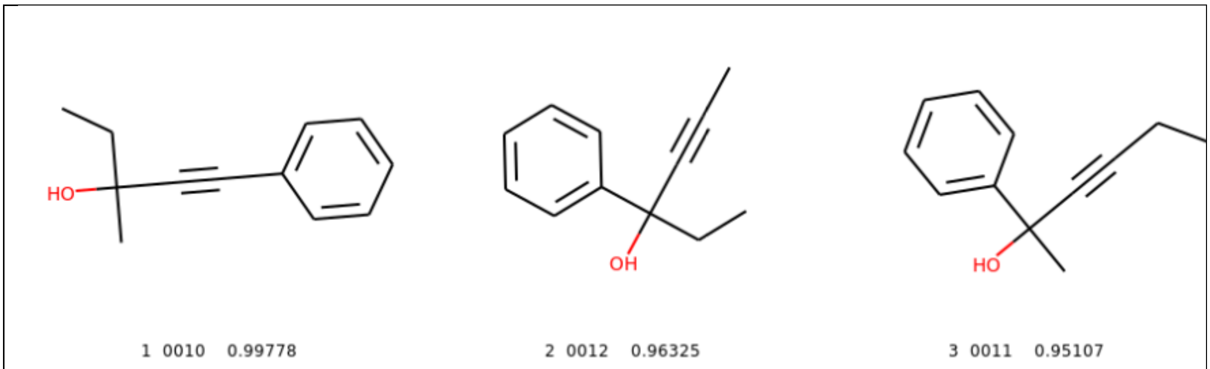

Figure 3. Three highest ranked isomers of $C_{12}H_{14}O$. The numbers below the depictions represent rank, id, and correlation coefficient, respectively

$C_{11}H_{14}N_2O$

$C_{11}H_{14}N_2O$ is Cytisine which is a natural product and a strong poison. As this is a large structure and complicated it will be discussed in some detail to illustrate how to obtain the spectroscopic filter data. The DEPT spectra provides the type of carbons (see Table 1). The ^{13}C chemical shift of 163.7 ppm suggests an ester or amide, but as the formula only contains one oxygen, the ester is excluded.

From the ^{1}H NMR spectrum (Figure 4) an OH or NH resonance can be seen. As the oxygen is part of the amide group, it has to be an NH as a broad resonance just below 2.5 ppm). Three double bond protons next to each other are seen at 7.3, 6.45 and 6.0 ppm. The CH_2 groups clearly have different chemical shifts (and they of course then show splitting due to coupling constants). This points to centers of chirality.

The Infra Red spectrum (see Figure 5) besides from confirming the presence of NH (~2900 cm^{-1}) and C=O groups (1650 cm^{-1}), the most informative part of the spectrum is from 2400 to 1700 cm-1 with no absorptions. This leaves out C#N and C#C triple bonds and -N=C=O groups. From the ^{1}H NMR spectrum no aldehyde (HC=O) groups are seen. As one O and one N is taken by the amide, no N=N and N=O groups can be present.

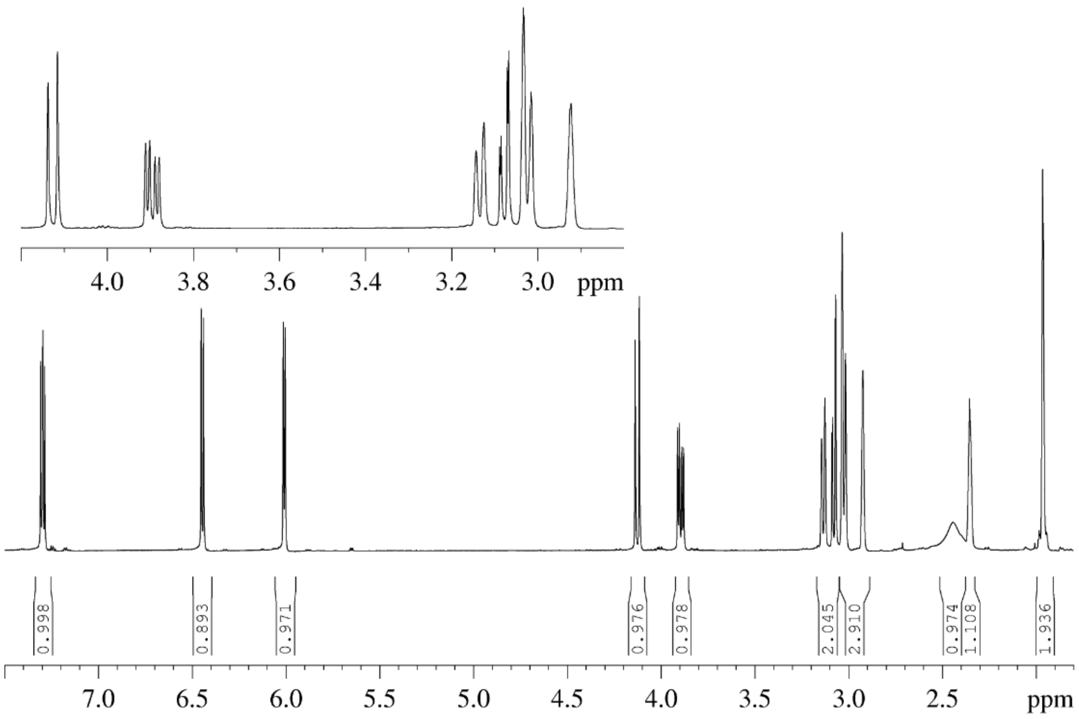

Figure 4. ^{1}H NMR spectrum of Cytisine taken from reference 18. Integrals are given below resonances.

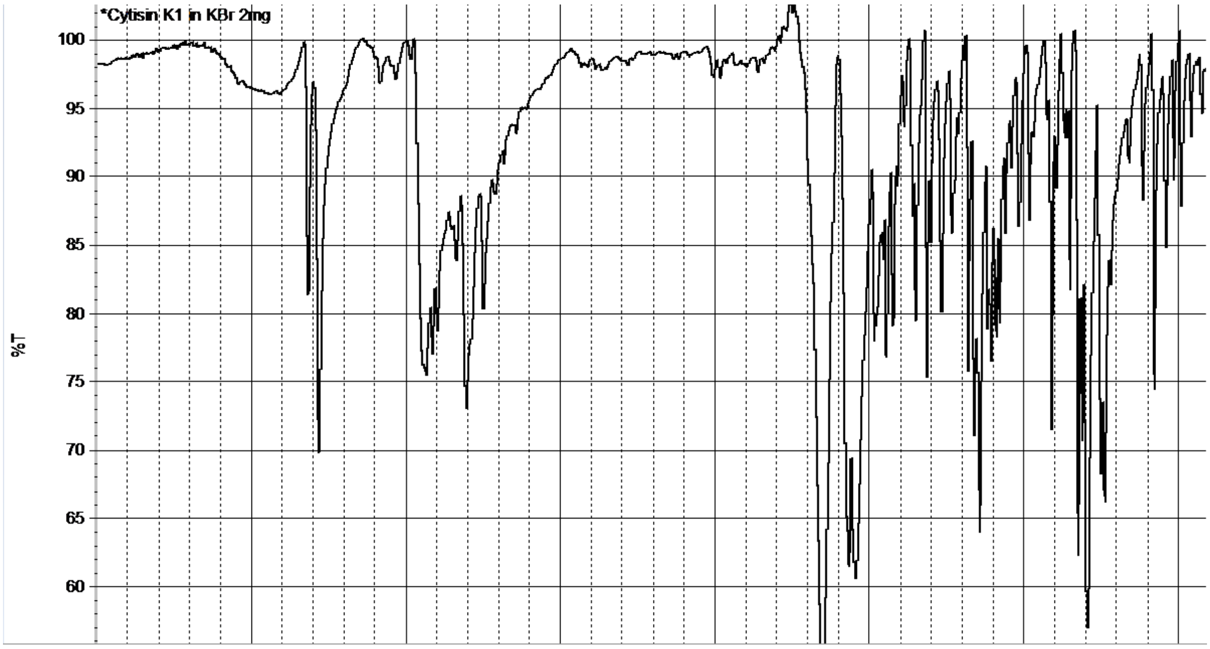

Figure 5. Infra Red spectrum of Cytisine taken from reference 18

All possible constitutional isomers cannot be generated with MOLGEN due to the storage limit of the computer used (10TB). Filtering was applied at the stage of structure generation with MOLGEN. The number of isomers generated with MOLGEN was 122 and the number is further reduced to 19 after filtering with SMARTS. For this case, puckering of the rings has been considered to find the most stable structures of each compounds. The results are shown in Figure 6. As can be seen, the difference in correlation coefficients of 0019 and 0011 is quite small for this case.

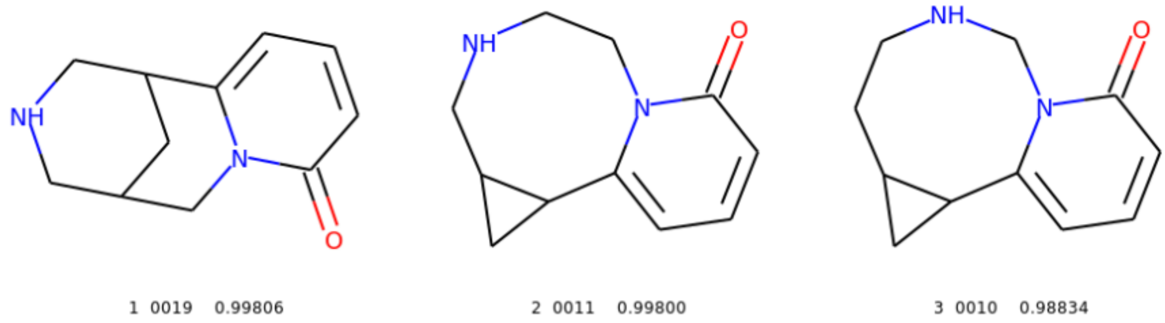

Figure 6. Three highest ranked isomers of $C_{11}H_{14}N_2O$. The numbers below the depictions represent rank, id, and correlation coefficient, respectively.

$C_8H_{12}N_4O_3$

$C_8H_{12}N_4O_3$ is the dipeptide, His-Gly. Again, filtering was applied at the stage of structure generation with MOLGEN. The number of isomers generated with MOLGEN was 33,594,278 and the number is further reduced to 54 after filtering with SMARTS. Even though $C_8H_{12}N_4O_3$ is a dipeptide, which could have two directional isomers (Gly-His and His-Gly), the right answer was found correctly. The results are presented in Figure 7.

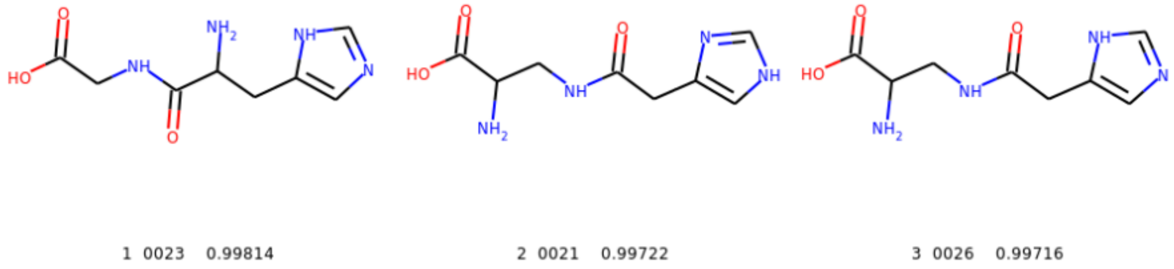

Figure 7. Three highest ranked isomers of $C_8H_{12}N_4O_3$. The numbers below the depictions represent rank, id, and correlation coefficient, respectively.

Limitations and Strengths

Structures of the five test cases have successfully been correctly been identified c. However, this protocol has a few limitations. Theoretically, MOLGEN could generate all possible connectivity isomers of a given molecular formula; however, the number of isomers increases as the number of non-hydrogen atoms in the molecules increases. It requires not only longer CPU time but also huge amount of disk space to store all isomer for those molecules. The constitutional isomer generation step with MOLGEN requires a fast computer with huge disk space to save all generated structures within reasonable time. For practical reasons, we only tested moderate size of molecules in this study. We used typically a desktop PC with Intel i7 CPU for this study. A faster computer with a parallelized algorithm of structure generation program will extend the capability of this protocol to bigger molecules. Ruddigkeit et. al. [32] generated GDB-17, which contains all possible molecules of up to 17 atoms of C, N, O, S and halogens. One can use this database instead of using MOLGEN. However, one should be aware that GDB-17 does not contain molecules with 3 or 4 membered rings. Secondly, in order to find the correct structure using correlation between experimental chemical shift and QM calculated chemical shielding constant, we have to have the right chiral structure and conformers, but this will increase the number of structures to be considered with QM calculation. As can be seen in the fourth column of Table 3, the difference in correlation coefficient between rank 1 and rank 2 is very small, which means that using a correct structure to calculate chemical shielding is very important.

Strength is clearly analyzing spectra with many rings, which can be difficult based on NMR data and especially using HMBC spectra. Furthermore, analysis of molecules with very few hydrogens should also be favored by the present technique.

Conclusions

We have successfully identified the structure of five molecules with molecular formula, using IR, ^{1}H and ^{13}C NMR data and aids in terms of computational tools. With this methodology, structure identification of novel compounds will be accelerated. Even though it has a few limitations at the current stage, using a faster computer and parallelized algorithm for structure enumeration can extend the size limit of molecules in consideration. A logic further step is to use high resolution MS data of fragments in for compounds in which certain parts of the molecule do not have many characteristic functional groups like glycosides of natural products.

Acknowledgements

KHC would like to thank KREONET (Korean Research Environment Open NETwork) which is managed and operated by KISTI (Korean Institute of Science and Technology Information). The authors wish to thank Professor Stefan Berger for making available the original drawing for Figures 4 and 5.

References

1. Meiler, J.; Meringer, M. Ranking Molgen Structure Proposals by 13C NMR chemical shift Prediction with ANALYZE. *MATCDY* **2002**, 45, 85-108.

2. Benecke, C.; Grund, R.; Hohberger, R.; Kerber, A.; Laue, R.; Wieland T. MOLGEN+, A generator of connectivity isomers and stereoisomers for molecular structure elucidation. *Analytica. Chimica. Acta.* **1995**, 314(3), 141-147

3. Wieland, T.; Kerber, A.; Laue, R. Principles of the Generation of Constitutional and Configurational Isomers. *J. Chem. Inf. Comput. Sci.* **1996**, 36(3), 413-419

4. Meiler, J.; Köck, M. Novel Methods of automated structure elucidation based on 13C NMR spectroscopy. *Magn. Reson. Chem.* **2004**, 42(12), 1042-1045

5. Meiler, J.; Will, M. Automated Structure Elucidation of Organic Molecules from 13C NMR Spectra Using Genetic Algorithms and Neural Networks. *J. Chem. Inf. Comput. Sci.* **2001**, 41(6), 1535-1546

6. Meiler, J.; Will, M. Genius: A Genetic Algorithm for Automated Structure Elucidation from 13C NMR Spectra. *J. Am. Chem. Soc.* **2002**, 124(9), 1868-1870

7. Lindel, T.; Junker, J.; Köck, M. Cocon: From NMR Correlation Data to Molecular Constitutions. *J. Mol. Model.* **1997**, 3(8), 364-368

8. Lindel, T.; Junker, J.; Köck, M. 2D-NMR-Guided Constitutional Analysis of Organic Compounds Employing the Computer Program COCON. *Eur. J. Org. Chem.* **1999**, 1999(3), 573-577

9. Köck, M.; Junker, J.; Maier, W.; Will, M.; Lindel, T. A COCON Analysis of Proton-Poor Heterocycles – Application of Carbon Chemical Shift Predictions for the Evaluation of Structural Proposals. *Eur. J. Org. Chem.* **1999**, 1999(3), 579-586

10. Junker, J.; Maier, W.; Lindel, T.; Köck, M. Computer-Assisted Constitutional Assignment of Large Molecules: Cocon Analysis of Ascomycin. *Org. Lett.* **1999**, 1(5), 737–740

11. Köck, M.; Junker, J.; Lindel, T. Impact of the 1H,15N-HMBC Experiment on the Constitutional Analysis of Alkaloids. *Org. Lett.* **1999**, 1(13), 2041–2044

12. SMARTS; A Language for Describing Molecular Patterns, available at https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (last accessed May 2019)

13. available at www.jeolusa.com/PRODUCTS/Mass-Spectrometers/AccuTOF-DART/MOTW-Spectra (last accessed May 2019)

14. Medana, C.; Santoro V.; Dal Bello, F.; Sala, C. Pazzi, M.; Sarro, M.; Calza, P. mass spectrometric fragmentation and photocatalytic transformation of nicotine and

cotinine. *Rapid Commun. Mass Spectrom*, **2016**, 30, 2617-2627.

15. Aborway,M.M *oxidative Bromination and Ring Expansion in Organic Chemistry,* Doctoral Thesis. University of Huddersfield, Great Britain. 2016.

16. Mol, H.G.J; Van Dam, R.C.J; Zomer, P.; Mulder, P.P.J. Screening of plant toxins in food, feed and botanicals using full-scan high-resolution (Orbitrap) mass spectrometry. *Food Additives & Contaminants: Part A*. **2011**, 28, 1405-1423.

17. Crews, P.; Rodríguez, J.; Jaspars, M. Organic Structure Analysis. Oxford University press. New York. **1998**.

18. Berger, J.; Sicker, D. Classics in spectroscopy. *Isolation and Structure Elucidation of Natural Products.* Wiley-VCH. Weinheim, Germany. 2009

19. Field, L.D.; Sternhell, S.; Kalman,J.R. *Organic Structures from Spectra. Fourth Edition.* Wiley, Chichester. Great Britain. 2008.

20. Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, 32(3), 244–255

21. Weiminger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28(1), 31–36

22. Weiminger, D.; Weiminger, A.; Weiminger, J.L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, 29(2), 97-101

23. Weiminger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.* **1990**, 30(3), 237–243

24. Heller, S.R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* **2015**, 7, 23

25. Cho, Y.S.; No, K.T.; Cho, K.H. yaInChI: Modified InChI string scheme for line notation of chemical structures. *SAR and QSAR in Environmental Research* **2012**, 23, 237-255

26. Landrum, G. RDKit Documentation, available at https://www.rdkit.org/ (last accessed May 2019)

27. Halgren, T.A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, 17, 490-519

28. Lodewyk, M.W.; Siebert, M.R.; Tantillo, D.J. Computational Prediction of 1H and 13C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. *Chem. Rev.* **2012**, 112(3), 1839–1862

29. Gaussian 16 Gaussian 16, Revision A.03, Frisch M. J.; Trucks G. W.; Schlegel H. B.; Scuseria G. E.; Robb M. A.; Cheeseman J. R.; Scalmani G.; Barone V.; Petersson G. A.; Nakatsuji H.; Li X.; Caricato M.; Marenich A. V.; Bloino J.; Janesko B. G.; Gomperts R.; Mennucci B.; Hratchian H. P.; Ortiz J. V.; Izmaylov A. F.; Sonnenberg J. L.; Williams-Young D.; Ding F.; Lipparini F.; Egidi F.; Goings J.; Peng B.; Petrone A.; Henderson T.; Ranasinghe D.; Zakrzewski V. G.; Gao J.; Rega N.; Zheng G.; Liang W.; Hada M.; Ehara M.; Toyota K.; Fukuda R., Hasegawa J.; Ishida M.; Nakajima T.; Honda Y.; Kitao O.; Nakai H.; Vreven T.; Throssell K.; Montgomery Jr. J. A.; Peralta J. E.; Ogliaro F.; Bearpark M. J.; Heyd J. J.; Brothers E. N.; Kudin K. N.; Staroverov V. N.; Keith T. A.; Kobayashi R.; Normand J.; Raghavachari K.; Rendell A. P.; Burant J. C.; Iyengar S. S.; Tomasi J.; Cossi M.; Millam J. M.; Klene M.; Adamo C.; Cammi R.; Ochterski J. W.; Martin R. L.; Morokuma K.; Farkas O.; Foresman J. B.; Fox D. J. Gaussian, Inc., Wallingford CT, 2016.

30. Scikit-learn; Machine Learning in Python, available at https://scikit-learn.org/stable (last accessed May 2019)

31. Anaconda Python ; available at https://www.anaconda.com/distribution/ (last accessed May 2019)

32. Ruddigkeit, L.; Deursen, R.; Blum, L.C.; Reymond, J.L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, 52(11), 2864–2875