Evaluation of Hitlists from IR Library Searches by the Concept of Maximum Common Substructures

Kurt VARMUZA,[1†] Nikolay T. KOCHEV,[2] and Plamen N. PENCHEV [2]

[1†] *Laboratory for Chemometrics, Institute of Food Chemistry, Vienna University of Technology, Getreidemarkt 9/160, A-1060 Vienna, Austria (E-mail: kvarmuza@email.tuwien.ac.at)*

[2] *Department of Analytical Chemistry, University of Plovdiv, Tsar Assen Street 24, BG-4000 Plovdiv, Bulgaria*

Hitlists from spectral similarity searches with IR spectra are used to generate lists of substructures that are characteristic for the structures of the unknowns. The applied method searches for maximum common substructures in all pairs of hitlits structures and ranks these substructures by a newly defined criterion. Examples demonstrate the advantage of the new ranking criterion as well as potentials and limits of the method.

(Received on August 9, 2001, Accepted on September 15, 2001)

Information about the unknown chemical structure of an organic compound can be obtained by comparing the infrared (IR) spectrum with reference spectra from a spectral library. The resulting hitlist contains compounds exhibiting the most similar spectra. If the unknown is present in the library then the correct answer often appears among the first hits and can be identified easily by considering additional restrictions such as volatility or origin of the investigated sample as well as results from other spectroscopic methods.[1-4] However, if the unknown is not contained in the spectral library a more detailed evaluation of the hitlist structures and spectra is necessary.[5, 6] This data interpretation is usually done by the spectroscopist. Computer-based methods - usually supposing that similar spectra indicate similar chemical structures - can support this work. Chemometric methods such as principal component analysis, PLS[7, 8] or multivariate classification can provide an insight into spectra-structure relationships and can give hints about the presence of particular substructures in the unknown.[9-13] Simulation of IR spectra[14] from given candidate structures can reduce the number of possible solutions.

This work deals with a method to extract relevant substructures from the molecular structures of hitlist compounds. A substructure is considered to be relevant if it is contained in the structure of the unknown and is helpful for building candidate structures - for instance by automatic isomer generation.[15, 16] For mass spectrometry (MS), methods have been developed based on a statistical evaluation of hitlist structures.[17] For MS,[18, 19] IR[20-23] and C-NMR,[24] methods based on the concept of maximum common substructures (MCS) have been described. This paper presents a summary of this method for IR and a new, more efficient ranking procedure for the found substructures.

Method

Spectral similarity search

The IR spectrum of a compound considered as unknown is compared with all reference spectra of a library. Based on previous investigations[21] the similarity $S(u,r)$ between the unknown (u) and a reference spectrum (r) is calculated from the correlation coefficient, R, of the absorbances, Eqs.(1 and 2).

$$R = \Sigma\,[A(r,k)\,A(u,k)] \,/\, [(\Sigma\,A^2(r,k)\,\Sigma\,A^2(u,k)]^{0.5} \quad (1)$$

$$S(u,r) = 999\,(R + 1)\,/2 \quad (2)$$

$A(u,k)$ and $A(r,k)$ are mean-centered absorbance units in wave number interval k, in the unknown and the reference spectrum, respectively. Because of mean-centering the sum of $A(.,k)$ is zero in each spectrum. Normalization used in Eq.(1) gives similarity values between 0 and 999. All examples have been calculated with 801 wave number intervals equally distributed between 500 and 3700 cm^{-1} and n = 50 hits have been considered for the evaluation of hitlist structures.

Maximum common substructure

The maximum common substructure (MCS) of two structures is the largest connected substructure present in both.[25-29] The size of a molecular structure or a MCS is measured by the number of non-hydrogen atoms.[30] In general it is possible that more than one MCS exist for a given pair of structures; the software used for this work determines only one of them. A newly developed software is in principle capable to determine all MCSs, which however often requires a very long computation time.[31] Examples for MCSs are shown in Fig. 1.

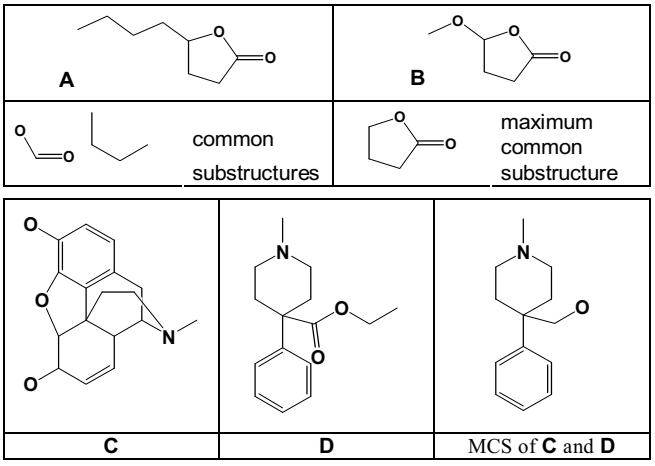

Fig. 1 Examples for maximum common substructures (MCSs). While the MCS for structures **A** and **B** is evident; the solution for **C** (morphine) and **D** (pethidine) is better searched by software.

A MCS is determined for each possible pair of hitlist structures. For a hitlist with n structures $n(n\text{-}1)/2$ pairs exist which is 1225 for $n = 50$. Duplicate MCSs are removed and the remaining MCSs are ranked with the purpose to obtain the relevant substructures with high rankings. The ranking criterion used up to now considers the size of a MCS and its frequency (that means in how many hitlist structures it is present), Eq.(3).[21]

$$R1(i) = (1 - f)\, n(i)/n + f\, a(i)/a(max) \quad (3)$$

$R1(i)$ is the ranking of MCS i; $n(i)$ is the number of hitlist structures containing MCS i as a substructure (frequency); $a(i)$ is the size of MCS i; $a(max)$ is the maximum size of the n investigated hitlist structures. The factor f (0 to 1) determines the mutual influence of size and frequency of a MCS on its ranking; extensive tests showed that a value of 0.3 is optimal.[20] The maximum value of 1 is reached for $R1$ if $n(i) = n$ and $a(i) = a(max)$.

A newly defined ranking criterion additionally considers in which of the hitlist structures a MCS is contained. It is evident that presence of a MCS in the first hits (which are the reference spectra with highest spectral similarity to the unknown) may count more than presence only in the last hits. The term $n(i)/n$ in Eq.(3) is replaced by one that realizes this idea, guiding to a more general definition for the ranking criterion, Eq.(4).

$$R2(i) = (1 - f)\, \{[n(i)(n+1) - \Sigma h(k,i)\,] / [n(n+1)/2]\} + f\, a(i)/a(max) \quad \text{with } k = 1 \text{ to } n(i) \quad (4)$$

$h(k,i)$ is the hitlist position of MCS i in its k-th occurrence. For example if MCS i is part of the hitlist structures 1, 3 and 4, then $n(i) = 3$, $h(1,i) = 1$, $h(2,i) = 3$, and $h(3,i) = 4$. The first term in Eq.(4) was constructed in order to obtain the maximum value 1 if $n(i) = n$ as in Eq.(3). Note that $n(n+1)/2$ is the sum of numbers 1 to n; if MCS i is present in all hitlist structures then $\Sigma h(k,i)$ is equal to this sum. If the $n(i)$ hits containing the MCS i are exactly in the mid positions of the hitlist then $R2$ is equal to $R1$ (see example below).

In Table 1 the two ranking criteria are compared using four hypothetical MCSs. Parameters fixed were $n = 10$, $a(i) = 5$, $a(max) = 7$, $f = 0.3$. MCS 1 and 2 have equal frequency $n(1) = n(2) = 5$; MCS 1 occurs in hits 1 to 5 and MCS 2 in hits 6 to 10. Ranking criterion $R1$ is equal for both MCSs because positions in the hitlist are not considered. Criterion $R2$, however, is larger for MCS 1 reflecting the occurrences in the first hits. MCS 3 has a frequency $n(3) = 7$ and it is part of the hitlist structures 4 to 10. $R1$ evaluates this MCS as the best because the frequency is highest. More reasonable is the result obtained by $R2$ because MCS 3 has a lower ranking than MCS 1 reflecting that MCS 3 is not part of the first three hitlist structures. For MCS 4 the hitlist structures containing this substructure are in the mid positions of the hitlist; in this case $R1$ and $R2$ are equal.

In real examples the highest ranked 10 to 20 MCSs are used as a set of substructures to characterize the molecular structure of the unknown. Extensive tests showed that substructures found by this method are mostly present in the structure of the unknown. However, exploitation of the resulting substructure set usually requires the evaluation by a spectroscopist.

The MCS approach is not limited by a pre-defined set of substructures but is self-adapting to the type and complexity of the molecular structures contained in the hitlist. Frequencies and sizes of the found substructures are measures for the applicability of the method to a particular problem. If only small MCSs are found and their frequencies are low then one may conclude that the spectral similarity search did not find sufficient hits with structures similar to that of the unknown. An application of the MCS approach to structure sets obtained by other types of database searches seems promising.

Table 1 Comparison of ranking criteria $R1$ and $R2$; $n = 10$, $a(i) = 5$, $a(max) = 7$, $f = 0.3$.

i	$n(i)$	$h(k.,i)$, $k = 1$ to $n(i)$	$R1(i)$	$R2(i)$
1	5	1, 2, 3, 4, 5	0.564	0.723
2	5	6, 7, 8, 9, 10	0.564	0.405
3	7	4, 5, 6, 7, 8, 9, 10	0.704	0.571
4	6	3, 4, 5, 6, 7, 8	0.634	0.634

Evaluation of characteristic substructure sets

A substructure derived from the hitlist structures is considered to be correct if it is contained in the structure of the unknown, otherwise it is considered to be wrong. For test compounds the obtained sets with characteristic substructures can be judged by a simple criterion, T, "top correct". T is defined as the number of substructures (MCSs) at the top of the ranked list that are all correct. For example, if the first five substructures are correct and the 6th is wrong, $T = 5$; if the first substructure is wrong then $T = 0$ independent from the correctness of all other substructures.

This criterion meets our experience that users primarily look at the top of the MCS list because for real unknowns no criteria are available the indicate whether a substructure is correct or not. It is therefore important to have as many as possible correct MCSs at the beginning of the ranked list.

Experimental

Data and software

The IR database used consisted of 13484 spectra together with the corresponding chemical structures; it is part of the SpecInfo database system.[32] The software used was IRSS[33] and ToSiM.[30] All work was done on personal computers running under Microsoft Windows 98, NT or 2000.

Comparison of ranking criteria

A random sample of 100 spectra was selected from the database. For each spectrum a hitlist with $n = 50$ entries was produced using the similarity criterion from Eq. (2); the spectrum used as unknown was excluded from the hitlist. From the hitlist structures MCSs were searched according to the method described above. The resulting MCSs were ranked either by criterion $R1$, Eq.(3) or $R2$, Eq.(4) with $f = 0.3$; the 50 highest ranked MCSs build a set of characteristic substructures. The substructure lists were evaluated by criterion T; high values for T are obtained if many substructures at the top of the list are all correct.

Results for T are in the range 0 to 50 for each of the ranking criteria. Using the new ranking criterion ($R2$) at least the first 16 MCSs were all correct in 26 test compounds; for the old ranking criterion ($R1$) in 21 test compounds. For nine and eight test compounds (applying $R2$ and $R1$ respectively) none of the found MCSs were correct. Histogram data in Table 2 show that almost 40% of the test compounds give substructure lists with 11 or more substructures at the top of the list being correct. In more than 50% of the test compounds 1 - 10 substructures at the top of the list are correct.

Table 2 Correctness of found MCSs using ranking criterion $R2$. T, number of correct substructures at the top of the ranked list. Total number of test compounds was 100.

T	% of investigated test compounds
0	9
1 - 10	52
11 - 20	24
21 - 30	7
31 - 40	3
41 - 50	5

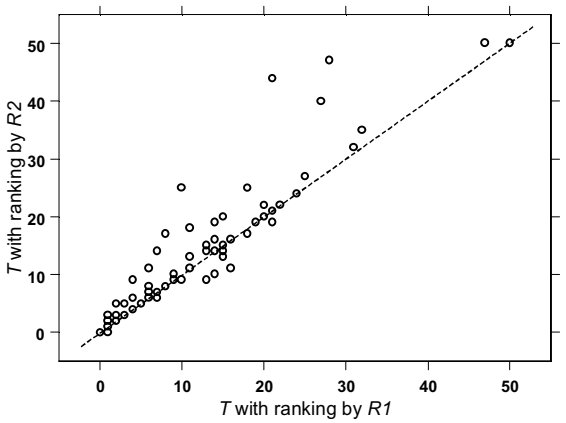

Fig. 2 Comparison of the correctness of the found substructures using criterion *T* (number of correct substructures at the top of the ranked list) for ranking by *R1* (Eq. 3) and *R2* (Eq. 4), respectively. Most of the test compounds are located above the broken line, demonstrating that *R2* is better than *R1*.

In a comparison of the *T* values with the Wilcoxon signed-rank test and the paired t-test both gave probabilities <0.001 for the zero hypothesis. Fig. 2 shows that in most cases a ranking by *R2* gives larger *T* than ranking by *R1*. The conclusion from this investigation is that the new ranking criterion (*R2*) yields significantly more correct substructures at the top of the ranked list than the old criterion (*R1*). However, this evaluation only considers the correctness of the found substructures but not their chemical information content.

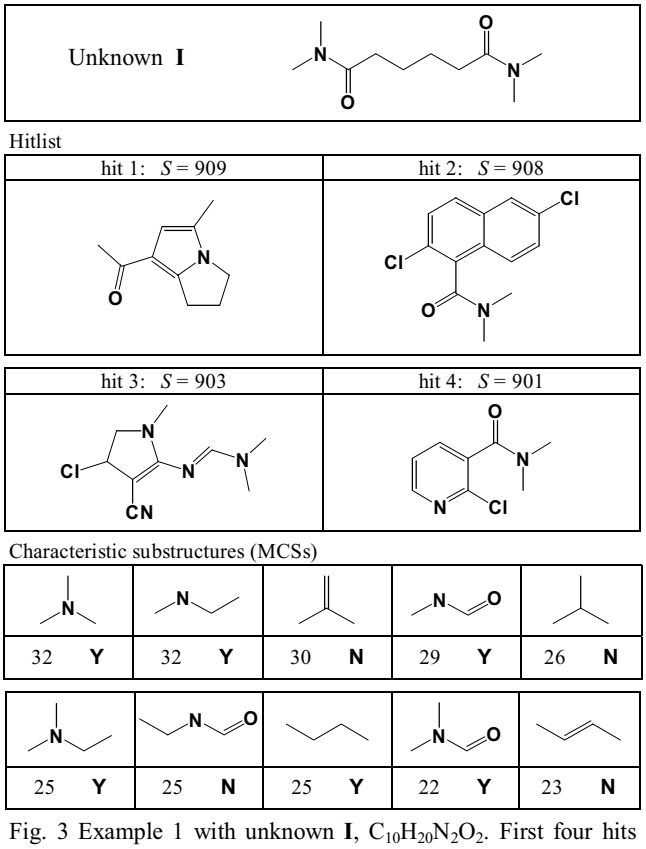

Fig. 3 Example 1 with unknown **I**, $C_{10}H_{20}N_2O_2$. First four hits from spectra similarity search are shown with spectral similarities, *S*. For ten characteristic substructures derived from 50 hitlist structures the frequencies in the hitlist structures are given; **Y** denotes a correct substructure, **N** a wrong one.

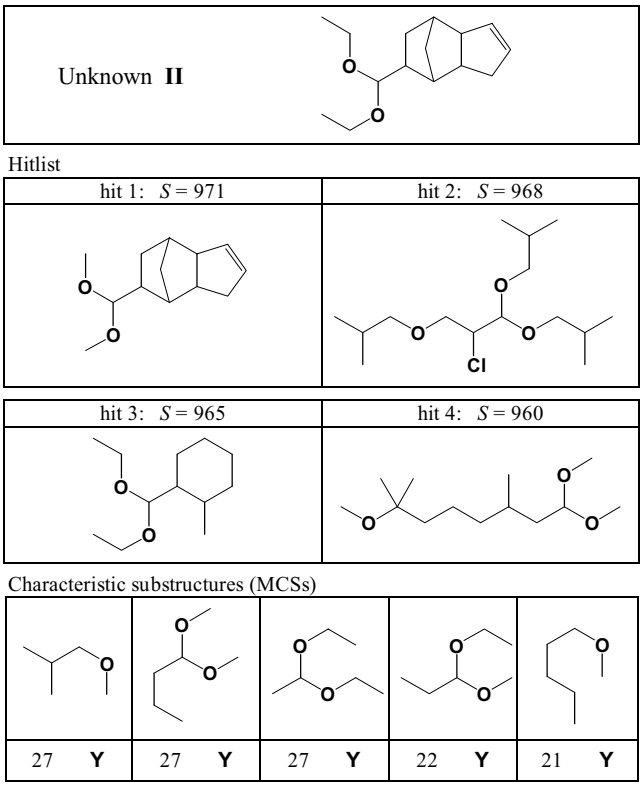

Fig. 4 Example 2 with unknown **II**, $C_{15}H_{22}O_2$. First four hits from spectra similarity search are shown with spectral similarities, *S*. Only substructures that are not part of other MCSs are given. The numbers are the frequencies in the used 50 hitlist structures; **Y** denotes a correct substructure.

Example 1

The IR spectrum of compound **I** in Fig. 3 was used for the spectra similarity search. The structures of the first four hits do not have a skeleton similar to that of the query structure, however, the functional groups are represented. Six of the ten best ranked MCSs are correct and four are wrong. This result demonstrates that a fully automatic evaluation and interpretation of the hitlist structures is not possible by the applied method. The erroneously predicted C-C-double bond can for instance be recognized by the absence of a characteristic C-H band between 3010 and 3100 cm^{-1}.

Example 2

For compound **II** in Fig. 4 the first four hits obtained by the spectra similarity search are shown. To avoid redundancies in the result all MCSs were removed that are substructures of other MCSs. This was performed by calculating the substructure isomorphism matrix.[34] Each MCS is considered as a target structure and as a substructure; matrix element *(i,j)* is 1 if MCS *i* is contained in MCS *j*, and is 0 otherwise. The remaining five MCSs are shown; all of them are correct. While the oxygen containing functional groups of the unknown are well represented, the ring structure could not be detected.

Example 3

With compound **III** in Fig. 5 the characteristic substructures are compared as obtained by using ranking criterion *R*1 or *R2*. The erroneous MCS 5 and MCS 6 in the list ranked by *R1* are moved to positions 9 and 7 when *R2* is applied, because the hitlist structures containing these MCSs are at the end of the hitlist. Two other MCSs gain a better ranking with *R2* so that the first six substructures become all correct.

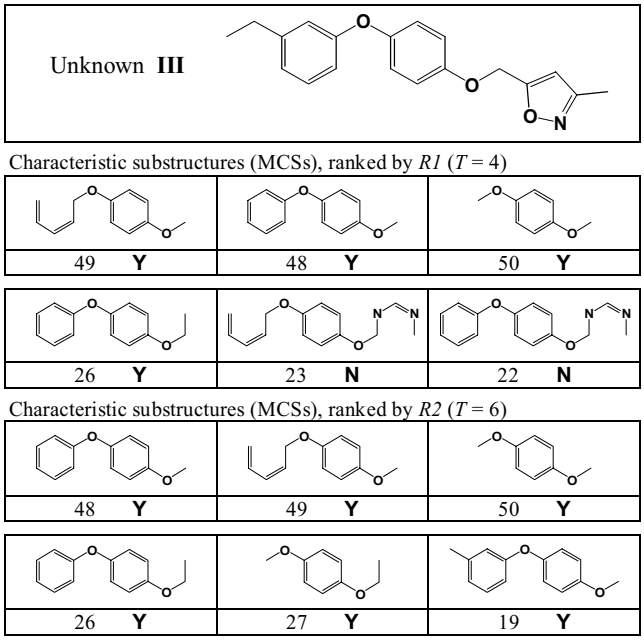

Fig. 5 Example 3 with unknown **III**, $C_{19}H_{19}NO_3$. Six substructures are shown either ranked by criterion *R1* or *R2*. The numbers are the frequencies in the 50 used hitlist structures; **Y** denotes a correct substructure, **N** a wrong one.

The query structure is well represented by the found substructures: both benzene rings are indicated, as well as the phenoxy group and part of the isoxazole ring.

Conclusion

Chemical structures in hitlists from IR spectra similarity searches can be evaluated by the described method based on maximum common substructures. Result is a list of substructures that are mostly contained in the query structure and provide useful hints for structure elucidation. The new measure for ranking these substructures places more correct substructures near the top of the list than the previously used one.

In general the obtained list with characteristic substructures has to be judged by a human expert in terms of spectroscopic relevance. Hidden correlations between substructures in the spectral library sometimes cause the prediction of substructures not present in the unknown. Usually, erroneous substructures can be detected by an inspection of the IR spectrum of the unknown or by complementary spectral data. Simulation of IR data for the suggested substructures - and comparison with the spectrum of the unknown - may automate this procedure to a great extent.

Acknowledgements

The authors thanks H. Scsibrany for his contributions, and R. Neudert (Chemical Concepts, Weinheim, Germany) as well as E. Pretsch (ETH Zurich, Switzerland) for providing the SpecInfo IR database. The work was supported by the Austrian Science Fund, project P14792-CHE.

References

1. T.L. Clerc, in *"Computer-enhanced analytical spectroscopy"*, ed. H.L.C. Meuzelaar and T.L. Isenhour, **1987**, Vol. 1, Plenum Press, New York, 145.
2. F. Ehrentreich, *Anal. Chim. Acta*, **1999**, *393*, 193.
3. H.J. Luinge, *Vib. Spectrosc.*, **1990**, *1*, 3.
4. C. Affolter, K. Baumann, J.T. Clerc, H. Schriber and E. Pretsch, *Mikrochim. Acta [Suppl.]*, **1997**, *14*, 143.
5. K. Baumann and J.T. Clerc, *Anal. Chim. Acta*, **1997**, *348*, 327.
6. F. Ehrentreich, *Fresenius J. Anal. Chem.*, **1997**, *359*, 56.
7. W. Werther and K. Varmuza, *Fresenius J. Anal. Chem.*, **1992**, *344*, 223.
8. H.J. Luinge, J.H. van der Maas and T. Visser, *Chemometrics Intell. Lab. Syst.*, **1995**, *28*, 129.
9. C. Klawun and C.L. Wilkins, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 69.
10. M.E. Munk and M.S. Madison, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 231.
11. M. Novic and J. Zupan, *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 454.
12. P.N. Penchev, G.N. Andreev and K. Varmuza, *Anal. Chim. Acta*, **1999**, *388*, 145.
13. D. Ricard, C. Cachet and D. Cabrol-Bass, *J. Chem. Inf. Comput. Sci.*, **1993**, *33*, 202.
14. J. Schuur and J. Gasteiger, *Anal. Chem.*, **1997**, *69*, 2398.
15. K. Funatsu and S.I. Sasaki, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 190.
16. A. Kerber and R. Laue, *"MOLGEN: Isomer Generator Software, 3.1"*, **1998**, University of Bayreuth, Institute for Mathematics II, Bayreuth.
17. S.E. Stein, *J. Am. Soc. Mass Spectrom.*, **1995**, *6*, 644.
18. M.M. Cone, R. Venkataraghavan and F.W. McLafferty, *J. Am. Chem. Soc.*, **1977**, *99*, 7668.
19. H. Scsibrany and K. Varmuza, *Fresenius J. Anal. Chem.*, **1992**, *344*, 220.
20. P.N. Penchev and K. Varmuza, *Comp. Chem.*, **2001**, *25*, 231.
21. K. Varmuza, P.N. Penchev and H. Scsibrany, *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 420.
22. K. Varmuza, P.N. Penchev and H. Scsibrany, *Vib. Spectrosc.*, **1999**, *19*, 407.
23. F. Ehrentreich, *Anal. Chim. Acta*, **2001**, *427*, 233.
24. L. Chen and W. Robien, *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 934.
25. J.J. McGregor and P. Willett, *J. Chem. Inf. Comput. Sci.*, **1981**, *21*, 137.
26. J.J. McGregor, *Software - Practice and Experience*, **1982**, *12*, 23.
27. Y. Takahashi, Y. Satoh, H. Suzuki and S.I. Sasaki, *Anal. Sci.*, **1987**, *3*, 23.
28. T. Wang and J. Zhou, *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 828.
29. J. Xu, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 25.
30. H. Scsibrany and K. Varmuza, in *"Software development in chemistry"*, ed. C. Jochum, **1994**, Vol. 8, Gesellschaft Deutscher Chemiker, Frankfurt am Main, 235.
31. N.T. Kotchev and P.N. Penchev, *unpublished*, **2001**.
32. SpecInfo, *"Spectroscopic information system"*, **1996**, Chemical Concepts (Wiley), Weinheim, Germany.
33. P.N. Penchev, N.T. Kochev and G.N. Andreev, *Comptes Rendus de l´Academie Bulgares des Sciences*, **1998**, *50*, 67.
34. K. Varmuza and H. Scsibrany, *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 308.