Available online at www.sciencedirect.com
SCIENCE DIRECT®
Tetrahedron Letters 45 (2004) 6931–6935
Tetrahedron Letters

Structure-based predictions of ^{1}H NMR chemical shifts of sesquiterpene lactones using neural networks

Fernando B. Da Costa,[a,b] Yuri Binev,[c] Johann Gasteiger[b] and João Aires-de-Sousa[c,*]

[a]*Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Av. do Café s/n, 14040-903 Ribeirão Preto, SP, Brazil*
[b]*Computer-Chemie-Centrum und Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstraße 25, D-91052 Erlangen, Germany*
[c]*REQUIMTE, CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal*

Received 3 May 2004; revised 15 July 2004; accepted 19 July 2004
Available online 6 August 2004

Abstract—In this work the prediction of ^{1}H NMR chemical shifts of CH_n protons of sesquiterpene lactones by means of neural networks is described. This method is based on the incorporation of experimental chemical shifts of protons of sesquiterpene lactones as additional memory of an associative neural network system previously trained with chemical shifts of other organic compounds. One advantage of this method is its ability to distinguish between CH_2 diastereotopic protons belonging to rigid substructures since stereochemistry is considered. This is achieved via the automatic conversion of the 2D structure diagram into a 3D molecular structure. The predicted ^{1}H NMR chemical shifts of the sesquiterpene lactones showed a high level of accuracy. This is the first report on a fully automatic proton assignment of structures of sesquiterpene lactones of an accuracy that allows its use in structure elucidation.

Nowadays there is an urgent need in innovative computer-based procedures that can assist the natural product chemist (phytochemist) in the process of structure determination of natural compounds. For several reasons, in many cases this process is the most time-consuming step in the entire course of the study of compounds from living organisms. The currently available computer-based procedures that can serve as useful tools in the structure determination of organic compounds are focused on spectra prediction, spectra interpretation, structure generation and NMR chemical shift prediction.[1,2] The methods on which they are based usually involve artificial intelligence in its broad sense.[2] Nevertheless, as only few of these procedures are exclusively focused on special classes of organic molecules one may say that the great majority can be classified as 'generalists'. So, the risk of inaccuracy is high and sometimes the phytochemists have no confidence in the ability of such computer-based procedures to succeed and doubt that they are really useful to solve their specific problems.

In this work, we describe for the first time the use of neural networks to estimate ^{1}H NMR chemical shifts of a special class of natural products, the sesquiterpene lactones (STLs). This tool is a combination of two different neural network approaches and comprises a new aid to help the phytochemist in the process of structure elucidation of STLs by providing highly accurate ^{1}H NMR data for 3D-structures taking into account stereochemistry as well as diastereotopic CH_2 protons. In the course of the conventional process of structure determination of natural compounds like the STLs, there is a step where it is important to the phytochemist to compare the ^{1}H NMR data of an unknown with those of reference structures. This step involves structure verification and structure confirmation and in many cases it may be time consuming. At this point the computer methods can play an important role, since they can provide the phytochemist with a large amount of spectral data in a short period of time. An important aspect in this particular situation is the quality of such data, which undoubtedly must reflect the level of its accuracy.

Keywords: Chemical shift prediction; ^{1}H NMR spectroscopy; Neural networks; Sesquiterpene lactones; Structure-based prediction.
* Corresponding author. Tel.: +351-21-2948300; fax: +351-21-2948550; e-mail: jas@fct.unl.pt

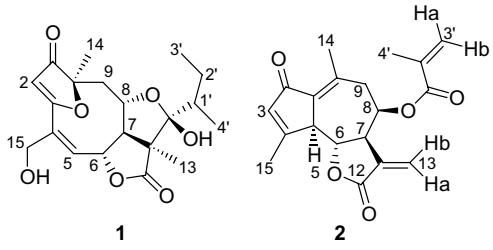
0040-4039/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.tetlet.2004.07.082

Although it can be easier with ^{13}C NMR data, it is extremely difficult to achieve a high level of accuracy with ^{1}H NMR spectra due to the problems inherent in this technique.[2] Especially when STLs are involved, to gain accuracy is a challenge because of the relatively great complexity of these structures.

The STLs are a special class of terpenoids comprising about 4000 structures with more than 30 skeleton subtypes and several substitutional features.[3,4] They are the typical metabolites from the large plant family Asteraceae, but are also present in species from Lauraceae, Apiaceae, Burseraceae and Magnoliaceae as well as in liverworts (Hepaticae).[4] Since the STLs show a vast array of biological activities,[3,5] have ecological importance[4,5] and are used as taxonomic markers in the family Asteraceae,[4] they have chemical, biological, medicinal and commercial interest. Thus, structure elucidation plays an important role in all of the studies concerning this class of compounds. The most widely used method in the structure elucidation of STLs is NMR spectroscopy and ^{1}H NMR spectral data analysis is usually the starting point, since it is one of the most informative techniques for this purpose. Several scientific publications including books,[6,7] journal articles[8] and reviews[9] have appeared during the last 30 years with regard to this subject. Furthermore, conformational studies based on minimizations of 3D structures as well as analysis of X-ray diffraction data are also accessible as supporting information.[7,10,11] Finally, an expert system has been developed to perform structure elucidation of STLs based on ^{13}C NMR data.[12] So far, there is not any work concerning the use of neural networks as a tool for structure elucidation of STLs based on NMR spectroscopy.

In this study, we used ensembles of previously trained feed-forward neural networks (FFNNs) for ^{1}H NMR chemical shift prediction.[13,14] The neural networks had been trained with experimental chemical shifts of hundreds of structures,[13] containing only one STL of the eudesmanolide subtype.[2] In that system, hydrogen atoms of the structures are represented by physicochemical, geometric and topological descriptors.[13] The method has been previously tested for general application to organic compounds and showed good results,[13] even in comparison to other methods for ^{1}H NMR chemical shift predictions.[14] Furthermore, it was demonstrated that predictions for a given group of compounds could be significantly improved if the FFNNs were bound to an additional memory of related compounds.[14] This memory is simply a list of protons represented by their descriptors together with the experimental chemical shifts. The result—an ensemble of neural networks together with the memory—is called an associative neural network (ASNN).[15] In this work, additional data comprising 392 experimental chemical shifts of 20 STLs have been added to the ASNNs as additional memory. It means that specific user-defined data were incorporated to an initial memory containing data of other classes of organic compounds.[14,15] This data set of STLs was elaborated based on published data and comprises several subtypes of skeletons as well as different substitutional features, like side chain esters and other functionalities. This procedure is necessary to improve the accuracy of the predictions in comparison to the original networks without the additional memory.[13,14] Certainly, this is the most important step in the way to improve the predictions with respect to a specific class of natural compound like the STLs.

In order to test the ability of the neural networks to predict ^{1}H NMR chemical shifts of STLs, two of them were selected from published data. The first is an eremantholide (15-hydroxyeremantholide B) (**1**)[16] and the second is a guaianolide (**2**).[17] Both compounds are not present in the memory. They were selected because of their structural complexity, the presence of chiral centers and diastereotopic CH_2 protons as well as the availability of the ^{1}H NMR data of all hydrogen atoms.[16,17] Furthermore, **1** and **2** as well as their closely related derivatives show biological activities and are challenging targets for total synthesis.[17–19] The ^{1}H NMR data of **1** and **2** were obtained in $CDCl_3$ at room temperature. In this study, protons of hydroxyl groups were not considered, since their chemical shifts are strongly affected by experimental conditions.



The predicted ^{1}H NMR chemical shifts obtained by neural networks for **1** and **2** with and without the memory, their ^{1}H NMR experimental data as well as the differences (Δ) between the experimental and the predicted values are listed in Table 1. The *mean absolute errors* for the predictions of the chemical shifts of all protons of **1** and **2** without the memory (Pred-FF) were 0.24 and 0.36 ppm, while the errors using the memory (Pred-Mem) were 0.14 and 0.23 ppm, respectively. These results show a high degree of improvement in the predictions using the memory and will be discussed below.

In the predicted values for **1** (Table 1), an interesting observation is with respect to H-9α and H-9β, because they are CH_2 diastereotopic protons and their different chemical shifts have been distinguished. The predictions for these protons after the memory has been employed (1.95 and 2.46 ppm, respectively) showed excellent agreement with the experimental data (2.02 and 2.31 ppm, respectively) and the observed differences (Δ) are small (Table 1). The prediction of the two olefinic protons H-2 and H-5 also showed good agreement with the experimental data. The difference (Δ) of the chemical shift of H-15 is 0.14 ppm, while other seven ones are 0.04 ppm or below (Table 1). These results are of high accuracy, since they are in the range of accepted experi-

Table 1. Experimental (Exp) ^{1}H NMR chemical shift values of the sesquiterpene lactones **1**[a] and the **2**[b], their predicted values without the enhanced memory (Pred-FF) and their predicted values using the enhanced memory (Pred-Mem) as well as the respective differences (Δ) between the experimental and the predicted values (all values are given in ppm)

H	**1**					**2**				
	Exp	Pred-FF	Δ	Pred-Mem	Δ	Exp	Pred-FF	Δ	Pred-Mem	Δ
2	5.72	5.77	0.05	5.74	0.02	—	—	—	—	—
3	—	—	—	—	—	6.23	5.62	−0.61	5.56	−0.67
5	6.32	6.19	−0.13	6.19	−0.13	3.54	3.50	−0.04	3.50	−0.04
6	5.01	4.59	−0.42	4.99	−0.02	4.09	4.11	0.02	4.33	0.24
7	2.86	3.48	0.62	3.55	0.69	3.15	2.96	−0.19	3.18	0.03
8	4.03	3.76	−0.27	3.99	−0.04	5.75	4.83	−0.92	5.15	−0.60
9α	2.02	2.43	0.41	1.95	−0.07	2.86	2.50	−0.36	3.18	0.32
9β	2.31	2.45	0.14	2.46	0.15	2.74	2.57	−0.17	2.50	−0.24
13a	1.31	1.46	0.15	1.34	0.03	6.24	6.32	0.08	6.33	0.09
13b	—	—	—	—	—	5.55	6.21	0.66	6.13	0.58
14	1.47	1.74	0.27	1.51	0.04	2.35	1.86	−0.49	2.25	−0.10
15	4.34	4.43	0.09	4.48	0.14	2.35	1.79	−0.56	2.17	−0.18
1′	2.02	1.78	0.24	2.03	0.01	—	—	—	—	—
2′a	1.71	1.45	−0.26	1.30	−0.41	—	—	—	—	—
2′b	0.96	1.45	0.49	1.30	0.34	—	—	—	—	—
3′a	0.95	1.00	0.05	0.91	−0.04	5.97	5.69	−0.28	6.03	0.06
3′b	—	—	—	—	—	5.57	6.12	0.55	5.56	−0.01
4′	1.04	1.11	0.07	1.01	−0.03	1.86	1.97	0.11	1.87	0.01

[a] Mean absolute error with Pred-FF = 0.24 and with Pred-Mem = 0.14 ppm.
[b] Mean absolute error with Pred-FF = 0.36 and with Pred-Mem = 0.23 ppm.

mental variations of the ^{1}H NMR technique. There are examples in the literature with natural compounds exhibiting such a variation in the chemical shift values of their protons when the spectra are run in different spectrometers. Only three differences (Δ) in the prediction of the chemical shifts of **1** using the memory are larger than 0.30 ppm. The first is related to the chemical shift of H-7 (Δ = 0.69 ppm), being the largest value in both compounds. However, in the ^{1}H NMR spectrum of **1**, H-7 is coupled to H-6 and H-8, giving rise to a typical double doublet that is easily assigned by the phytochemist.[16] The other two predicted chemical shifts in which large differences were observed are those of the two H-2′ of the butyl side chain (Δ = 0.41 and 0.34 ppm). In the ^{1}H NMR spectrum of **1**, the signals of H-2′ are the only two multiplets, which show coupling between each other and also with H-1′ and H-3′ and they can be inferred by spectral analysis.[16] These diastereotopic CH_2 protons were not distinguished by the system (Table 1) because stereochemistry is only considered for protons belonging to rigid substructures.[2,13,14] It should be mentioned that these two H-2′ protons have not been experimentally assigned in the structure.[16] As already mentioned, the comparison of the ^{1}H NMR data of an unknown compound with those of reference structures is an approach that is done at a certain point in the entire course of structure determination of the unknown. It is important to keep in mind that it is only in this step that structure verification and structure confirmation are necessary, since the phytochemist has proposals for the possible candidates. According to the purpose of this work, it is exactly at this point that the prediction of ^{1}H NMR chemical shifts is a valid and helpful tool. If we consider these facts and analyze the complexity of the structure of **1** and the obtained results, it can be stated that large differences in the prediction of only three chemical shifts are not able to cause any problem related to misinterpretation of the entire set of data.

The predicted values for **2** (Table 1) have also a high degree of accuracy. In the structure of **2** there is a pair of diastereotopic CH_2 protons at C-9 that have been distinguished by the networks and the predictions of their chemical shifts using the memory are in accordance with the experimental data (Table 1). This case illustrates how the information contained in the enhanced memory could improve the prediction of stereochemical effects. As observed in the diastereotopic protons of **1**, in those of **2** the high and the low field experimental values of the CH_2 protons have the correspondent predicted values in the same order. Other positive aspects should be mentioned. First, is the distinction of the chemical shifts of the olefinic *gem*-protons at C-13, a key point in the structure elucidation of all α,β-unsaturated STLs with an exocyclic methylene group. This difference in the chemical shifts of H-13a and H-13b in 6,7-lactonized compounds is due to the deshielding effect of the carbonyl group at C-12 on H-13a, which has low field absorption. A second aspect is related to the methacrylate moiety, the side chain ester attached to C-8. All the predictions for the hydrogen atoms of this ester were quite accurate and the olefinic *gem*-protons at C-3′ have also been distinguished. Finally, the H-5 and H-7 signals, which have close experimental chemical shift values (3.54 and 3.15 ppm, respectively), were predicted with a high level of accuracy according to the experimental data. The largest difference (Δ = 0.67 ppm) in the prediction of the chemical shifts of **2** is related to H-3, the olefinic proton located on a carbon atom between a carbonyl group and an sp^2 carbon atom. Nevertheless, this assignment is also easy to be performed by the phytochemist, since the signal of H-3 is the remaining olefinic signal that appears in the spectrum after both H-13 and

the two H-3′ protons, which can be simply assigned by exclusion.

These results were obtained either by an ensemble of FFNNs,[20] or by an ASNN system with a memory of ca. 4000 chemical shifts of general organic compounds (including the chemical shifts that had been used to train the FFNNs) enhanced with 392 chemical shifts of STLs. Additional experiments were performed with a memory consisting exclusively of the STLs (and the chemical shifts that had been used for training the FFNNs). In this case, the errors observed were 0.18 and 0.22 ppm, respectively for **1** and **2**, which are generally better than those obtained by the FFNNs (0.24 and 0.36 ppm), but slightly inferior to those with the larger memory. When the FFNNs were retrained with the data from the 20 STLs (and the data that had been used for the initial training) errors of 0.19 and 0.31 ppm were obtained for **1** and **2**. These results are better than those obtained by the initial FFNNs, but worse than those obtained using ASNNs. As in previous work,[14] re-training has not started with the individual FFNNs taking random weights, but with the weights of the previously trained FFNNs.

These results show that the use of FFNN ensembles in combination with the memory of an ASNN system centered on a special class of natural compound, as the STLs, yields highly accurate results for the ^{1}H NMR chemical shift prediction of protons of individual structures. In this case, the combination of a general with a specific memory was very helpful to achieve more accurate results. The level of improvement of the errors for both structures is around 60%. Another aspect to be analyzed is that the speed of the overall process is very high. In this work, the predictions for each structure were achieved in less than 3 s using a common PC (Pentium III 600 MHz, 256 MB RAM). Such performance is very important when assignments of protons of several compounds are considered. The percentage of large differences ($\Delta \geqslant 0.30$ ppm) that were observed between the experimental and the predicted chemical shifts (Pred-Mem) of the protons of **1** and **2** is very low. The interpretation of such differences can be achieved through the analysis of the individual spectra and they are not capable of hindering the correct assignment of all protons of both compounds by the phytochemist. The great majority of the differences (Δ) is very small, being in the range of the normal experimental variations of the ^{1}H NMR technique. The 3D representation of the structures together with the assignment of diastereotopic CH_2 protons is a mandatory requirement to achieve good results as those in this study, since such protons are highly influenced by their 3D environment.

The addition of data to the memory of ASNNs is a strategy used in different areas of research, such as chemoinformatics, medicinal and physical chemistry.[15] For example, it has been successfully used for the prediction of physicochemical properties of several chemical compounds, like lipophilicity and aqueous solubility.[21] It was shown that the incorporation of user-defined additional data is helpful to achieve better performance for predictions, being also an important tool for fast development and updating new models by nonexperts. In this work, the obtained results corroborate those observations.

Acknowledgements

F.B.C. is grateful to the Alexander von Humboldt Stiftung (Germany) for a research fellowship at CCC and also for a Europe Research Fellowship to REQUIMTE in Portugal. Y.B. acknowledges Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for a post-doctoral grant under the POCTI program (SFRH/BPD/7162/2001). Chemical Concepts GmbH is acknowledged for providing most of the ^{1}H NMR experimental data contained in the larger memory.

References and notes

1. Steinbeck, C. In *Computer Assisted Structure Elucidation*, Gasteiger, J.; Engel, T., Eds.; Handbook of Chemoinformatics; Wiley-VCH: New York, 2003; Vol. 3, Chapter 2.3, pp 1378–1406.
2. Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J. *Anal. Chem.* **2002**, *74*, 80–90.
3. Schmidt, T. J. *Curr. Org. Chem.* **1999**, *3*, 577–608.
4. Seaman, F. C. *Bot. Rev.* **1982**, *48*, 123–551.
5. Picman, A. K. *Biochem. Syst. Ecol.* **1986**, *14*, 255–281.
6. Yoshioka, H.; Mabry, T. J.; Timmermann, B. N. *Sesquiterpene Lactones: Chemistry, NMR and Plant Distribution*; University of Tokyo Press: Tokyo, 1973.
7. Fischer, N. H.; Oliver, E. J.; Fischer, H. D. In *The Biogenesis and Chemistry of Sesquiterpene Lactones*; Herz, W.; Griesebach, H.; Kirby, G. W., Eds.; Progress in the Chemistry of Organic Natural Products; Springer-Verlag: Viena, 1979; Vol. 38, pp 47–390.
8. Samek, Z.; Yoshioka, H.; Mabry, T. J.; Irwin, M. A.; Geissman, T. A. *Tetrahedron* **1971**, *27*, 3317–3322.
9. Buděšínský, M.; Šaman, D. *Ann. R. NMR S* **1995**, *30*, 231–475.
10. Watson, W. H.; Kashyap, R. P. *J. Org. Chem.* **1986**, *51*, 2521–2524.
11. Schmidt, T. J. *J. Mol. Struct.* **1996**, *385*, 99–112.
12. Emerenciano, V. P.; Rodrigues, G. V.; Macari, P. A. T.; Vestri, S. A.; Borges, J. H. G.; Gastmans, J. P.; Fromanteau, D. L. G. *Spectroscopy* **1994**, *12*, 91–98.
13. Binev, Y.; Aires-de-Sousa, J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 940–945, A web interface for FFNN prediction of chemical shifts (without additional memory) is freely accessible at http://www.dq.fct.unl.pt/spinus and http://www2.chemie.uni-erlangen.de/services/spinus.
14. Binev, Y.; Corvo, M.; Aires-de-Sousa, J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 946–949.
15. Tetko, I. V. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717–728.
16. Vichnewski, W.; Skrochy, C. A.; Nasi, A. M. T. T.; Lopes, J. L. C.; Herz, W. *Phytochemistry* **1999**, *50*, 317–320.
17. Schorr, K.; García-Piñeres, A. J.; Siedle, B.; Merfort, I.; Da Costa, F. B. *Phytochemistry* **2002**, *60*, 733–740.
18. Koch, E.; Klaas, C. A.; Rüngeler, P.; Castro, V.; Mora, G.; Vichnewski, W.; Merfort, I. *Biochem. Pharmacol.* **2001**, *62*, 795–801.
19. Takao, K.; Ochiai, H.; Hashizuka, T.; Koshimura, H.; Tadano, K.; Ogawa, S. *Tetrahedron Lett.* **1995**, *36*, 1487–1490.

20. Performance of the FFNNs (without memory) relatively to STLs could be assessed by predicting the 392 chemical shifts for the 20 STLs. The mean absolute errors for aliphatic protons in nonrigid substructures, nonaromatic pi protons, and aliphatic protons in rigid substructures were 0.14, 0.22, and 0.42 ppm, respectively.
21. Tetko, I. V.; Tanchuk, V. Y. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136–1145.