Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 490 (2003) 313–324
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Spectral similarity versus structural similarity: infrared spectroscopy

K. Varmuza*, M. Karlovits, W. Demuth

Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology, Getreidemarkt 9/166, A-1060 Vienna, Austria

Accepted 21 May 2003

Abstract

A new method is described for evaluation of spectral similarity searches. Aim of the method is to measure the similarity between the chemical structures of query compounds and the found reference compounds (hits). A high structural similarity is essential if the query is not present in the spectral library. Similarity of chemical structures was measured by the Tanimoto index, calculated from 1365 binary substructure descriptors. The method has been applied to several 1000 hitlists from searches in an infrared (IR) spectra database containing 13,484 compounds. Hitlists with highest structure information were obtained using a similarity measure based on the correlation coefficient computed from mean centered absorbance units. Frequency distributions of spectral and structural similarities have been investigated and a threshold for the spectral similarity has been derived that in general gives hitlists exhibiting significant chemical structure similarities with the query.

Keywords: Spectral library search; IR spectra; Tanimoto index; Substructure descriptors; Interpretative power

1. Introduction

Information about the chemical structure of a compound is difficult to extract from the infrared (IR) spectrum or the mass spectrum (MS) because of the complicated and widely unknown relationships between these spectral data and the chemical structure. Aim of spectra evaluation is often the *identification* of a compound, assuming its spectrum is already known and available; but may also be the *interpretation* in terms of the unknown chemical structure with the spectrum of the unknown not available as a reference [1,2]. Identification of compounds is performed best by library search methods based on spectral similarities; a number of spectral databases and powerful software products are available for this purpose and are routinely used. The more challenging problem is the interpretation of spectra, which still is a research topic in spectroscopy, chemometrics and computer chemistry. Up to now, no comprehensive solutions are available and the suggested methods are not used in practice.

The potential of spectral library search methods is not clear in cases they are used to interpret spectra from unknowns that are not present in the library. An *interpretative power* is claimed by some spectroscopic database systems, however, systematic investigations are rare that try to quantify the similarity between the chemical structures of hitlist compounds and the chemical structure of query compounds. Recent

* Corresponding author. Tel.: +43-1-58801-16060; fax: +43-1-58801-16091.
E-mail address: kvarmuza@email.tuwien.ac.at (K. Varmuza).

0003-2670/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0003-2670(03)00668-8

approaches for a computer-assisted evaluation or exploitation of hitlists in IR apply different strategies and can be summarized as follows. Hitlists from IR spectral similarity searches have been evaluated by extracting large and frequent substructures (maximum common substructures) in the hitlist compounds; such substructures can be useful in generating candidate structures for the query compound [3–9]. Fuzzy logic together with spectral similarity search have been applied for IR spectra interpretation [10]. For the purpose of high-throughput screening in searches of biological active compounds the similarities of FT-IR spectra have been compared with structural similarities using principal component analysis (PCA) [11]. Linked similarity searches for structures and IR spectra have been applied for the prediction of IR spectra [12]. Neural networks have been trained with IR spectra and a special code for 3D structures with the aim to predict the 3D structure from IR data [13]. Considerable effort has been put into the development of multivariate classifiers, based on discriminant analysis, PLS, and neural networks, that recognize substructures from IR data [14–19]. PLS and Kohonen mapping have been applied for cluster analysis of spectral and structural data and for classification [20,21].

All these methods rely on the assumption that similar spectra indicate similar chemical structures—although it is for instance known that a given substructure (functional group) may result in IR peaks at different wave numbers. Strategies for an evaluation of spectral library search methods with respect to their capability of finding identical or similar compounds have been discussed and summarized by Clerc et al. [22] and Luinge [2]. A general theory of similarity measures for library search systems has been developed [23]. The performance of finding identical and similar reference compounds was estimated by using small sets of user-selected compounds [24].

In the present work a large number of hitlists have been investigated with the aim of quantifying the similarity between chemical structure of query compounds and the chemical structures of the found hits. The newly developed method was applied for comparisons of different spectra similarity criteria and different spectral resolutions. The statistical frequency distributions of spectral and structural similarities were analyzed.

2. Methods

A spectral library search hitlist contains reference compounds with spectra that are most similar to the spectrum of a given query spectrum. The interpretative power of a spectral library search system is the ability to produce hitlists with chemical structures that are very similar to the structure of query compounds. For a systematic evaluation of library search systems it is necessary to define similarity criteria for spectra as well as for chemical structures. Similarity concepts are always relative and refer to some specific context [25]; nevertheless they are necessary and useful if relationships between available data (spectra) and desired data (chemical structures) cannot be described sufficiently by formal models.

2.1. Similarity of IR spectra

An infrared spectrum is characterized by a vector $\boldsymbol{x}$ with component $x(j)$ being the averaged absorbance in wave number interval j, scaled to the range of 0–1. The similarity between two IR spectra, represented by vectors $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$, has been characterized by four different similarity measures [4]. All similarity measures are scaled to the range of 0–999, with the maximum obtained for $\boldsymbol{x}_A = \boldsymbol{x}_B$. The number of intervals, k, was typical 801 for a wave number range of 500–3700 cm^{-1}, corresponding to an interval width of 4 cm^{-1}. Vector notation is used here as far as possible; all vectors (for instance $\boldsymbol{x}$) are considered as column vectors; transposed vectors (for instance $\boldsymbol{x}^T$) are row vectors; $\mathbf{1}$ is a vector of appropriate size with all elements being 1.

1. Spectral similarity COR is based on the correlation coefficient calculated from the mean centered absorbances as follows. Let $\bar{x}_A$ be the arithmetic mean of the absorbances $x_A(j)$ in spectrum A, and $\bar{x}_B$ the corresponding mean for spectrum B. The mean centered absorbance vectors of spectra A and B are given by

$$\boldsymbol{z}_A = \boldsymbol{x}_A - \mathbf{1} \cdot \bar{x}_A \quad (1)$$

$$\boldsymbol{z}_B = \boldsymbol{x}_B - \mathbf{1} \cdot \bar{x}_B \quad (2)$$

respectively, and the correlation coefficient is

$$r = \frac{\boldsymbol{z}_A^T \cdot \boldsymbol{z}_B}{||\boldsymbol{z}_A|| \cdot ||\boldsymbol{z}_B||} \quad (3)$$

Note that the Euclidean norm $||\boldsymbol{z}||$ is equivalent to the length $\left[\sum z(j)^2\right]^{0.5}$ $(j = 1, \ldots, k)$ of the vector. Appropriate scaling of r gives the spectra similarity measure in the range of 0–999 as

$$\mathrm{COR} = \frac{999(r+1)}{2} \quad (4)$$

2. Spectral similarity MAD is based on the mean of the absolute absorbance differences, scaled to the range 0–999. A vector $\boldsymbol{d}$, containing the absolute absorbance differences, is given by

$$\boldsymbol{d} = |\boldsymbol{x}_A - \boldsymbol{x}_B| \quad (5)$$

and the spectra similarity measure by

$$\mathrm{MAD} = 999\left[1 - \frac{\boldsymbol{d}^T \cdot \boldsymbol{1}}{k}\right] \quad (6)$$

Note that $\boldsymbol{d}^T \cdot \boldsymbol{1}$ is equal to the sum of all $d(j)$ with $j = 1, \ldots, k$.

3. Spectral similarity measure MSD is based on the mean of the squared absorbance differences, given by

$$\mathrm{MSD} = 999\left[1 - \left(\frac{\boldsymbol{d}^T \cdot \boldsymbol{d}}{k}\right)^{0.5}\right] \quad (7)$$

4. Spectral similarity measure DPN is proportional to the dot product of the absorbance vectors $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$, each normalized to unit length; it corresponds to the correlation coefficient of the original absorbances.

$$\mathrm{DPN} = \frac{999\boldsymbol{x}_A^T \cdot \boldsymbol{x}_B}{||\boldsymbol{x}_A|| \cdot ||\boldsymbol{x}_B||} \quad (8)$$

2.2. Similarity of chemical structures

Chemical 2D structures were available as connection tables. Each chemical structure has been characterized by a vector $\boldsymbol{y}$ with components $y(j)$ being binary substructure descriptors. A set of 1365 substructures (see Section 3.2) was defined for this purpose; $y(j)$ is 1 if substructure j is present in the molecule and 0 otherwise [26,27]. The similarity of two structures, represented by vectors $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$, was characterized by the widely-used Tanimoto index [28], t, also called Jaccard similarity [29].

$$t_{A,B} = \frac{\sum \mathrm{AND}[y_A(j), y_B(j)]}{\sum \mathrm{OR}[y_A(j), y_B(j)]} \quad \text{with } j = 1, \ldots, 1365 \quad (9)$$

AND$[\cdot, \cdot]$ is the result of the logical AND of the given two binary variables, and OR$[\cdot, \cdot]$ the result of the logical OR. The summation is calculated over all used descriptors. Eq. (9) can be written in vector notation as

$$t_{A,B} = \frac{\boldsymbol{y}_A^T \cdot \boldsymbol{y}_B}{\boldsymbol{y}_A^T \cdot \boldsymbol{1} + \boldsymbol{y}_B^T \cdot \boldsymbol{1} - \boldsymbol{y}_A^T \cdot \boldsymbol{y}_B} \quad (10)$$

The Tanimoto index is in the range of 0–1; the value 1 is obtained if all descriptors are pairwise equal. The Tanimoto index only considers substructures that are contained at least in one of the compared structures. Details about the software SubMat and the substructures used for the calculation of binary substructure descriptors are given in Sections 3.1 and 3.2.

2.3. Similarity between query structure and hitlist structures

A hitlist obtained by a spectral similarity search contains h hits which are the database compounds exhibiting spectra most similar to the query spectrum. The hitlist is ranked by decreasing spectra similarity; the first hit is the reference spectrum most similar to the query spectrum. The size, h, of a hitlist is typical 10–100. The quality of a hitlist in terms of high structural similarity with the query structure has been measured by an averaged Tanimoto index, $t_Q(h)$ that describes the structural similarity between a query structure Q, and the hitlist structures 1 to h.

$$t_Q(h) = \left(\frac{1}{h}\right)\sum t_{Q,i} \quad \text{with } i = 1, \ldots, h \quad (11)$$

If spectral and structural similarities are related then, in general, the first hit exhibits a large (in an ideal case the maximum possible) structural similarity with the query; that means $t_Q(h)$ decreases with increasing h.

To characterize the performance of a spectral similarity search method, a set of hitlists has to be considered, obtained from n randomly selected query

compounds. Averaging $t_Q(h)$ over all hitlists gives grand means $T(h)$ of Tanimoto indices, ranging between 0 and 1.

$$T(h) = \left(\frac{1}{n}\right) \sum t_Q(h) \quad \text{with } Q = 1, \ldots, n \qquad (12)$$

Plots of $T(h)$ versus the number of considered hits, h, are used in this work to characterize and to compare spectra similarity search methods.

3. Experimental and results

3.1. Database and software

The IR database used consists of 13,484 compounds and is part of the SpecInfo system [30]. Each IR spectrum contains 801 absorbance values in the spectral range 500–3700 cm^{-1}, corresponding to a sampling interval of 4 cm^{-1}. Absorbances within each spectrum have been scaled to the integer number range 0–255, corresponding to a resolution of 8 bit. Chemical structures of the database compounds were available in Molfile format [31,32].

Binary substructure descriptors were calculated by software SubMat [27,33] running under operating systems Microsoft Windows 95/98/NT/2000. SubMat uses two input files, one for the molecular structures, the other for the substructures; both have to be in Molfile format. The output file is in text format and contains a matrix with the descriptor vectors $\boldsymbol{y}$ for the molecular structures. Current limits for SubMat are 127 explicitly defined atoms and 255 bonds per structure. Typical computing time with a Pentium IV 1.8 GHz is 3 s for 100 molecular structures and 1000 substructures. SubMat has two operating modes. One is the interactive mode with a Windows style graphical user interface. The other is the remote mode by calling SubMat from another program and supplying necessary parameters via a command file in text format. Dedicated temporary files (so called semaphore files) are used for the communication between the programs providing messages for progress, error and termination. The remote mode of SubMat has been extensively used together with Matlab programs.

Software written in Matlab 6.0 and Visual C++ has been used for structure and spectra file handling, spectral similarity searches and statistics.

3.2. Substructures

A set of 1365 generally applicable substructures has been defined for the characterization of organic molecular structures by binary substructure descriptors [27]. The non-hydrogen elements used in the substructures were C, N, O, S, P, F, Cl, Br, I, B, and Si; furthermore, a pseudo element A for hetero atoms and a pseudo element Q for non-H atoms were allowed. The bond types used were single, double, triple, aromatic, and not-defined. A part of the substructures was built systematically; others were defined on the basis of chemical and spectroscopic ideas. The substructures are collected in eight groups as follows.

Group 1 defines the presence of a minimum number of atoms of selected elements as $N_{1–6}$, $O_{1–6}$, $S_{1–6}$, $P_{1–6}$, $F_{1–3}$, $Cl_{1–3}$, $Br_{1–3}$, $I_{1–3}$, $B_{1–3}$, Si, $A_{1–6}$ (46 substructures). *Group 2* contains 78 two-atom substructures made from elements C, N, O, S, F, Cl, Br, I, A, Q by applying the five bond types defined above and considering the chemical valence rules. Examples are C–N, C=O, A–A. *Group 3* contains 404 non-aromatic, single rings with three to eight ring atoms. Heterocyclic rings of size three, four, and five, containing C, N, O were exhaustively built by the isomer generator software MOLGEN [34] resulting in 262 substructures; 116 of them were used because they are present in more than 20 compounds of the Crossfire Beilstein database system [35]. *Group 4* contains 93 substructures with not aromatic condensed rings. *Group 5* contains 97 substructures with at least one aromatic ring (benzene ring or N-aromatic ring). *Group 6* contains 39 bridged ring systems. *Group 7* contains 418 non-cyclic substructures built from elements C, N, O, Q; for instance all possible tree structures with three or four atoms (at least one C atom) are included. *Group 8* contains153 functional groups not present in other substructure groups; these substructures mainly contain N or/and O but also S, halogens and Si.

3.3. Structural properties of the IR database

The ranges of common elements in the 13,484 compounds are $C_{0–50}$, $H_{0–78}$, $N_{0–10}$, $O_{0–13}$, $F_{0–32}$, $Cl_{0–7}$, $S_{0–5}$; molecular weights are between 18 and 962. The number of compounds containing at least one hetero atom is 6278, containing a benzene ring 8499, any type of ring 11,526. From the originally defined 1365

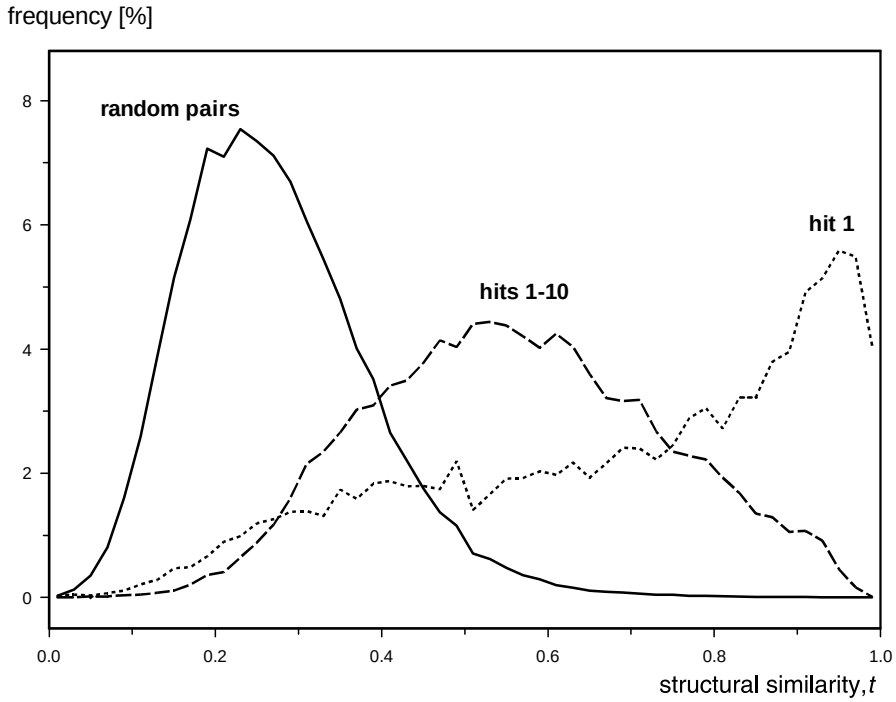

Fig. 1. Frequency distribution of Tanimoto indices, *t*, for 10^6 pairs of randomly selected different structures, the first hit, and the first 10 hits (13,484 query compounds). The frequency is given in percentage for 50 intervals of *t*, each 0.02 units wide.

substructures a set of 1065 (78%) is present in the compounds of the database. A few small substructures are contained in more than 99% of the database structures, however, more than half of the substructures are contained in only 5% of the database structures. The number of substructures per database structure varies between 2 and 212 with a median of 75. The diversity of the structures was characterized by the calculation of the Tanimoto indices for 10^6 pairs of randomly selected different structures [27]. The frequency distribution in Fig. 1 shows a skewed bell-shaped distribution with a mean of 0.27, and a standard deviation of 0.11. Tanimoto indices above 0.6 can be considered to indicate a significantly high structural similarity because only 1% of the random pairs yield higher values.

3.4. Comparison of spectral similarity measures

Random samples of 200 query compounds were selected from the database. For each query compound a hitlist containing 50 compounds with the most similar spectra was determined; the compound identical to the query compound was excluded from the hitlist. Fig. 2 shows the grand means *T(h)* of Tanimoto indices for five parallel experiments, each with 200 query compounds, using the spectral similarity COR (Eq. (4)). The results demonstrate that in general the structures of the first hits have highest similarity with the structure of the query compounds; for the first hit the averaged Tanimoto index is approximately 0.7. The maximum differences of *T(h)* between parallel experiments are only between 0.02 and 0.05.

The results obtained for spectra similarity hitlists can be compared with two extreme situations. One extreme are pseudo hitlists containing randomly selected compounds from the database. In this case the structural similarity of hitlist compounds corresponds to the average Tanimoto index within the database, which is 0.27 as described above. This value is considerable lower than the structural similarities obtained by spectral similarity searches.

The other extreme are hitlists containing the database compounds with highest structural similarity to the query compounds. Such pseudo hitlists exhibit

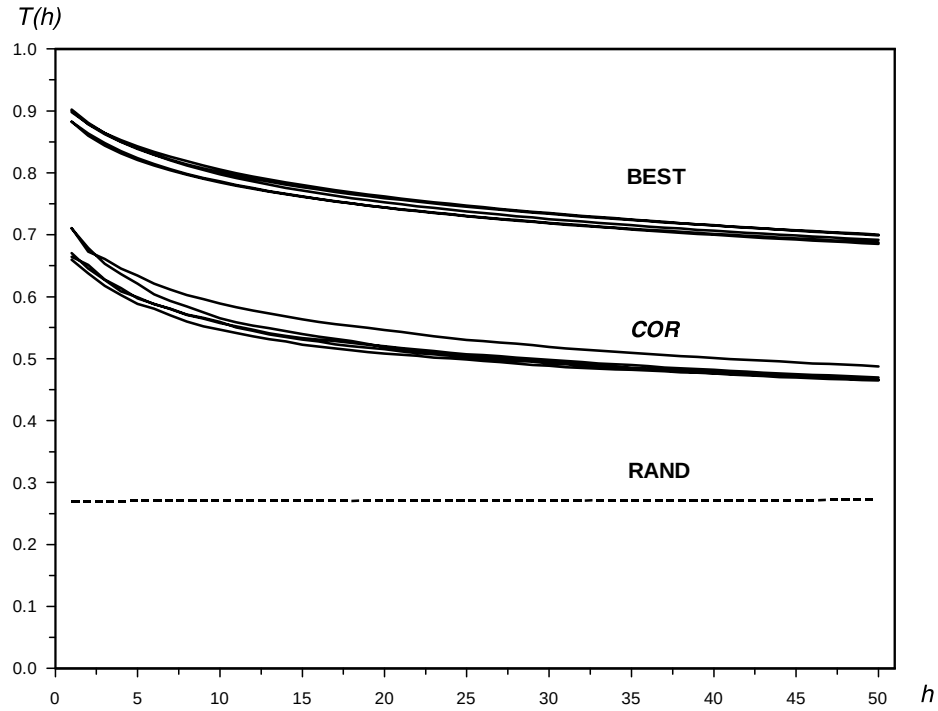

Fig. 2. Similarity *T(h)* between query structures and the corresponding *h* hitlist structures for spectra similarity measure COR, spectral resolution 4 cm^{-1}, and five parallel experiments each with 200 randomly selected query compounds. BEST, pseudo hitlists containing the database structures with maximum structural similarity to query structures; RAND, randomly selected pseudo hitlists, equivalent to the mean of Tanimoto indices in the database.

the maximum structural similarities that are possible with the used database; they can be determined only in tests with query compounds of known structure. As shown in Fig. 2 these maximum values are considerable higher than those obtained from hitlists based on spectral similarity. They define an upper limit given by the composition of the used database. In general, this limit cannot be achieved by a spectral similarity search because spectral data do not contain full structure information, and furthermore the used spectra similarity measure utilizes only a part of the inherent structure information.

The performances of the used four spectra similarity measures are compared in Fig. 3. Similarity measure COR is best, closely followed by DPN; worse results have been obtained by MSD and MAD. Also in a previous study [4]—using a maximum common substructure approach for the evaluation of hitlists—the criterion COR yielded best results.

3.5. Spectral resolution

The effect of spectral resolution on the performance of library searches [24,36–38] or substructure classifiers [15,16,19,21] has been reported in several papers. Typical sampling intervals used are between 4 and 32 cm^{-1}, corresponding to 800–100 wave number intervals (points per spectrum) between 500 and 3700 cm^{-1}. A reduced spectral resolution makes the similarity more tolerant and has therefore been proposed for similarity searches (search for compounds with similar structures) and classification searches (recognition of substance class from hitlist data) [24]. For the development of substructure classifiers by artificial neural networks a resolution of 14 cm^{-1} (256 points in the range 400–4000 cm^{-1}) was found to perform better than a resolution of 5.6 cm^{-1} [39].

In this work wave number intervals between 4 and 512 cm^{-1} have been applied to study the effect

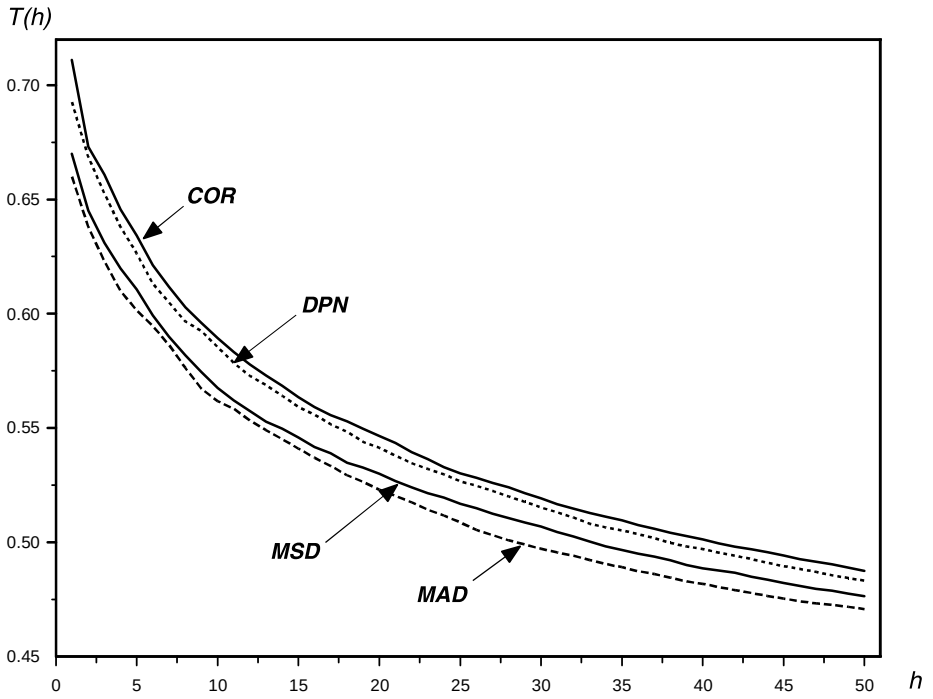

Fig. 3. Similarity *T(h)* between query structures and the corresponding *h* hitlist structures for four different spectra similarity measures, COR (Eq. (4)), DPN (Eq. (8)), MSD (Eq. (7)), and MAD (Eq. (6)).

of spectral resolution on the structural similarity of hitlists. The results shown in Fig. 4 have been obtained from a random sample containing 200 query compounds and using spectral similarity COR. A reduction of the resolution to 16 or even 32 cm^{-1} has no negative effect, however, 64 cm^{-1} or larger wave number intervals decrease the performance significantly. In contrast to above mentioned works an optimum resolution was not found for the used data. Obviously the detailed characterization of chemical structures by several hundred descriptors also requires a detailed coding of the spectra in order to achieve high structural similarities.

3.6. Hitlist example

Fig. 5 shows results from a spectrum similarity search and a corresponding structure similarity search for the query compound 3-aminobenzyl alcohol. The first five hits from the spectral similarity search have high spectral similarities between 912 and 868, but diverse Tanimoto indices between 0.96 and 0.14; the latter are of course not known for a unknown query compound. The five database compounds with structures most similar to the query are also shown; their Tanimoto indices are between 0.96 and 0.85. Note that the two best structures have been found in the spectral similarity search as hits 3 and 4. The hitlist spectra as well as the hitlist structures clearly indicate two groups of compounds, demonstrating that high spectral similarity not necessarily implies high structural similarity. The overall spectral similarity, COR, yields high values for spectra of aminobenzyl alcohols but also for spectra of some pyrazole compounds. This situation is not uncommon in spectral library searches and requires a critical evaluation of the hitlist. The spectroscopist may recognize the differences between aminobenzyl alcohols and pyrazoles in the spectral range 1000–1400 cm^{-1}—as well as characteristic bands for *ortho*-, *meta*-, and *para*-substitution below 900 cm^{-1}. Considering this, the correct substance class of the unknown will be found. Such type of hitlist evaluation can be supported by multivariate exploratory data analysis. For instance, PCA can be applied to spectral or to structural data of the hitlist,

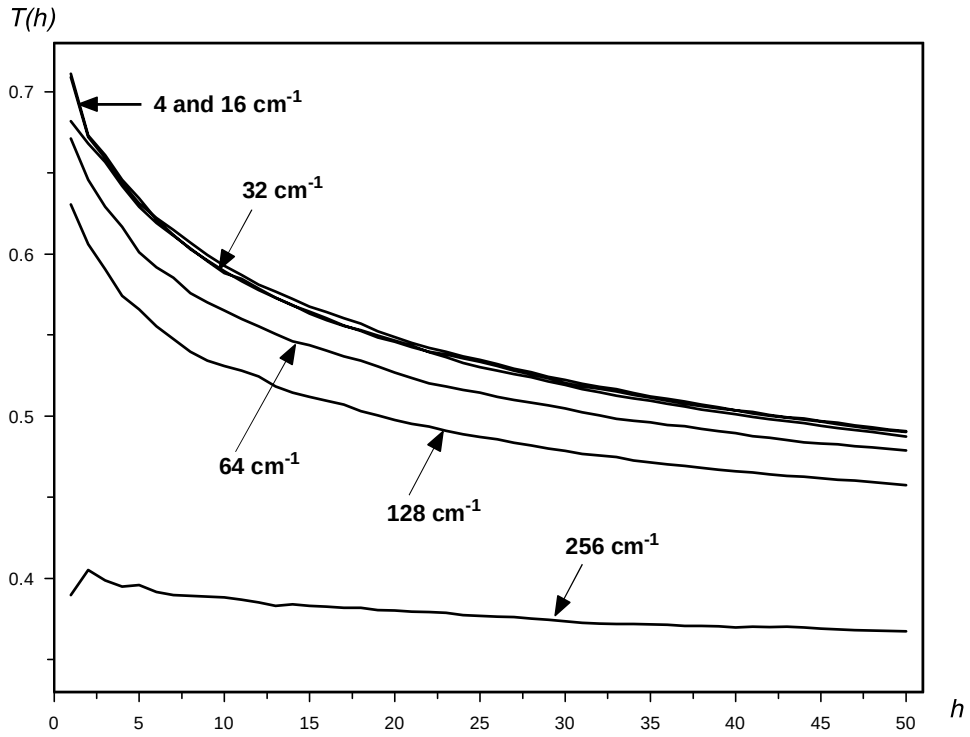

Fig. 4. Effect of spectral resolution (varied between 4 and 256 cm^{-1}) on the similarity *T(h)* between query structures and the corresponding *h* hitlist structures for spectra similarity measure COR.

and PLS can be applied to both. The resulting scatter plots are for instance appropriate for a visual cluster analysis of the hitlist compounds which assists to find the correct substance class of the unknown [20,27].

3.7. Relationship between spectral and structural similarity

The example in Section 3.6 demonstrates that an easy to describe, general relationship between spectral similarity and structural similarity does not exist. Some insight into this problem can be provided from frequency distributions of structural similarities, *t* (Tanimoto index) and spectral similarities, COR. As already discussed in Section 3.3. Fig. 1 contains the zero distribution of *t*, obtained from 10^6 randomly selected pairs of reference compounds. Ninety-nine percent of the random pairs have a structural similarity below 0.6. Considerable higher structural similarities occur for structure pairs built from a query structure and the structure of the corresponding first hit (curve "hit 1"). The maximum of this distribution is at $t = 0.95$; however, about one third of the queries gave first hits with rather poor structural similarities below 0.6. The averaged Tanimoto index for the first 10 hits (curve "hits 1–10") shows a wide distribution with 90% of the queries between $t = 0.29$ and 0.88. The latter two frequency distributions have been determined by using each of the 13,484 database compounds as a query compound.

A counterpart to Fig. 1 are frequency distributions of the spectral similarity, shown in Fig. 6. Again a zero distribution has been estimated from 10^6 randomly selected pairs of reference compounds; its mean is at COR = 625; 95% of the random pairs have spectral similarities below 800. Curve "hit 1" describes the distribution of similarities between query spectra and spectra of the corresponding first hits; 95% of the values are above 800, and the overlap with the zero distribution is small. Curve "hits 1–10" is the distribution for an averaged COR of the first 10 hits. Also these two distributions have been determined by using

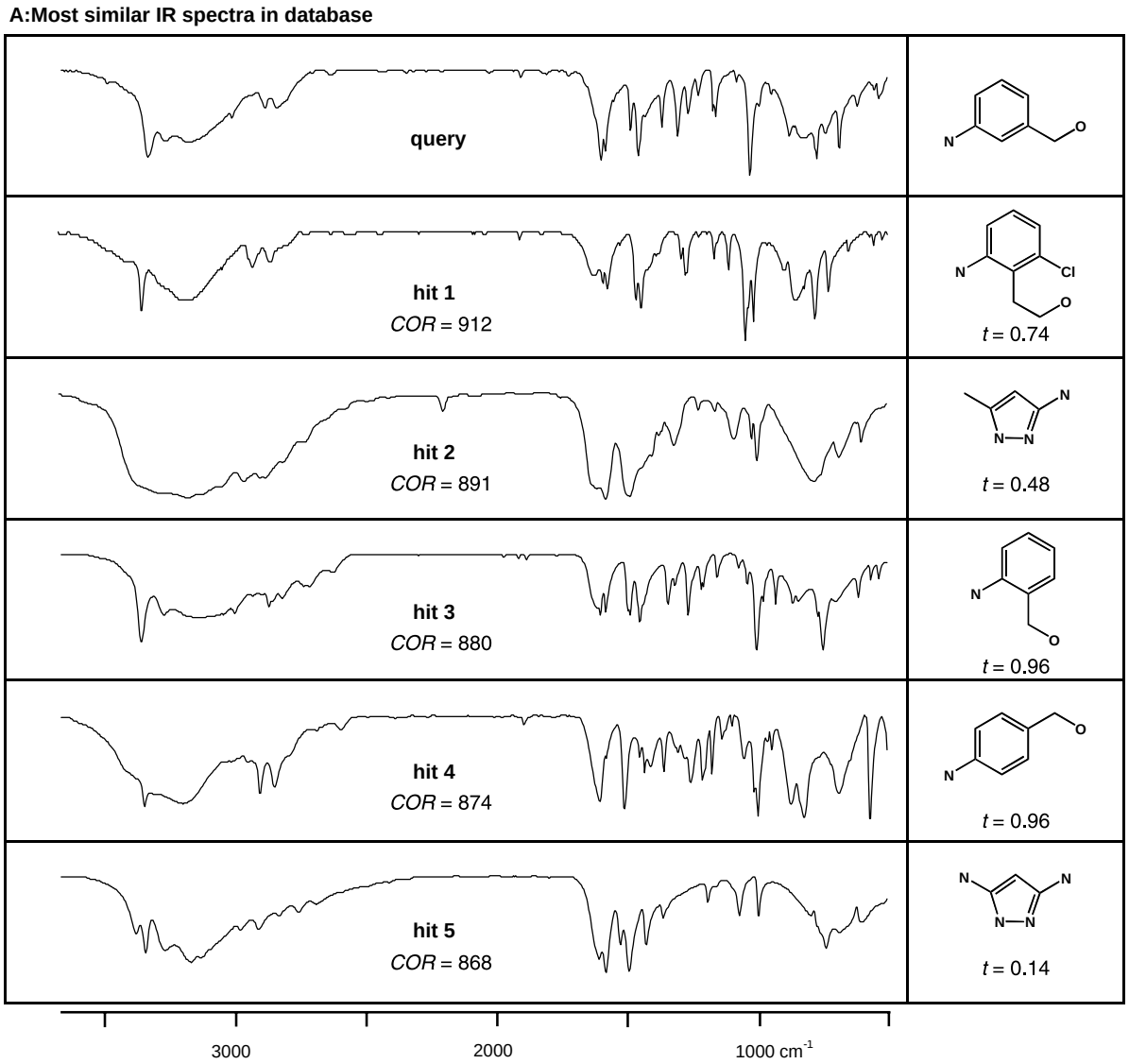

Fig. 5. Search results for query compound 3-aminobenzyl alcohol. (A) Query and first five hits from spectra similarity search; (B) five database compounds with structures most similar to the structure of the query compound. COR, spectral similarity; *t*, structural similarity (Tanimoto index).

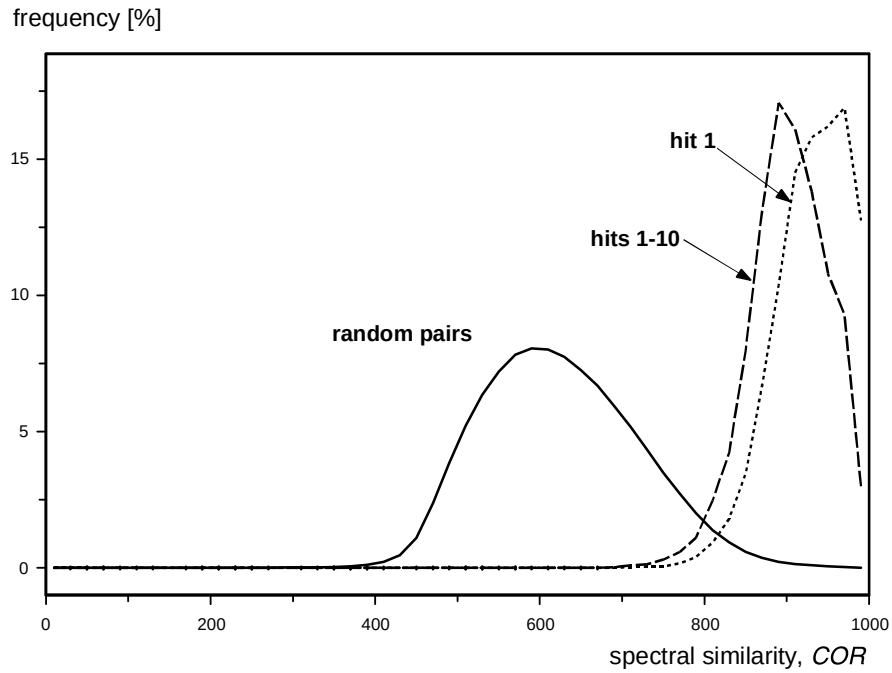

Fig. 6. Frequency distribution of spectral similarity, COR, for 10^6 pairs of randomly selected different structures, the first hit, and the first 10 hits (13,484 query compounds). The frequency is given in percentage for 50 intervals of COR, each 20 units wide.

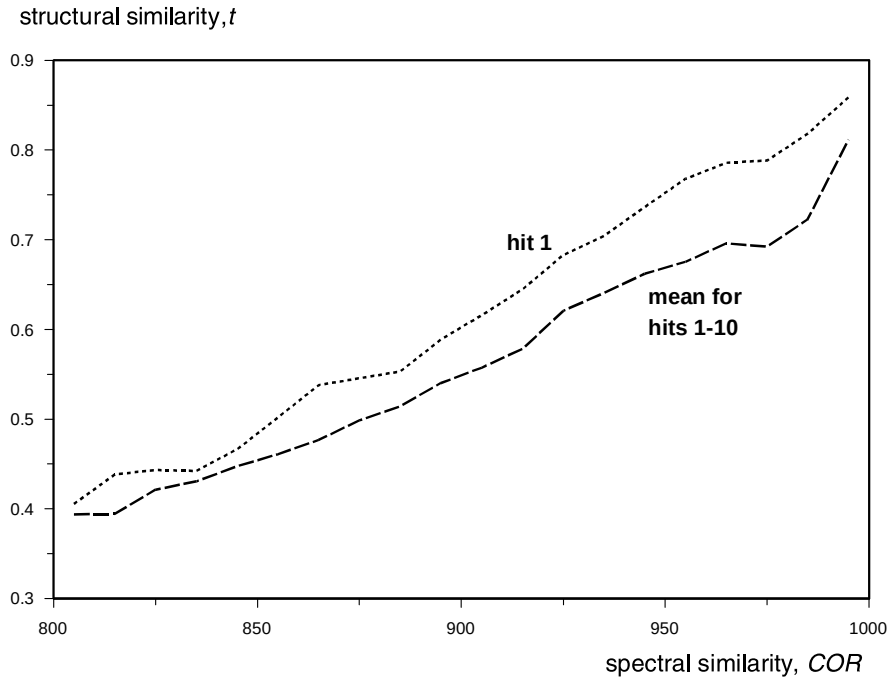

Fig. 7. Relationship between spectral similarity, COR, and structural similarity, *t*, for the first hit and the first 10 hits.

each of the 13,484 database compounds as a query compound.

Finally, the spectral similarity (COR) of the first hit has been plotted versus the averaged structural similarity (t) of the first hit (Fig. 7). For this purpose the region 800–1000 of COR was divided into 20 intervals; in each interval the resulting Tanimoto indices have been averaged and plotted versus the center of the interval. The general trend is an increase of the structural similarity with increasing spectral similarity of the first hit. The same result is obtained for the mean structural similarity of hits 1–10. For practical applications one can conclude from Fig. 7 that in general a value for COR above 900 is necessary to obtain a Tanimoto index of 0.6 or higher which has a probability of less than 1% for randomly selected pairs and therefore can be considered as a significant high structural similarity. However, rather large standard deviations appear for the structural similarity of the first hit. For instance, 1120 test compounds yield a spectral similarity between 960 and 970 for the first hit; the structural similarity between the query and the first hit has a mean of 0.79 and a standard deviation of 0.18. That means the reference compounds of some hitlists correspond almost to a random selection, while for others they are near the optimum selection.

4. Conclusions

A new method has been developed for a quantitative evaluation of the *interpretative power* of spectra similarity search methods (spectral library searches). A high interpretative power gives hitlists that contain compounds with structures that are very similar to the query structure; that means the library search method is useful even in cases the query compound is not contained in the database.

A quantitative evaluation of the interpretative power not only requires a strictly defined spectral similarity but also an unequivocal definition for the similarity of chemical structures. It is obvious that chemical structure similarity heavily depends on the context of using the structures. Nevertheless, general applicable measures for chemical structure similarity are successfully used for instance to characterize the diversity of structural libraries. A widely used measure is the Tanimoto index based on binary substructure descriptor; this criterion was also applied in this work. Tests showed that the created set of 1365 substructures gives calculated structure similarities that mostly correspond to the chemist's impression of similarity. If necessary this concept of structure similarity can be adapted to special needs by using another more appropriate set of substructures.

The results obtained for an IR database can be summarized as follows. The first hits (corresponding to the most similar spectra) yield highest structural similarity with the query compound. Among the four investigated spectra similarity measures COR performed best (correlation coefficient of mean-centered absorbances). The resolution of IR spectra can be reduced to about 200 data points for a wave number range of 500–3700 cm^{-1} (corresponding to an interval width of 32 cm^{-1}) without loss of interpretative power. A high similarity between the query structure and the structures of the first 10 hits can be expected, if the spectral similarity between the query and the first hit is above 900. Spectral similarities below 800 indicate that the query has no sufficient similar spectra in the database and therefore the hitlist will in general not contain reliable structure information. An application of the described method to mass spectra is in progress.

Although it is well known to practicing spectroscopists that compounds with similar chemical structures often have similar IR spectra (and hopefully vice versa), it is desirable to check and to quantify this hypothesis. The method presented makes it possible to test new spectra search algorithms for their capability to produce hitlists with highest structure information. From such hitlists substructures that are relevant for the unknown can be extracted, either by statistical evaluation, a maximum common substructure approach, or by manual inspection. This is a complementary way to the more global approaches of knowledge-based systems or spectral classifiers, to achieve structural restrictions that can be used in systematic structure elucidation.

Acknowledgements

The authors thank R. Neudert and E. Pretsch for providing the SpecInfo IR database, A. Kerber and R. Laue for the isomer generator software MOLGEN, as well as P. Penchev, H. Scsibrany, and S. Qehaja

for collaboration. Two unknown reviewers made constructive suggestions. The work was supported by the Austrian Science Fund, project P14792-CHE.

References

[1] J.T. Clerc, in: H.L.C. Meuzelaar, T.L. Isenhour (Eds.), Computer-Enhanced Analytical Spectroscopy, Plenum Press, New York, 1987, pp. 145–162.
[2] H.J. Luinge, Vibr. Spectrosc. 1 (1990) 3–18.
[3] K.S. Lebedev, Zh. Analit. Khim. 48 (1993) 851–863.
[4] K. Varmuza, P.N. Penchev, H. Scsibrany, J. Chem. Inform. Comput. Sci. 38 (1998) 420–427.
[5] K. Varmuza, P.N. Penchev, H. Scsibrany, Vibr. Spectrosc. 19 (1999) 407–412.
[6] P. Penchev, K. Varmuza, Comp. Chem. 25 (2001) 231–237.
[7] K. Varmuza, N.T. Kochev, P. Penchev, Anal. Sci. 17 (2001) i659–i662.
[8] B.G. Derendyaev, L.I. Makarov, T.F. Bogdanova, V.N. Piottukh-Peletskii, J. Struct. Chem. 42 (2001) 271–280.
[9] F. Ehrentreich, Anal. Chim. Acta 427 (2001) 233–244.
[10] F. Ehrentreich, Anal. Chim. Acta 393 (1999) 193–200.
[11] V. Schoonjans, F. Questier, Q. Guo, Y. Van der Heyden, D.L. Massart, J. Pharm. Biomed. Anal. 24 (2001) 613–627.
[12] K. Baumann, J.T. Clerc, Anal. Chim. Acta 348 (1997) 327–343.
[13] M.C. Hemmer, J. Gasteiger, Anal. Chim. Acta 420 (2000) 145–154.
[14] D. Cabrol-Bass, C. Cachet, C. Cleva, A. Eghbaldar, T.P. Forrest, Can. J. Chem. 73 (1995) 1412–1426.
[15] C. Klawun, C.L. Wilkins, J. Chem. Inform. Comput. Sci. 36 (1996) 69–91.
[16] H.J. Luinge, J.H. van der Maas, T. Visser, Chemom. Intell. Lab. Syst. 28 (1995) 129–138.
[17] M.E. Munk, M.S. Madison, E.W. Robb, Mikrochim. Acta (Wien) 11 (1991) 505–514.
[18] M.E. Munk, M.S. Madison, E.W. Robb, J. Chem. Inform. Comput. Sci. 36 (1996) 231–238.
[19] D. Ricard, C. Cachet, D. Cabrol-Bass, J. Chem. Inform. Comput. Sci. 33 (1993) 202–210.
[20] W. Werther, K. Varmuza, Fresenius J. Anal. Chem. 344 (1992) 223–226.
[21] M. Novic, J. Zupan, J. Chem. Inform. Comput. Sci. 35 (1995) 454–466.
[22] J.T. Clerc, E. Pretsch, M. Zürcher, Mikrochim. Acta (Wien) 11 (1986) 217–242.
[23] M. Zürcher, J.T. Clerc, M. Farkas, E. Pretsch, Anal. Chim. Acta 206 (1988) 161–172.
[24] C. Affolter, J.T. Clerc, Fresenius J. Anal. Chem. 344 (1992) 136–139.
[25] D.H. Rouvray, J. Chem. Inform. Comput. Sci. 34 (1994) 446–452.
[26] K. Varmuza, H. Scsibrany, J. Chem. Inform. Comput. Sci. 40 (2000) 308–313.
[27] H. Scsibrany, M. Karlovits, W. Demuth, F. Müller, K. Varmuza, Chemom. Intell. Lab. Syst., in print.
[28] P. Willett, Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, UK, 1987.
[29] B.G.M. Vandeginste, D.L. Massart, L.C.M. Buydens, S. De Jong, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998.
[30] Chemical Concepts, Spectroscopic Database System SpecInfo, http://www.chemicalconcepts.com/, Weinheim, Germany, 1996.
[31] A. Dalby, J.G. Nourse, W.D. Hounshell, A.K.I. Gushurst, D.L. Grier, B.A. Leland, J. Laufer, J. Chem. Inform. Comput. Sci. 32 (1992) 244–255.
[32] MDL-Information-Systems-Inc., CT file format, http://www.mdli.com/downloads/literature/ctfile.pdf, San Leandro, CA, USA, 2002.
[33] K. Varmuza, H. Scsibrany, Software SubMat, Generation of Binary Substructure Descriptors for Chemical Structures, Laboratory for Chemometrics, Vienna University of Technology, http://www.lcm.tuwien.ac.at, Vienna, Austria, 2002.
[34] A. Kerber, R. Laue, Software MOLGEN (Isomer generator software), Institute for Mathematics II, University of Bayreuth, http://www.mathe2.uni-bayreuth.de/molgen4/, Bayreuth, Germany, 2000.
[35] MDL-Information-Systems-Inc., Database Crossfire Beilstein, http://www.beilstein.com/products/xfire/, Frankfurt am Main, Germany, 1997.
[36] F. Ehrentreich, S.G. Nikolov, M. Wolkenstein, H. Hutter, Mikrochim. Acta 128 (1998) 241–250.
[37] S.R. Lowry, D.A. Huppler, C.R. Anderson, J. Chem. Inform. Comput. Sci. 25 (1985) 235–241.
[38] P.M. Owens, T.L. Isenhour, Anal. Chem. 55 (1983) 1548–1553.
[39] E.W. Robb, M.E. Munk, Mikrochim. Acta (Wien) 1 (1990) 131–155.