Interpretive search in a ^{13}C-NMR spectral library of plant compounds

S. Nachkova, S. Milenkova, P. Bozov and P. Penchev

Abstract: *An algorithm for interpretive library search in a database of ^{13}C NMR spectra is tested with a newly composed spectral library. Two hundred and ten spectra of compounds isolated from plants are searched in this library of 1000 spectra of natural compounds. A use is made of a specially designed rank that sorts the substructures according to their reliability. A kind of reliability function is derived from the rank values of the retrieved substructures. The obtained results are highly informative and can be used in the process of structure elucidation.*

Key words: *Interpretive library search; C-13 NMR spectral database.*

INTRODUCTION

Computer searching in spectral libraries of fully assigned ^{13}C NMR spectra of substructures or full structures is an indispensable part of the structure elucidation [1-2]. The library search in spectral databases has two main goals: identification of the unknown compound if its spectrum is among the reference spectra (the so called *identity search*), or obtaining a list of compounds whose spectra are most similar to that of the unknown (*similarity search*). As the spectrum reflects in great extend the structure of a compound, the result hitlist can be used for deriving some conclusions for the unknown's structure, and this is usually done through manual inspection of the hit structures by the chemist or with the aid of some computer algorithms as that of maximum common substructure [3]. Particularly, the ^{13}C NMR spectrum reflects the nature of the skeletal backbone of the organic compound, information not as easily derivable by other spectroscopic techniques.

When the library does not contain structures similar enough to the unknown one (i.e. library spectra are less similar to the unknown one) the structure inferences derived from the hitlist can be less informative or even incorrect. In this case, the approach of *interpretive library search* can be applied more successfully [4-5]. This approach derives from the proposition that if an unknown and a reference spectrum share a subspectrum in common, the substructure assigned to the reference subspectrum is also present in the unknown. In our approach, each such substructure is a connected part of the reference structure, and each carbon atom in the substructure has at least one assigned signal of the unknown. The practice shows that some of substructures retrieved by the interpretive library search are incorrect (i.e. not present in the unknown's structure), that is why a preliminary derived reliability function is needed to evaluate the substructure correctness. As a final result, the obtained plausible substructures are ordered by their decreasing prediction accuracy (reliability), and part of them can be selected for use in the structure generation process [1].

Such multivariate function is obtained for a library of 38 225 ^{13}C NMR spectra of organic and natural compounds [5] and it has been proved that the interpretive search produces correct and highly informative results [5-6]. The function uses various parameters characterizing each of the inferred substructures [5] but some of the variables depend on the library size. Because of that, when searching in another library, the estimated prediction accuracy does not differentiate between correct and incorrect substructures. Obviously, that hinders the application of the method with various spectral libraries. That fact is shown in the present paper in which a newly composed spectral library of 1000 spectra of plant compounds is used. To overcome the problem, the multivariate function is replaced by a use of a specially designed rank that sorts the substructures according to their reliability. The rank uses some of the substructure variables described in [5].

METHODS

The method for interpretative library search is implemented into Windows-based user-friendly program, called *InferCNMR* [5]. The program input consists of chemical shift and multiplicity of each signal in the ^{13}C-NMR spectrum of the unknown compound, as well as the molecular formula of the organic compound. The retrieved substructures are presented embedded into the reference structures and, in this way, are explicitly defined in terms of atom type, hydrogen multiplicity and bond type. The substructures are sorted according to their accuracy (calculated by a multivariate function [5]) or by their rank. Additionally, a reliability function is derived from the rank values of the retrieved substructures. The parameters which restrict the search algorithm are the tolerance of signal matching (*Tol*) in ppm and minimum number of the carbon atoms in the inferred (retrieved) substructures (*m.n.c.*); for this study the former is set between 0.0 – 2.0 ppm with a step of 0.1 ppm and the latter to 6.

In this study, it is used a newly composed spectral library of 1000 fully assigned ^{13}C NMR spectra of plant compounds. The structures of the compounds are represented as 2D connectivity tables with x and y atom coordinates, and each carbon atom of the reference compound has a single chemical shift assigned to it. The spectra are taken from Phytochemistry journal (years 2002-2006, volumes 61-67). The library header information includes the compound names, used solvents and bibliographic source data. The library is named PHYCHEM.

Two sets, the so-called *learning set* (LS) and *test set* (TS), each of 100 fully-assigned ^{13}C NMR spectra, are composed. The spectra are taken from Phytochemistry journal (years 2001-2002, volumes 58-60) and the corresponding compounds are also isolated from plants. As we started a collaboration with one of the coauthors (P.B.) who has a 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. [7-11]. These compounds are well representative to check the efficiency of the interpretive search and it has been proved previously [6] that the interpretive search together with multivariate reliability function produces very reliable and highly informative results for them.

All these 210 spectra in the learning, test and validation sets are not contained in the PHYCHEM library. They are searched in it with the algorithm described earlier [5].

RESULTS AND DISCUSSION

The 100 spectra from learning set produce hitlists of size ranging from 3 to 858 substructures (151 in average). The hitlist size for 100 test set spectra varies from 8 to 800 substructures (134 in average).

To estimate the performance of the reliability function, the substructures from the result hitlists of the learning and test sets are separately gathered into two lists. The latter are composed of 15 136 and 13 424 substructures, respectively, and the correct substructures in them are 47.5% and 46.1%. All the substructures are processed by the previously developed reliability function and the two lists are sorted separately according to the substructure accuracy. Then, estimates are calculated for the accuracy threshold values of 90%, 95 and 99%. The corresponding 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 in Table 1 and as can be seen they are far lower than the corresponding threshold values. This means that *the reliability function works poorly*. This is in fact not so surprising because the function uses variables (parameters) that depend on the library size, and the present reliability function was assembled with the aid of library search in 38 225 spectra [5].

Table 1. Estimation of substructure accuracy.

Set[a]	A_T = 90 % [b]			A_T = 95 %			A_T = 99 %		
	n.c.s.[d]	n.i.s.[d]	A_E, %[c]	n.c.s.	n.i.s.	A_E, %	n.c.s.	n.i.s.	A_E, %
LS	4 470	3 168	58.5	1 966	827	70.4	669	231	74.3
TS	3 750	2 792	57.3	1 592	697	69.6	544	222	71.0

a) LS = learning set; TS = test set. b) A_T = threshold accuracy. c) A_E = estimated accuracy. d) n.c.s. and n.i.s. = number of correct and incorrect substructures, correspondingly.

In order to solve the above mentioned problem, a new function can be built, but this requires comprehensive statistics and is a very time- and resource-consuming procedure. On the other side, the authors intend to extend the library with additional spectra: up to now, other 500 spectra are gathered from the same journal. That is why the function must be carefully prepared and it has to use only substructure parameters (original or derived from them) that are independent of the library size.

The alternative of a reliability function is a usage of some sort of rank that is composed from substructure parameters. The rank just sorts the correct substructures in the beginning of the list and is not related to any output accuracy value that depends on the portability of the reliability function. The composition of a rank is faster, does not need any comprehensive statistics, and its usefulness can be proved just with the aid of small representative sets. Additionally, a sort of reliability function can be derived very fast from the rank value of all substructures in the sorted list.

All substructure parameters from Tables 1 and 2 in [5], 48 in number, are screened for their discriminative power between the correct and incorrect substructures in the learning set results. As the most discriminating are found the four parameters listed in Table 2. (Keep in mind that some of the parameter short names are changed when compared with the previous paper [5].) Additionally, the tolerance (*Tol*) used by signal match is a very good parameter because more correct substructures are generated by smaller tolerance. It has a strong correlation with the parameter *sRMSD* and that is why only one of them can be used in the rank.

Table 2. Substructure parameters used in the rank.

sRS: the number of all substructures retrieved in a search at a given tolerance.
sSO: the number of occurrences of a particular substructure produced in a search at a given tolerance in all *sRS* substructures.
sLO: the number of reference compounds in the library containing the particular substructure produced by a search at a given tolerance.
sFV: number of free valences in the substructure
sRMSD: minimum root mean standard deviation of substructure matched signals

Several combinations of parameters from Table 2 together with other parameters are checked and it is found as the most effective for LS spectra the rank given by Eq. 1

$$Rank = f\frac{sLO \cdot sSO}{LibSz} * \frac{sNA}{sFV} + (1-f)(2-Tol), \qquad \text{(Eq. 1)}$$

where *LibSz* is the size of the spectral library (1000 in our case), *sNA* is the number of atoms in the substructure, and *f* is a user-adjustable factor (weight) ranging between 0.0 and 1.0.

The parameters are placed in the numerator or denominator, or with plus or minus sign, depending on their positive or negative discrimination: the higher rank value the more correct is estimated the substructure. The parameter *sNA* is placed in the numerator with the aim of “pushing” larger substructures to the top of the sorted list; in [5] it has been used the reverse ratio, *sFV/ sNA*, as a reliability function variable.

A new reliability function is built by the use of rank values of all substructures in the learning set. The procedure is the same as those described in [5] with the only difference that the rank value replaces the artificial neural network's output value. In the same way as in [5] the recalls at accuracy of 90%, 95% and 99% (R_{90}, R_{95} and R_{99}) are calculated. The factor f in Eq. 1 is varied between 0.0 and 1.0 in ten steps and a maximum is searched of the three mentioned recalls. The curves f vs. R_{90}, f vs. R_{95} and f vs. R_{99} look very similar in both sets, LS and TS, and all six have maximum at f = 0.6. The corresponding calculated recalls for both sets are very close and high; for LS: R_{90} = 45.2%, R_{95} = 37.5% and R_{99} = 25.6%; for TS: R_{90} = 41.8%, R_{95} = 34.0% and R_{99} = 25.3%.

The rank is applied to the LS and TS spectra: in 95 out of 100 hitlists in LS sorted by the rank the first ranked substructure is correct; in TS there are 92 out of 100 sorted hitlists with first substructure correct.

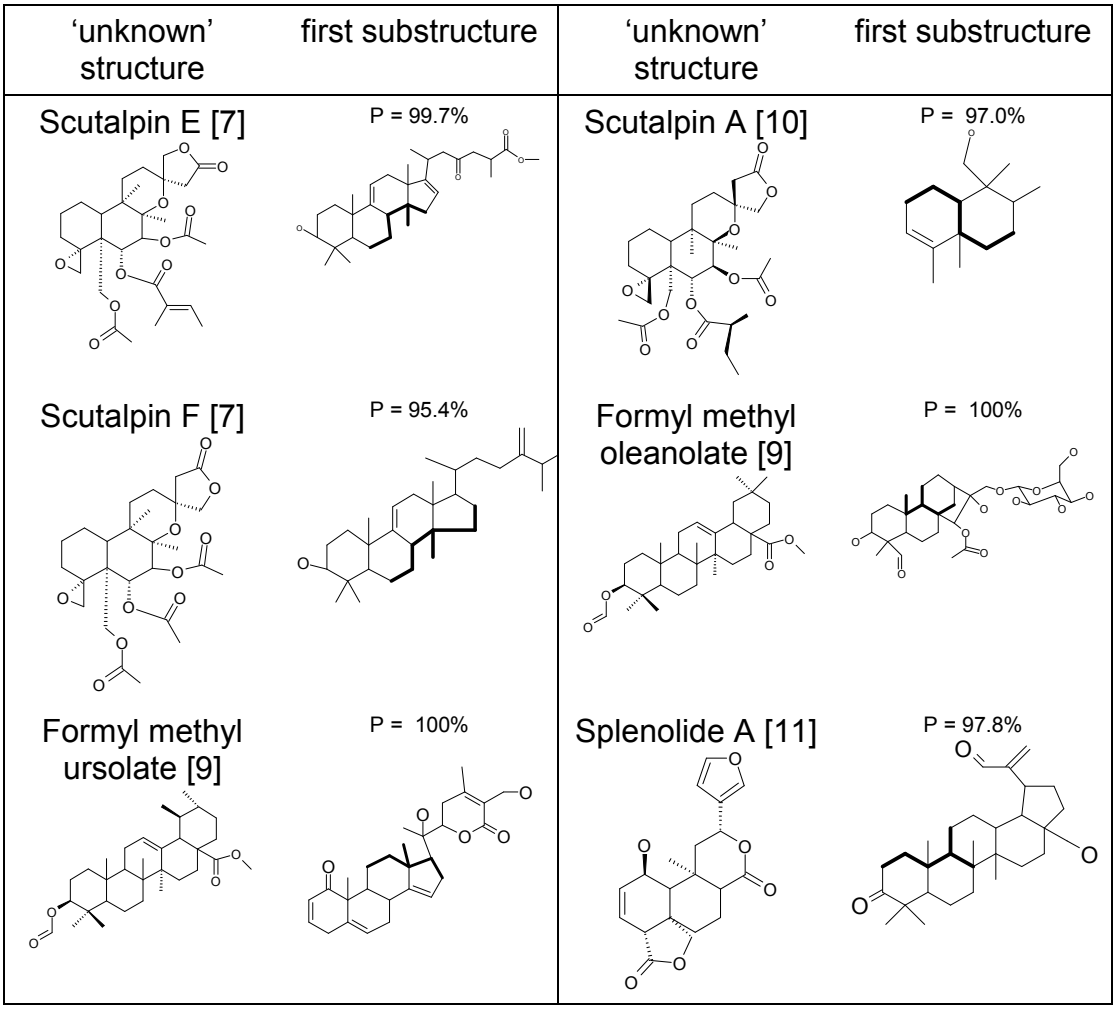

Figure 1. Validation set results which are correct.

The ten spectra from validation set predicted from 37 to 680 substructures (194 in average). In four of the cases the first substructure is incorrect. The six correct first substructures are given together with query structures in Figure 1; the former are represented embedded into the corresponding reference compound structure.

CONCLUSIONS

As a whole, these results turned out to be very satisfactory for test set compounds; the very poor performance for the ten validation compounds can not be explained and this pose a great challenge to the practical application of the method to the compounds similar to them. One possible explanation is that the validation set is a quite small one for the statistics to be reliable enough.

REFERENCES

[1] Gray, N.; Computer-Assisted Structure Elucidation. John Wiley, 1986.

[2] Munk, M. Computer-Based Structure Determination: Then and Now. J. Chem. Inf. Comput. Sci., 1998, 38, 997-1009.

[3] Chen, L.; W. Robien; Application of the Maximum Common Substructure Algorithm to Automatic Interpretation of ^{13}C-NMR Spectra. J. Chem. Inf. Comput. Sci., 1994, 34, 934-941.

[4] Shelley, C.; Munk, M. Computer Prediction of Substructures from Carbon-13 Nuclear Magnetic Resonance Spectra. Anal. Chem. 1982, 54, 516-521.

[5] Penchev, P.; Schulz K.-P., Munk M.; INFERCNMR: A ^{13}C NMR Interpretive Library Search System. J. Chem. Inf. Model., 2012, 52, 1513-1528.

[6] Nachkova, S.; S. Milenkova, P. Bozov and P. Penchev; Interpretive Library Search in a Database of Fully Assigned ^{13}C-NMR Spectra. Sent for publication in Scientific researches of the Union of Scientists in Bulgaria- Plovdiv, series B, 2012.

[7] Bozov, P.; Papanov G.; Malakov P.; Neo-Clerodane Diterpenoids from Scutellaria Alpina. Phytochem., 1995, 35, 1285-1288.

[8] Malakov, P.; Bozov P., Papanov G.; Neo-Clerodane Diterpenoid from Scutellaria Orientalis Subsp. Pinnatifida. Phytochem., 1997, 46, 587-589.

[9] Papanov, G.; Bozov P., Malakov P.; Triterpenoids from Lavandula Spica. Phytochem., 1992, 31, 1424-1426.

[10] Bozov, P.; Malakov P.; Papanov G.; De La Torre M.; Rodrigues B., Perales A.; Scutalpin A, A Neo-Clerodane Diterpene from Scutellaria Alpina. Phytochem., 1993, 34, 453-456.

[11] Merkova, S.; Bozov P.; Iliev I.; Chemical Constituents of the Aerial Parts of Salvia Splendenes. Annuaire de L'Universite de Sofia "St. Climent Ohridski", Faculte de Chimie, 2011, 102/103, 279-284.

Acknowledgment

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

About the authors:

Correspondence author: Stefka Nachkova, Assist. Prof., Chemical Faculty, University of Plovdiv, E-mail: stefka@uni-plovdiv.net.

Sofia Milenkova, Student, Math High School "Acad. Kiril Popov", Plovdiv.

Petko Bozov, Assist. Prof., PhD, Biological Faculty, University of Plovdiv.

Plamen Penchev, Assoc. Prof., PhD, Faculty of Chemistry, University of Plovdiv.

This paper has been reviewed