IMPROVED REALISATION OF MAXIMUM COMMON SUBSTRUCTURE CONCEPT FOR STRUCTURE ELUCIDATION

N. Kochev [a], P. Penchev [a] *, G. Andreev [a], K. Varmuza [b]

[a] Department of Analytical Chemistry, University of Plovdiv, 4000 Plovdiv, Bulgaria

[b] Vienna University of Technology, Laboratory for Chemometrics, Institute of Food Chemistry and Food Technology, Getreidemarkt 9/160, A-1060 Vienna, Austria

Abstract

An improved realisation of an algorithm for deriving a set of maximum common substructures (MCSs) is described. The MCSs set is obtained from the structures of a hitlist that results from a library search of an infrared spectrum in a collection of 13484 spectra. The algorithm calculates more than one MCS of a given pair of structures if existing and gives additionally common substructures (CSs) up to a predefined level for their size. A new ranking of the MCSs and CSs is proposed that is more efficient for the purposes of structure elucidation of organic compounds than previous ones.

Introduction

The library search in computer collections of spectra is a routine technique for the identification of organic compounds [1]. If the query compound has a spectrum in the collection the so called *identity* search is performed [2] which often gives the corresponding reference spectrum of the compound on the first or second position in the hitlist; otherwise a correct compound identification is impossible. The hypothesis usually applied by spectroscopist in this case is that the similar spectra indicate similar chemical structures. Based on this assumption the spectroscopist evaluates structures and spectra in the hitlist with the aim to construct plausible candidates for the unknown molecular structure or at least to get hints which substructures may be present or absent in it. Related automatic procedures are the *k-nearest neighbours* method and approaches based on the concept of maximum common substructure (MCS) [3-5].

In this paper we describe an improved realisation of the MCS concept that is implemented in the program IRSS [6] containing a newly developed algorithm [7] and an example that demonstrate the achieved effectiveness.

Method

The maximum common substructure of two chemical structures represented as connectivity tables is defined as the largest connected substructure that is present in the two given structures. The MCS of a set of n hitlist structures is usually very small and insignificant. In this case substructures characteristic for the hitlist compounds can be derived by comparing all pairs of the hitlist structures [4,5]. First, the MCS for each of the $n(n\text{-}1)/2$ possible pairs of hitlist structures is determined, and then for each obtained MCS_i the number of occurrences, n_i (frequency), in the n hitlist structures is counted. Finally the MCSs are ordered by their decreasing ranking weight R_i as defined in equation (1) [4,5]. The ranking considers both the frequency and the size of the substructures (which is given by the number of non-hydrogen atoms).

* Corresponding author, e-mail address: plamen@argon.acad.bg

$$R_i = (1 - f)\, n_i / n + f A_i / A_{max}; \qquad (1)$$

A_i is the number of non-hydrogen atoms in MCS_i; A_{max} is the maximum number of non-hydrogen atoms in the n investigated molecular structures; f is a user-adjustable factor ranging between 0 and 1.

This ranking does not consider the positions of the hitlist structures which contain the i^{th} MCS. For example, it can happen that only the last 40 from all 50 hitlist structures contain a given MCS thus giving a high rank because of a large n_i; a case which sounds inadequately for an experience spectroscopist. To correct this "position" problem we proposed the use of a changed rank given by equation (2).

$$R_i = (1 - f)\, \Sigma 2\,(n + 1 - p_i) / n(n+1) + f A_i / A_{max}; \qquad (2)$$

where p_i is position of the hitlist structure in the hitlist, and the other parameters have the same meaning as those in equation (1). $n + 1$ in the numerator ensures the last position in the hitlist, n, to be counted with a weight of 1.

The newly proposed rank is a generalised case of the old one. If the hitlist structures that contain a particular MCS are just in the middle of the hitlist, it can be easily proved that the new rank gives exactly the value of the old one.

In both equations, if f is set to zero, the sizes of the derived MCSs do not affect the ranking; in this case the rank from equation (1) is determined entirely by the occurrence of corresponding MCS, and the rank from equation (2) entirely by the positions of the hits containing this MCS. If f is 1 only the size is considered.

When spectroscopists make inspection of a hitlist, they probably never think in terms of "maximum common substructures present in the hitlist structures". They rather derive structural fragments smaller than particular MCSs but which are more reasonable from the spectroscopic point of view. These smaller fragments, called by us *common substructures* (CSs), are also more frequent than MCSs among the hitlist structures, and this makes them a promising extension of the MCS set. In the newly developed implementation of the MCS algorithm the user can set the allowed maximal difference, L, between number of atoms in a MCS and the number of atoms in additionally derived CSs. For example, if L is set to 3, all CSs having up to three atoms less than the MCS will be given as characteristic substructures for the corresponding pair of hitlist structures; however, only if they are not contained in the MCS.

The isomorphism of substructures and the determination of MCSs and CSs is controlled by several parameters whose influence was studied before [5]. Their optimal values are used by us in the present study. Only the influence of the new rank and the additional CSs on the quality of the results was explored, see *Results and Discussion*.

Results and Discussion

To evaluate the effectiveness of the proposed new realization of the MCS concept, 10 compounds having reference spectra in the library were regarded as 'unknown' and their IR spectra were searched in 13484 IR spectra of the *SpecInfo* data base [4,5]. For comparison the 10 compounds were those used before [5]. The resulting hitlists containing 51 compounds were obtained by *correlation coefficient* full-curve spectral searches. The first entry of each hitlist (the corresponding reference spectrum) was removed giving the hitlist size equal to 50. The structures of each hitlist were processed by five different MCS algorithms described in Table 2 and designated by letters A, B, C, D, and E. Parameter L determines whether the CSs are output together with the MCSs (L > 0) or only the MCSs are

presented in the results (L = 0). The two rankings defined in equations (1) and (2) were used in the algorithms as indicated in the last row of Table 2. Letter *A* designates the algorithm, previously described and implemented in the software ToSiM [8]. The algorithm *B* is analogous to *A* but implemented in the software IRSS.

Table 1. Ten compounds used as unknowns. Size of molecules is measured by the number of non-hydrogen atoms; T_{18} is the number of first correct MCSs from all 18 MCSs obtained.

No	Compound name <CAS registry number>	Size	T_{18} A	B	C	D	E
1	Butylamine <109-73-9>	5	1	1	1	1	1
2	1-Pentanol, 5-bromo <34626-51-2>	7	5	5	5	5	5
3	Tetrahydropyran-4-methanol <14774-37-9>	8	5	5	5	8	8
4	1,3-Dioxolane-4-methanol, 2-vinyl- <4313-32-0>	9	18	17	18	15	17
5	Benzene, 1-methoxy-3-(1-propenyl)- <20112-91-8>	11	3	3	3	4	4
6	1-Amino-naphthalene <134-32-7>	11	1	2	2	6	6
7	1-Hexanone, 1-(3-pyridinyl)- <81418-03-3>	13	4	5	4	5	5
8	4,7-Methano-1H-indene, 6-(diethoxymethyl)-3A,4,5,6,7,7A-hexane <67633-93-6>	17	18	18	18	18	18
9	1H-1,2,4-Triazole, 1-[2-[3-(2-fluoro phenyl)-2-methylpropoxy]-3,3-dimethyl-1-butenyl]- <101975-44-4>	23	6	7	7	7	9
10	Proline, 1-benzoyl-4-(2,5-dichloro benzoyl)-3-(1,1-dimethylethyl)-5-phenyl, ethyl ester <103430-68-8>	38	18	18	18	18	18

Table 2. Applied MCS algorithms, designated with letters A, B, C, D, and E.

algorithm	A	B	C	D	E
software used	ToSiM	IRSS	IRSS	IRSS	IRSS
CS level, L	0	0	10	0	10
ranking	equation (1)	equation (1)	equation (1)	equation (2)	equation (2)

As there is no criterion capable to decide which of the obtained MCSs are correct when a real spectrum interpretation is performed, we have chosen the number of first correct MCSs, T_{18}, (Table 1) as a figure of merit for algorithms' effectiveness. Our experience show that the user considers MCSs for structure elucidation beginning with the first MCS,

followed by the second and so on. That is why it is important to have as many as possible correct MCS at the beginning of their ranked list. The last five columns in Table 1 summarise the results from tests performed with the five MCS algorithms. The algorithms were statistically compared with the Wilcoxon matched-pairs test [9] applied for the T_{18} values. This test was chosen because the normality of data distribution is doubtful and only a small sample size (10) is available.

Algorithm A versus algorithm B

As it was mentioned, algorithms A and B are analogous implementations in two different programs [6,8]. The results in Table 1 show only a few small differences between the performances; the Wilcoxon test does not give a significant difference (94% probability for the zero hypothesis). The main reason for the differences is the fact that algorithm B usually gives more than one MCS for a hit pair, if existing. The additionally found MCSs proved to be highly characteristic for the hitlists structures; they add new relevant structural information that improves the MCS method effectiveness. Another reason to compare both algorithms is the fact that algorithm B (software IRSS) is a newly developed and its correctness and performance has to be confirmed by comparison with the previously realized software ToSim. In this work algorithm B is regarded as the main MCS algorithm which is compared with other ones.

Algorithm C versus algorithm B

Comparison between algorithms C and B does not show a significant statistical differences, Table 1. Algorithm C combines the full MCS search with an additional calculation of CSs down to 10 atoms smaller than a MCS (L=10). It appeared that nearly half of the additional CSs are not contained in the query structure. One of the reasons for this is that they are obtained from pairs of hitlist structures which are at the end positions in the hitlist. From this result one can expect an improvement when using the new ranking defined in equation (2).

Algorithm D versus algorithm B

Comparison of algorithms B and D observes the effectiveness of the new proposed ranking, equation (2), and its influence upon the generated characteristic structures. The Wilcoxon test gives a probability of 62% for the zero hypothesis; however, D yields better results than B in 3 cases while in the 6 others the results are identical. The improvements observed by algorithm D was expected as it is mentioned in the introduction of the new ranking which takes into account not only the frequency of the found MCS among the hit structures but also considers their positions in the hitlist as well. These positions reflect strongly the similarities between an unknown spectrum and the reference spectra in the hitlist.

Algorithm E versus algorithm B

Comparison of the T_{18} values for B and E algorithms from Table 1 shows additional improvement of the MCS algorithm. There is no negative difference $T_{18}(E) - T_{18}(E)$, and the statistical significance of the difference is 91%. Algorithm E is an approach combining the new ranking (equation 2) and the generation of additional common structures along with the generated MCSs (L =10). As was mentioned before these additional CSs are not always correct hints for the query structure but ranking them with equation (2) puts them at the beginning of the ranked list of substructures. This fact makes them more reliable from the spectroscopic point of view. Another question not answered in this work is that CSs are, as a

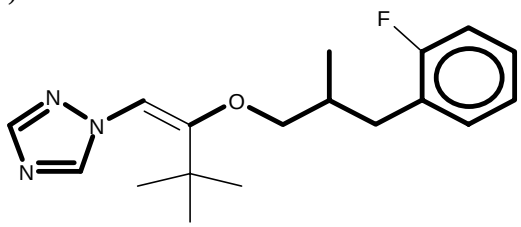
rule, smaller fragments than MCSs thus decreasing their significance in the structure elucidation process. Additional criteria would be required to evaluate fully the usage of them.

Example of Application

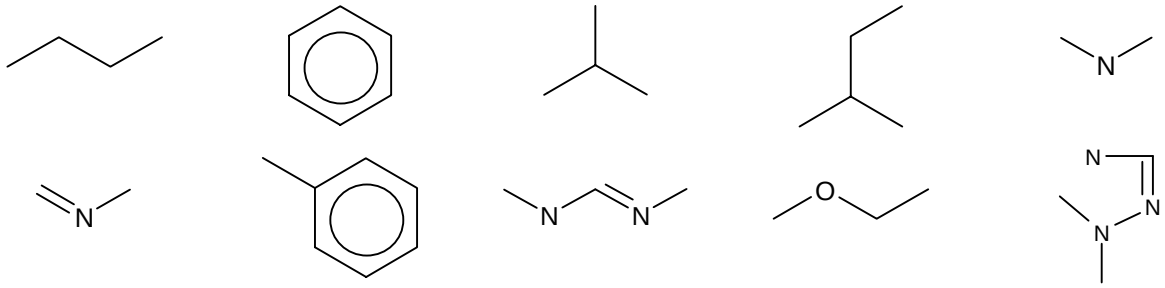
The IR spectrum of the compound in figure 1a is searched in a collection of 13484 spectra. The hitlist is processed with algorithm E. The results are given in figure 1b. As can be seen nearly the whole structure (figure 1a) is covered by the predicted fragments.

Figure 1. a) The unknown molecular structure with the predicted structural fragments in it (marked with bold bonds).
b) The first ten MCSs predicted by algorithm E. The hitlist with 50 entries is obtained by correlation coefficient full-curve search.

a)



b)



References

1. H.J. Luinge; *Automated interpretation of vibrational spectra*. Vib. Spectrosc. **1**, 3-18, 1990.
2. J.T.Clerc; Automated spectra interpretation and library search systems. In*: Computer-enhanced analytical spectroscopy.* H.L.C. Meuzelaar and T.L. Isenhour, (Eds.), Plenum, New York, 1987, p. 145-162.
3. H. Scsibrany and K. Varmuza; *Common substructures in groups of compounds exhibiting similar mass spectra*. Fresenius J. Anal. Chem. **344**, 220-222, 1992.
4. K. Varmuza, P.N. Penchev and H. Scsibrany; *Maximum common substructures of organic compounds exhibiting similar infrared spectra.* J. Chem. Inf. Comput. Sci. **38**, 420-427, 1998.
5. P.N. Penchev and K. Varmuza*; Characteristic substructures in sets of organic compounds with similar infrared spectra*. Computers & Chemistry, **25**, 231–237, 2001.
6. P.N. Penchev, N.T. Kochev and G.N. Andreev; *IRSS: A Programme System for Infrared Library Search*. Compt. Rend. Acad. Bulg. Sci., **51**, 67-70, 1998.
7. N.T. Kochev, P.N. Penchev; *Implementation of an Algorithm for Maximum Common Substructure Determination*. 2001 (to be published).
8. H. Scsibrany and K. Varmuza; ToSiM: PC-Software for the Investigation of Topological Similarities in Molecules. In *Software Development in Chemistry;* Jochum, C. (Ed.); Gesellschaft Deutscher Chemiker. Frankfurt am Main, Vol. **8**, 1994, p. 235-249.
9. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michote, L. Kaufman; *Chemometrics: A Textbook*. Elsevier, Amsterdam, 1988.