DOI 10.1007/s10600-015-1475-5
Chemistry of Natural Compounds, Vol. 51, No. 5, September, 2015

INTERPRETIVE LIBRARY SEARCH OF PLANT COMPOUND SPECTRA IN A ^{13}C NMR DATABASE[#]

S. Nachkova,[1] S. Milenkova,[1] P. Penchev,[1*] and P. Bozov[2]

The derivation of structural characteristics of a compound of unknown structure from its spectral data is a central procedure for modern structure elucidation. Computer searching in spectral libraries of fully-assigned ^{13}C NMR spectra of substructures or full structures is an essential part of structure elucidation and has been widely applied because this type of spectra reflects the nature of the skeletal backbone of an organic compound, information not as readily available by other spectroscopic techniques [1]. In this paper we describe an extensive test of a previously developed method for interpretive search in spectral libraries of fully-assigned ^{13}C NMR spectra [2]. The method is implemented into a Windows-based user-friendly program, called Infer C NMR.

The program input consists of the ^{13}C NMR spectrum of the unknown compound (chemical shift and multiplicity of each signal) and molecular formula. The search algorithm retrieves a list of connected substructures from the reference compounds in such a way that only atoms with matched signals are included into the inferred substructures. The substructures are explicitly defined in terms of atom type, hydrogen multiplicity, and bond type. They are presented embedded into the reference structures and are sorted according to their reliability (estimation of their correctness, usually called accuracy). This accuracy is calculated by a multivariate function that was obtained in advance by comprehensive statistics. The parameters that restrict the search algorithm are the tolerance of signal matching (Tol) in ppm and the minimum number of carbon atoms in the inferred (retrieved) substructures (m.n.c.); the latter is set to six for this study.

Although our program for interpretive library search is intended to serve as a useful stand-alone application for the spectroscopist, its output can be sent as input to a computer-enhanced structure elucidation system, such as SESAMI [3]. In this mode, one or more of the retrieved substructures act as constraints on the structure generation process, serving to reduce the number of plausible alternative structures that are presented to the chemist by the structure generator. The greater the number and information content of the constraints, the greater the efficiency of the structure generator and the fewer the structures produced. It is important to recognize that if a substructure predicted as present is handed to the structure generator, every structure output will contain that substructure. Thus, if even only one of the retrieved substructures used as constraints is incorrect, every structure produced by structure generator will be invalid (the worst scenario) or no output structures will be generated because the constraints contradict each other or other spectral data (a better scenario). That is why the output from the interpretive ^{13}C NMR library search must have two very important features: high information content and high reliability.

As described in a previous paper [2], the accuracy function was tested with a large validation set of nearly 12,740 spectra by leave-one-out cross-validation. These spectra were part of the library and some of them are not natural compounds but smaller ones produced by chemical synthesis. That is why the present test gives a better estimation of real-world capabilities of the interpretive library search.

One hundred and four ^{13}C NMR spectra of compounds isolated from plants and published in *Phytochemistry* (year 2002, volumes 58–59) were searched in a library of 38 225 fully-assigned ^{13}C NMR spectra. Four spectra retrieved no substructures.

[#]Dedicated to the 85th birthday of Prof. Morton Munk, ASU, Arizona, USA.

1) Department of Analytical and Computer Chemistry, University of Plovdiv, 4000, Plovdiv, Bulgaria, e-mail: plamen@uni-plovdiv.net; 2) Department of Biochemistry and Microbiology, University of Plovdiv, 4000, Plovdiv, Bulgaria.
Published in *Khimiya Prirodnykh Soedinenii*, No. 5, September–October, 2015, pp. 852–854. Original article submitted February 14, 2014.

TABLE 1. Estimation of Substructure Accuracy

Threshold accuracy, %	N.a.s.	N.c.s.	N.i.s.	Estimated accuracy, %
with 100 spectra from *Phytochemistry* (year 2002, vols. **58**–**59**)				
90	24 131	22 482	1649	93.2
95	16 609	15 910	699	95.8
99	7969	7868	101	98.7
with 10 spectra taken from [4–8]				
90	3214	2923	291	91.0
95	2252	2172	80	96.5
99	1086	1084	2	99.8

N.a.s.: number of all substructures predicted with an accuracy higher than given threshold; N.c.s. and N.i.s.: number of correct and incorrect substructures, respectively.

The remaining 100 spectra predicted from 2 to 7518 substructures (874 on average) with the accuracy of the first ranked substructure ranging from 77.2% to 100% (97.0% on average). The first ranked substructure (it is with the highest accuracy) is usually used by structure elucidation; in 79 of cases it was correct. The 21 incorrect first substructures were predicted with accuracy ranging from 77.2% to 96.1% (89.5% on average), and the 79 correct first substructures were predicted with accuracy ranging from 89.0 to 100% (99.0% on average); the accuracy intervals for both classes – that of correct and that of incorrect first substructures – overlap each other. But if an accuracy of 97% is set as a threshold value, no one substructure ranked as first would be incorrect. In this case, 72 of the correct predicted first substructures have an accuracy higher than this threshold, which gives a 91.1% (=72/79) recall of the correct first substructures.

For the estimation of substructure accuracy, the retrieved substructures in all 100 output lists were processed. The estimate is calculated as the number of correct substructures as percent of all substructures that have an accuracy higher than a given threshold value. The estimates are listed for threshold values of 90%, 95%, and 99% in the last column of Table 1. As can be seen, the estimates are close to the corresponding threshold values (compare the first and last columns). This means that the reliability function can be used when the method is applied to structure elucidation of natural compounds. Additionally, Table 1 as a whole casts a glance at the meaning of the substructures accuracy.

As we started a collaboration with one of the coauthors (P.B.), who has practical experience for isolation of plant compounds and their structure elucidation, we additionally composed a small validation set of spectra of 10 compounds isolated by P.B. [4–8]. These ten compounds were isolated from plants and are suitable for checking the efficiency of the interpretive search. The corresponding estimation of accuracy with these ten spectra is given in Table 1. The estimates are very close to the corresponding thresholds and even exceed the thresholds.

The ten test spectra predicted from 154 to 5878 substructures (1720 in average), with the accuracy of the first ranked substructure ranging from 95.5 to 100% (98.8% in average). Only in one of the cases is the first substructure incorrect; its accuracy is 95.5%, below the 97% threshold discussed above. The remaining nine correct first substructures were predicted with accuracy ranging from 97.1 to 100.0% (99.1% in average); then the accuracy of all correct first substructures exceed the 97% threshold, i.e., we have 100% recall.

The information content of retrieved substructures can be estimated by their efficiency in the structure generation process when they are entered as GOODLIST fragments. By "efficiency" here is meant the reduction of the number of plausible structures that are the output of the generator. Schriber and Pretsch revealed some common regularities for that type of application [9]. For example, a good-list fragment containing a less frequent element is more efficient than another one with an element of higher occurrence [9], but it is obvious that the most trivial one – the size of the fragment – is the most important factor for structure generator efficiency.

The test compound structures are given together with the first inferred substructures in Fig. 1; the latter are represented embedded into the corresponding reference compound structure. The size of the ten query structures (calculated as the number of heavy atoms) ranges from 25 to 43 (34 on average), and that of the ten generated first substructures from 6 to 20 (9 on average). The size of the nine correct inferred first substructures varies from 17 to 56% of that of the corresponding 'unknown' structure. This constitutes a large amount of structural information; moreover, in some cases the oxygen atoms in the substructure increase its information content [9], as is the case in two-third of the correct substructures.

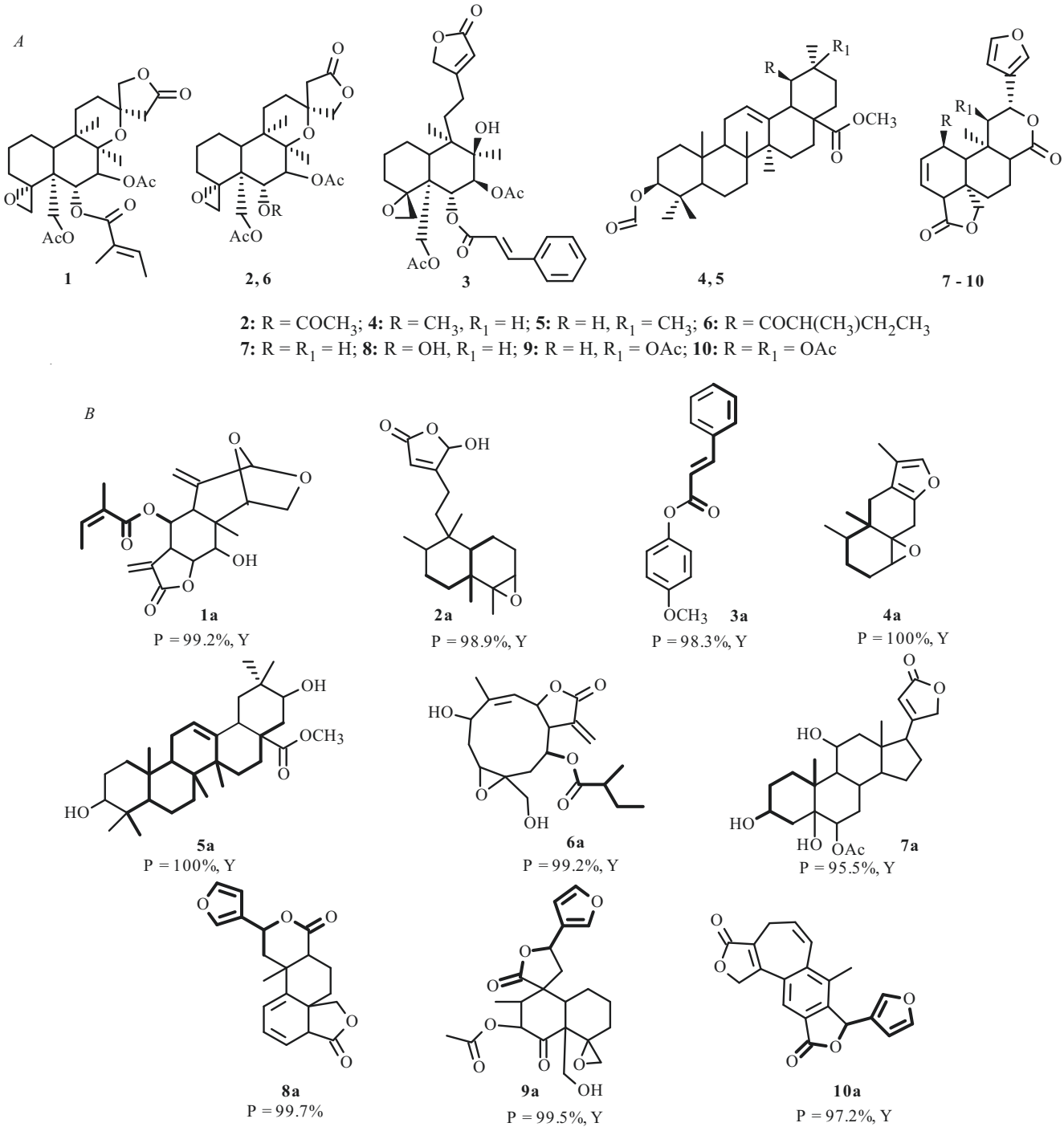

Fig. 1. Independent evaluation set results. In row (*A*) the so-called 'unknown' structures are shown; for them the compound name and spectrum source are given. In rows (*B*) are shown the corresponding substructure generated with highest precision, embedded into the reference structure (bold bonds); for them the precision and correctness (Y/N) are given. Compounds: scutalpin E (**1**) [4], scutalpin F (**2**) [4], scutorientalin E (**3**) [5], formyl methyl ursolate (**4**) [6], formyl methyl oleanolate (**5**) [6], scutalpin A (**6**) [7], salviarin (**7**) [8], splenolide A (**8**) [8], splenolide B (**9**) [8], and splendidin (**10**) [8].

A thorough look at the retrieved substructures reveals some specifics of the interpretive library search. First, if a side chain is present in the 'unknown' structure, this molecular part is inferred as a substructure with a high reliability, as is the case here with scutalpins A and E (2-methylbutanoic ester moiety) and splenolides A and B and splendidin (3-furanyl moiety). Second, as the algorithm expands the substructure within the magnetically nonequivalent reference structure′s atoms, the inferred substructure with the scutorientalin E spectrum has a 'half' benzene ring. Nevertheless, an experienced chemist will deduce that the inferred substructure is a cinnamic ester moiety. Third, if the reference and unknown compounds are

similar in structure, the inferred substructure is large. This result is encountered for formyl methyl oleanolate – the inferred substructure has 20 heavy atoms, which constitute 56% of all the query structure′s atoms. As can be seen, the query and reference structure differ only at two sites whose surrounding is excluded from the reference, thus remaining the atoms of the retrieved substructure.

The test with 100 spectra allows more reliable statistics to be obtained. Altogether, there were 87354 substructures retrieved by these 100 search sessions; 37488 (43%) of the substructures are correct. The size of all retrieved substructures varies from 6 to 42 (9.2 in average) and that of the correct ones also from 6 to 42 (but 8.8 on average). The correlation coefficient between the accuracy and size of the correct substructures is –0.08054 which is statistically significant: $t_{kr} = |–15.64| > 2.58 = t(0.01, 37486)$. That means that, on average, the higher the accuracy of the correct substructure, the smaller its size, thus indicating an adverse relation between them. The same relation is confirmed by the comparison of both average values (9.2 and 8.8) mentioned above. Despite this, the size of the 79 correct inferred first substructures – that have the highest precision – varies from 14 to 74% (33% on average) of that of the corresponding 'unknown' structure.

As a whole, these results turned out to be very satisfactory. They proved that the interpretive library search in a large database of fully-assigned ^{13}C NMR spectra of organic compounds can be used for structure elucidation of natural compounds. The designed reliability function for ranking the retrieved substructures performs pretty well, giving at first places reliable and informative substructures.

The program and the database could be freely downloaded from: http://kosnos.com/spectroscopy/iris/.

ACKNOWLEDGMENT

This work was supported by the Bulgarian National Science Fund, Contract DDWU02/37.

REFERENCES

1. N. Gray, *Computer-Assisted Structure Elucidation*, John Wiley, New York, 1986.
2. P. Penchev, K.-P. Schulz, and M. Munk, *J. Chem. Inf. Model*., **52**, 1513 (2012).
3. A. Korytko, K.-P. Schulz, M. Madison, and M. Munk, J. *Chem. Inf. Comput. Sci*., **43**, 1434 (2003).
4. P. Bozov, G. Papanov, and P. Malakov, *Phytochemistry*, **35**, 1285 (1995).
5. P. Malakov, P. Bozov, and G. Papanov, *Phytochemistry*, **46**, 587 (1997).
6. G. Papanov, P. Bozov, and P. Malakov, *Phytochemistry*, **31**, 1424 (1992).
7. P. Bozov, P. Malakov, G. Papanov, M. De La Torre, B. Rodrigues, and A. Perales, *Phytochemistry*, **34**, 453 (1993).
8. S. Merkova, P. Bozov, and I. Iliev, *Annuaire de L′Universite de Sofia "St. Climent Ohridski,"* Faculte de Chimie, **102/103**, 279 (2011).
9. H. Schriber and E. Pretsch, *J. Chem. Inform. Comput. Sci*., **37**, 879 (1997).