Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 516 (2004) 75–85
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Spectral similarity versus structural similarity: mass spectrometry

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

Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology, Getreidemarkt 9/166-2, A-1060 Vienna, Austria

Received 19 January 2004; received in revised form 9 April 2004; accepted 15 April 2004

Abstract

A recently described method [Anal. Chim. Acta 490 (2003) 313] for the evaluation of spectral similarity searches has been applied to low resolution mass spectra. Aim of the method is to measure the similarity between the chemical structures of query compounds and the chemical structures of the found reference compounds (hits). A high structural similarity is desirable if the query is not present in the spectral library. Similarity of chemical structures has been measured by the Tanimoto index, calculated from 1365 binary substructure descriptors. The method has been applied to sets of 200–10,000 hitlists obtained with different search methods from a database containing 106,955 compounds. Hitlists with highest structure information have been obtained by using a similarity measure based on the correlation coefficient computed either from spectral features, or from the cubic root of peak intensities for masses up to 200. Frequency distributions of spectral and structural similarities have been investigated and a threshold for the spectral similarity has been derived that in general yields hitlists with structures that are very similar to the query structure. An example compares different search methods.

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

1. Introduction

Mass spectral library search is widely used for the identification of organic compounds. A measured low resolution mass spectrum—mostly obtained by electron impact ionization—is compared with the reference spectra of a database. The resulting hitlist contains the reference spectra most similar to the spectrum of the unknown. If the unknown is present in the library the correct compound often shows a high spectral similarity and appears as the first hit or among the first hits [1,2]. A number of mass spectra databases and powerful software products are available for this purpose and are routinely used. However, if the unknown is not present in the library, spectra similarity hitlists may be less useful. An *interpretative power* is claimed by some mass spectra database systems, meaning that the hitlist structures are often similar to the structure of the query and thereby allow the extraction of essential structure information about the unknown.

Three approaches for interpretative library search systems for mass spectra have essentially contributed to this field. (1) Starting in the seventies, McLafferty and coworkers developed STIRS (self-training and interpretive retrieval system) [3–10]. Several match factors have been defined, each representing a particular type of spectral similarity, and from the resulting hitlist the presence of characteristic substructures in the unknown has been predicted. (2) Around 1980 Henneberg and co-workers [11–14] developed the SISCOM algorithm (search for identical and similar compounds) that has been implemented in the mass spectra database system MassLib [15]. The spectral similarity used has been optimized (but not fully published) with the aim to obtain hitlists containing relevant chemical structure information. By a comparison of the probabilities of substance classes in the database and in the hitlist, structure properties of the unknown can be predicted [16]. (3) A comprehensive investigation by Stein [17] for the NIST Mass Spectral Database and comparisons with other search systems revealed parameters for an optimum transformation of peak intensities. In general, best recognition of substructures from hitlists has been obtained by using a spectra similarity calculated with the square root of the peak intensities. In contrary to search strategies for best identifications of compounds [18,19] a

* 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 © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2004.04.014

weighting of the peak intensities by mass did not improve the results. Also no advantages have been found for a restriction of the considered mass range; however, an increase of the library size improved the classification of substructures.

Common to all three approaches is the fact that different spectra similarity measures are necessary for an optimum identification of compounds, on the one hand, or an optimum interpretative power, on the other hand. Most studies focus on the recognition of single substructures but do not consider the overall similarity of chemical structures.

Recently, a new method has been described to evaluate the interpretative power of spectral similarity searches, and has been applied to infrared spectra [20]. An application of this approach to low resolution mass spectral data is reported here. The similarity between chemical structures has been measured by comparing binary fingerprint vectors that indicate presence or absence of 1365 substructures. The amount of structural information in hitlists has been characterized by an averaged structure similarity calculated from the Tanimoto indices between query and the first hits. Sets of 200–10,000 hitlists for randomly selected query compounds have been investigated with the aim to find good parameters of search methods.

The work is organized as follows. In Section 2 the used similarity measures for mass spectra and for chemical structures are described, and a criterion is defined for the quality of a hitlist. In Section 3 the influences of different weighting schemes for peak intensities and of different mass ranges are reported. Furthermore, the use of spectral features instead of peak intensities for the calculation of spectral similarities is tested. Frequency distributions of spectral and structural similarities have been analyzed, and finally, an example compares different spectral search methods.

2. Methods

A hitlist from a spectral library search contains reference compounds from a database with spectra most similar to the query spectrum. The interpretative power of a spectral search system is the ability of producing hitlists with chemical structures very similar to the structures of query compounds. The used similarity criteria for mass spectra and for chemical structures are described in the following subsections.

2.1. Similarity of mass spectra

A low resolution mass spectrum can be characterized by a vector $\boldsymbol{x}$ with components x_j, with $j = 1, \ldots, k$, and x_j, for instance, being the peak intensities at selected masses m, or weighted peak intensities as described in Section 2.1.2. The vector components may also be spectral features derived from the peak intensity data as described in Section 2.1.3. Many similarity criteria for mass spectra—more or less complicated and more or less documented—have been suggested and are implemented in commercial software [1].

A hitlist obtained by a spectral similarity search contains H hits ranked by decreasing spectral similarity with the query spectrum; the first hit is the reference spectrum most similar to the query spectrum; H is typically between 5 and 50.

2.1.1. Correlation coefficient

A widely used similarity measure—and part of many spectra library search systems—is based on the correlation coefficient, r, given by

$$r = \frac{\boldsymbol{x}_\mathrm{A}^\mathrm{T} \cdot \boldsymbol{x}_\mathrm{B}}{||\boldsymbol{x}_\mathrm{A}|| \cdot ||\boldsymbol{x}_\mathrm{B}||} \tag{1}$$

with $\boldsymbol{x}_\mathrm{A}$ and $\boldsymbol{x}_\mathrm{B}$ being column vectors representing the compared mass spectra A and B, respectively. The Euclidean norm $||\boldsymbol{x}||$ is equivalent to the length of the vector, given by $(\Sigma x_j^2)^{0.5}$ with $j = 1, \ldots, k$. The similarity r can be interpreted in different ways: r is the cosine of the angle between vectors $\boldsymbol{x}_\mathrm{A}$ and $\boldsymbol{x}_\mathrm{B}$; r is the correlation coefficient for a regression line passing the origin, for k points with coordinates $x_{\mathrm{A}j}$ and $x_{\mathrm{B}j}$. It also corresponds to the correlation coefficient of two mean-centered variables given by the components of the vectors $\boldsymbol{x}_\mathrm{A}$ and $\boldsymbol{x}_\mathrm{B}$; r has also been called normalized dot product. Because the vector components (peak intensities, weighted peak intensities, spectral features) used in this work are always positive, r is in the range 0 to +1. For a better readability the used spectral similarity S is scaled to the range 0–1000.

$$S = 1000r \tag{2}$$

2.1.2. Weighting of peak intensities and selection of masses

In a first approach the peak intensities (in percent of the base peak) are used as vector components in the calculation of a spectral similarity. Different weighting schemes—as described by Stein and Scott [18] for compound identification—have been applied to evaluate the significances of mass numbers and peak intensities. A vector component x_j is calculated from peak intensity I_m at mass m by

$$x_j = m^a I_m^b \tag{3}$$

Exponent a has been varied between 0 and 2, and exponent b between 0.01 and 2.

An appropriate selection of masses used for the calculation of spectral similarities may be of great influence, and therefore the upper mass limit has been varied between 100 and 1000. Furthermore, three strategies as implemented in some commercial search systems have been tested [1]. (1) Mode IDENTITY uses all masses covered by the two compared mass spectra, and is successfully used for the identification of compounds if rather pure query spectra are available. (2) Mode FIT uses only masses with a peak in the reference spectrum. This mode is useful if the query spectrum is from a mixture of compounds. The spectral similarity indicates to which extent a reference spectrum is part of the query spectrum. (3) Mode RFIT uses only

masses with a peak in the query spectrum, and has been suggested for unknowns probably not contained in the library. The similarity indicates to which extent the query spectrum is part of the reference spectrum and hopefully the structure of it is part of the reference structure.

2.1.3. Spectral features

A spectral feature is a characteristic number that can be automatically computed from a spectrum. A spectrum is represented by a set of spectral features (feature vector). Aim of the data transformation is to obtain a set of variables that are closer related to chemical structure properties than the original spectral data. Often, nonlinear mathematical transformations are applied, considering spectroscopic facts to some extent. Advantages of this approach for multivariate data analyses of mass spectra have been shown in several applications [21–25]. Early successful uses of spectral features in mass spectra similarity searches have been described, for instance, by Clerc et al. [26,27], and by McLafferty and co-workers [8,10]. Later applications have been reported by Drablos [28], Lebedew and Cabrol-Bass [29], and Varmuza et al. [30].

Most spectral features used have already been described [25], so only a brief overview is given here (Table 1). The number of mass spectral features has been extended to 862, and the spectral similarity S has been calculated from feature vectors. New features have been developed for isotope peak patterns and for characteristic peak groups. All features are scaled to the range 0–100. Let I_m be the intensity of a peak at mass m, normalized to the base peak with an intensity of 100%.

Group 1 contains simple features that are equal to peak intensities at selected single masses. The features in group 2 are peak intensities at single masses but normalized to the local ion current, with the aim to emphasize isolated peaks [26]. The local ion current is the sum of peak intensities in a mass interval $\pm\ \Delta m$ around a considered mass m. Group 3 contains features defined as averaged peak intensities of mass intervals. Features in group 4 are based on logarithmic intensity ratios of peak intensities, given by $\ln I_m/I_{m+\Delta m}$. Peak intensities below 1% are set to 1 avoiding divisions by zero. Group 5 contains features obtained by modulo-14 summation, that means peak intensities at masses with a difference of 14 are summed, resulting in 14 features. Features in group 6 are based on an autocorrelation function of peak intensities for defined mass differences and mass intervals. The spectra type features in group 7 characterize the relative peak intensities in the low mass range, the base peak intensity in percentage of the total intensity sum, and the proportion of peak intensities summed at even mass numbers.

Group 8 comprises features that indicate the presence of specified isotope peak patterns (or any other target peak pattern) located anywhere in the mass spectrum. A feature, x, of this type is calculated by

$$x = 100 \max(r_m^3) \tag{4}$$

with r_m being the correlation coefficient—as defined in Eq. (1)—calculated from the peak intensities of a given target pattern and the peak intensities in the spectrum, starting at mass m. The target pattern is shifted across the spectrum and the maximum correlation coefficient gives the feature [31,32].

Group 9 contains features that indicate the joint presence of peaks at defined mass numbers, for instance, characteristic peak series as given by McLafferty [33]. Because the joint presence of peaks is more important than their intensities, scaled intensities I^c $(0 < c < 1)$ are used in the calculation of the feature, instead of I. Furthermore, intensities equal or smaller than an intensity threshold I_0 are set to zero, and the transformed intensities are scaled to the range 0–100. The resulting intensity, I_{scaled}, is calculated by

$$\begin{aligned} I_{\text{scaled}} &= \frac{100(I - I_0)^c}{(100 - I_0)^c} \quad \text{if} \quad I > I_0 \\ I_{\text{scaled}} &= 0 \qquad\qquad\qquad\ \ \text{if} \quad I \le I_0 \end{aligned} \tag{5}$$

Features, x, of this type are calculated by

$$x = \left(\frac{1}{g}\right) \sum I_{\text{scaled},m} \quad \text{(sum over all } g \text{ target peaks)} \tag{6}$$

Table 1
Mass spectral features

Group no.	Description	z	b
1	Intensities at masses 12, 13, 15, 17, 19–27, 29–31, 33–200	184	0.333
2	Intensities normalized to local ion current for $\Delta m = \pm 3$ at masses 12, 13, 15, 17, 19–27, 29–31, 33–200	184	2
3	Averaged intensities of mass intervals 33–50, 51–70, 71–100, 101–150	4	0.333
4	Logarithmic intensity ratios for mass differences of 1 and 2, and lower masses 39–150	224	1[a]
5	Modulo-14 summation for mass intervals 31–120, 121–800, 31–800	42	0.333
6	Autocorrelation for mass differences 1, 2, 14–60, and mass intervals 31–120, 100–800, 31–800	147	1
7	Spectra type	3	0.333
8	Isotope peak patterns for Cl_1–Cl_5, Br_1–Br_5, and Cl_xBr_y $(x + y = 2, 3, 4, 5)$ up to mass 800	20	1[a]
9	Characteristic peak groups	54	1[a]
Sum		862	–

z, number of features; b, optimum exponent for preceding peak intensity transformation I^b.
[a] Intensity transformation not reasonable for this group or included in feature definition.

with m being a mass of the target peak pattern, and g the number of peaks in the target peak pattern. Tests showed that these features, calculated with $c = 0.2$ and an intensity threshold, I_0, of 5% of the base peak intensity, improved the interpretative power of spectral searches.

2.2. Similarity of chemical structures

Representation of chemical structures and calculation of structural similarities has been performed by the same methods as used in the previous study with infrared spectra [20]. Each chemical structure (two-dimensionally encoded) has been characterized by a vector $\boldsymbol{y}$ with components y_j being binary substructure descriptors. A set of 1365 substructures has been defined for this purpose; y_j is 1 if substructure j is present in the molecule and 0 otherwise [34]. The similarity of two chemical structures, represented by vectors $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$, has been measured by Tanimoto index, $t_{A,B}$ [35]

$$t_{A,B} = \frac{\sum \mathrm{AND}[y_{Aj}, y_{Bj}]}{\sum \mathrm{OR}[y_{Aj}, y_{Bj}]} \quad \text{with} \quad j = 1 \cdots 1365 \tag{7}$$

AND[·, ·] is the result of the logical AND, and OR[·, ·] the result of the logical OR of two binary variables. The Tanimoto index is in the range 0–1, with value 1 if all descriptors are pairwise equal.

The used 1365 substructures have been described elsewhere [20,34] and only a summary is given here. A part of the substructures has been built systematically by using the isomer generator software MOLGEN [36]; others have been defined on the basis of chemical and spectroscopic ideas. Group 1 contains 46 substructures that indicate presence of a certain minimum number of atoms of the elements N, O, S, P, F, Cl, Br, I, B, Si, or any hetero atom. Group 2 contains 78 two-atom substructures. Group 3 contains 404 non-aromatic single rings including heterocycles. Group 4 contains 93 substructures with non-aromatic condensed rings. Group 5 contains 97 substructures with a benzene ring or an *N*-aromatic ring. Group 6 contains 39 bridged ring systems. Group 7 contains 418 non-cyclic substructures, most of them systematically generated. Group 8 contains 153 functional groups not present in other substructure groups.

A quality measure for hitlists has been derived from the Tanimoto indices $t_{q,h}$ calculated for a query structure, q, and hitlist structure h ($h = 1$ to H). The averaged Tanimoto indices, $t_{q,1-h}$ for the first h hits (for instance, $t_{q,1-5}$) has been used as a qualifier for single hitlists. For an evaluation of spectra similarity search methods, random samples with n query spectra have been used. A search method has been characterized by the grand mean of the n values for $t_{q,1-h}$ defined by [20]

$$T(h) = \left(\frac{1}{n}\right) \sum t_{q,1-h}, \quad q = 1 \cdots n \tag{8}$$

Throughout this study $T(5)$ has been used to compare methods, which is the averaged structural similarity between query structures and the structures of the corresponding first five hits, averaged over n (200–10,000) randomly selected query compounds. The query compound itself has always been found as the first hit, and has been excluded from the hitlist in order to simulate a situation with the unknown not contained in the library.

2.3. Database and software

The mass spectra database used consists of 106,955 compounds and is part of the NIST Mass Spectral Database [37]. Nominal molecular masses are between 2 and 1218 with a mean of 271.6; the ranges of common elements in the compounds are $C_{0–72}$, $H_{0–218}$, $N_{0–12}$, $O_{0–25}$, $F_{0–45}$, $Cl_{0–12}$, $Br_{0–12}$, $S_{0–14}$,$P_{0–10}$. Chemical structures were encoded in the Molfile format [38,39]. The mass spectra contain low resolution peak list data with integer mass numbers and intensities in the range 1–9999.

The 1365 substructures used to characterize molecular structures have been also encoded in Molfile format. Binary substructure descriptors have been calculated by software SubMat [34,40]; typical computing time with a Pentium IV 2 GHz is 40 ms for one molecular structure and 1365 substructures. SubMat can work stand-alone or can be started and controlled from other software, for instance, from a Matlab program. Software for structure and spectra handling, and for spectral and structural similarity searches has been written in Matlab 6.0 or Microsoft Visual C++.

From the originally defined 1365 substructures a set of 1259 (92.2%) is present in the compounds of the database [34]. A few small substructures are contained in more than 99% of the database structures, however, 68% of the substructures are contained in only 5% of the database structures. The number of substructures per database structure varies between 0 and 287 with a median of 78.

3. Results

For first investigations five random samples (query sets) each containing 200 query compounds have been selected from the database. For each query compound a hitlist containing 50 compounds has been determined. The section is organized as follows. First, peak intensities are used with different weighting schemes and mass ranges; second, results obtained with spectral features are reported; third, selected best methods are extensively tested with 10,000 hitlists; finally, an example compares hitlists obtained by different spectra similarity searches.

3.1. Similarity searches using peak intensities

Fig. 1 shows results obtained by a standard search technique (encoded by M1000) using peak intensities in percent of base peak at all mass numbers (mode IDENTITY). The averaged structural similarities $T(h)$ are highest for the first hits and reach a maximum of 0.6. The differences of $T(h)$

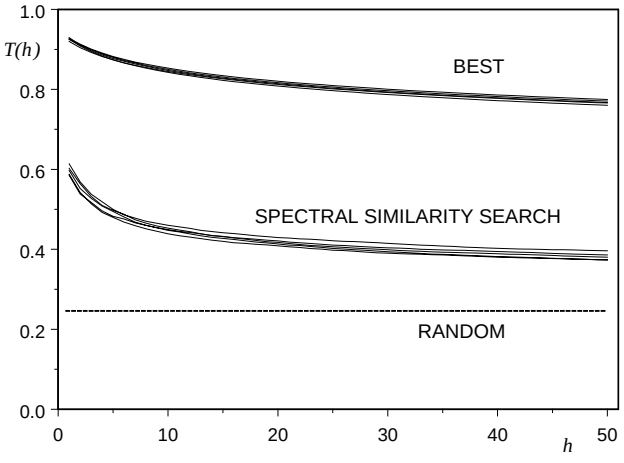


Fig. 1. Averaged similarity $T(h)$ between query structures and the corresponding h hitlist structures for spectra similarity search method M1000. Results from five parallel experiments each with 200 randomly selected query compounds are shown. BEST, pseudo hitlists containing the database structures with maximum structural similarity to query structure; RANDOM, randomly selected pseudo hitlists.

between the five parallel experiments (each 200 queries) are not larger than 0.03. These results can be compared with two extreme cases. One is the result from pseudo hitlists containing randomly selected compounds from the database. In this case the structural similarity between hitlist compounds and the query corresponds to the average Tanimoto index of randomly selected structure pairs from the database, which is 0.247. The structural similarities obtained by spectral similarity searches are considerable higher than this minimum value. The other extreme is the result from hitlists containing the database compounds with highest structural similarity with the query compounds. Such pseudo hitlists exhibit the maximum structural similarities that are possible with the used database; they can be determined only in tests with known query compounds. These maximum values are considerable higher than those from hitlists obtained by mass spectral similarity. They are an upper limit, given by the composition of the database. In general, this limit cannot be achieved by a spectral similarity search because spectral data do not contain full structure information, and any spectra similarity measure is capable of utilizing only a part of the inherent structure information.

Parameters for the mass spectral similarity—as described in Section 2.1.2—are the exponents a and b for mass and intensity, respectively, the considered mass range, and the applied mass selection mode. Also an intensity threshold for the used peaks may influence the performance. An exhaustive search for an optimum set of parameters or the application of optimization methods would be very time-consuming; therefore a good set of these parameters has been determined as follows. For each tested set of parameters a random sample with 200 query compounds has been used, and quality measure $T(5)$ has been applied. Table 2 shows the results for varying the exponent a between 0 and 2, and the exponent b between 0.5 and 2, applied to the mass intervals 1–200 and 1–1000, and separately tested for the three modes IDENTITY, FIT, and RFIT. Conclusions from these experiments are as follows. (1) The IDENTITY mode is best, while the FIT mode gives results that are similar to random hitlists. Mode RFIT is close to IDENTITY but is in all investigated cases below IDENTITY. (2) The mass should not be included into weighting, because $a = 0$ is best. (3) The dynamic range of the peak intensities should be reduced, because $b = 0.5$ (square root of intensities) is best. (4) The mass range 1–200 is better than 1–1000 in all cases shown for IDENTITY and RFIT.

Based on these results the influences of exponent b (for intensity transformation) and of the mass range have been investigated in more detail. Fig. 2 shows the effect of changing exponent b between 0.01 and 2 for $a = 0$ or 1, and for masses up to 200 or up to 1000. Independently from the value of exponent a and the two mass ranges, the best values for exponent b are 0.333 or 0.5. For $a = 0$ the influence of an intensity threshold on the optimum value for b

Table 2
Influence of weighting peak intensities by $m^a I^b$, different upper mass limits (m/z 200 or 1000), and mode of mass selection (IDENTITY, FIT, RFIT)

a	b	Grand mean Tanimoto index, $T(5)$					
		IDENTITY		FIT		RFIT	
		m/z 1–1000	m/z 1–200	m/z 1–1000	m/z 1–200	m/z 1–1000	m/z 1–200
0	0.5	0.555	0.577	0.329	0.265	0.473	0.489
0	1	0.480	0.516	0.260	0.268	0.449	0.471
0	2	0.404	0.423	0.244	0.266	0.400	0.409
1	0.5	0.493	0.531	0.281	0.269	0.432	0.465
1	1	0.467	0.502	0.246	0.273	0.431	0.462
1	2	0.411	0.435	0.235	0.270	0.386	0.405
2	0.5	0.448	0.499	0.243	0.259	0.392	0.437
2	1	0.430	0.476	0.237	0.269	0.406	0.446
2	2	0.408	0.429	0.231	0.304	0.382	0.402

A random sample with 200 query compounds has been used. $T(5)$ is the averaged structural similarity between the query and the first five hits.

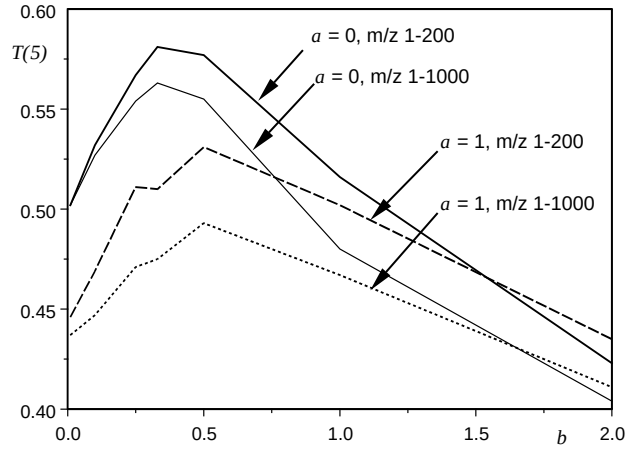


Fig. 2. Influence of weighting peak intensities by $m^a I^b$, for upper mass limits of 200 and 1000. Random sample with 200 query compounds; mode IDENTITY; $T(5)$, average structural similarity between query and first five hits.

has been checked. Peaks have been deleted that are smaller than 1 or 5% of the base peak intensity. For both thresholds and both mass ranges the optimum for b is 0.333, as found without applying a threshold. For instance, for a mass range up to 200 and a threshold of 5%, $T(5)$ has the values 0.516, 0.529, and 0.525 for $b = 0.1$, 0.333, and 0.5, respectively. The corresponding values for $T(5)$ without a threshold are 0.532, 0.581, and 0.577, showing a slight decrease of the performance if peaks smaller than 5% are deleted.

Fig. 3 shows the influence of the upper mass limit, varied between 100 and 300. Independently from exponent b—exponent a was kept constant at the previously found optimum value zero—an upper mass limit of 200 has been found to be optimal.

In summary, best hitlists in the sense of high structural similarity with the query compound, have been obtained with the cubic root of peak intensities, masses up to 200, and the IDENTITY mode (this method is encoded by M200). The averaged structural similarity $T(5)$ for this method has been determined for five query sets each containing 200 randomly selected compounds. Results are 0.581, 0.584, 0.600, 0.574, and 0.585, with a mean of 0.585; the last number corresponds to 65.1% of the best result possible with the used database. More complicated spectra similarity search methods have to compete with this value.

3.2. Similarity searches using spectral features

Transformation of the mass spectra into the 862 spectral features, as described in Section 2.1.3, and application of S as the similarity measure gives averaged structure similarities $T(5)$ of 0.588, 0.594, 0.590, 0.581, and 0.588 for the five query sets (each containing 200 queries). The mean of 0.588 is almost identical with the corresponding result, 0.585, obtained with the cubic root of peak intensities up to mass 200.

Considering the improvements obtained by applying weighting schemes for intensities and a reduced mass range, the same transformations have been tested as a preceding step before the calculation of features. The transformations have been optimized with a set of 200 queries and then the results have been confirmed with 1000 queries. No improvements could be achieved by reducing the upper mass limit to 200 or 300. However, advantages have been found for intensity transformations I^b. Optimum values for exponent b have been searched separately for each feature group, testing the values 0.25, 0.333, 0.5, 1, and 2. For feature group 1 with the features being peak intensities at selected masses, $b = 0.333$ was found to be best, corresponding to the results in Section 3.1. The same transformation has

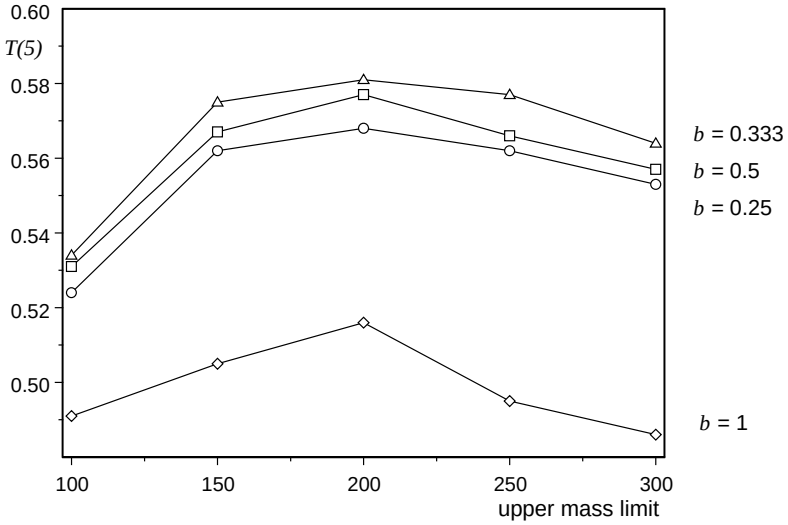


Fig. 3. Influence of upper mass limit for different exponents b (for intensity transformation) and exponent a set to 0. Random sample with 200 query compounds; mode IDENTITY; $T(5)$, average structural similarity between query and first five hits.

been found to be best for modulo-14 features and some other groups. For group 2 (intensities normalized to local ion current) an exponent of 2 gave best results; this value discriminates small peaks. The combined use of the best values for *b*, as given in Table 1, increased the structural similarity between queries and hits (method F862B). The structure similarity measure *T*(5) for the five query sets is 0.600, 0.612, 0.611, 0.600, and 0.601, respectively, with a mean of 0.605. This is a significant increase compared with the value 0.585 obtained by method M200; the statistical t-test gives a probability of 0.035 for the zero hypothesis (both methods equal, $n = 1000$).

In summary, use of spectral features in mass spectra similarity searches yields better hitlists—in the sense of high structural similarity with the query compound—than using weighted peak intensities.

3.3. Evaluation of selected best search methods

Three of the tested spectra similarity search methods have been evaluated in more detail using up to 10,000 hitlists. The methods selected are (1) the standard method M1000 using peak intensities *I* in percent base peak intensity and all masses 1–1000; (2) method M200 using transformed peak intensities $I^{0.333}$ up to mass 200; and (3) method F862B using 862 spectral features calculated from transformed peak intensities I^b as given in Table 1.

3.3.1. Intensity threshold

Mass spectral peaks with small intensities may be caused by impurities in the sample or by background of the instrument. The influence of deleting peaks that are smaller than 1, 2, or 5% of the base peak intensity has been investigated with 1000 randomly selected query compounds. Reference spectra and query spectra have been treated in the same way; deleting no peaks corresponds to a threshold of 0.01% because the peak intensities in the database are between 1 and 9.999. Fig. 4 shows that the performance, measured by *T*(5), decreases with increasing intensity threshold. This effect is pronounced for methods F862B and M200, while method M1000 only shows a small decrease at a threshold of 5%. The large number of peaks used by M1000 makes the method more robust than the two others. However, method F862B gives at all tested thresholds higher structural similarities than methods M200 and M1000.

3.3.2. Distribution of the structural similarity

The diversity of the structures has been characterized by a calculation of the Tanimoto indices for 10,000 pairs of randomly selected different structures. The obtained frequency distribution has a skewed bell shape with a mean of 0.247, and a standard deviation of 0.115 (Fig. 5). Tanimoto indices, *t*, above 0.60 indicate a significantly high structural similarity because only 1% of the random pairs yield higher values. The distributions of the structure similarity between query and first hit have been estimated from 10,000 hitlists (for randomly selected query compounds). Methods M200 and F862B give similar distributions, showing larger areas for $t > 0.6$ than the distribution for method M1000. The percentage of hitlists with $t > 0.6$ for query and first hit is 64 and 67% for the methods M200 and F862B, respectively, but only 49% for method M1000.

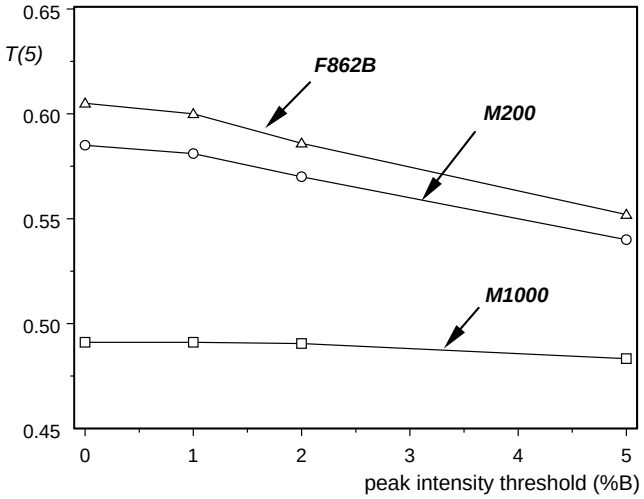


Fig. 4. Influence of peak intensity threshold on the averaged structure similarity. Random sample with 1000 query compounds; *T*(5), average structural similarity between query and first five hits; methods compared are M1000, M200, and F862B.

Fig. 6 shows the averaged structural similarity, *T*(*h*), between query and hits as a function of the number of considered hits, *h*, for the three selected methods. These results have been obtained from 10,000 hitlists. *T*(*h*) is highest for the first hits and reaches a maximum of 0.7; method F862B is best.

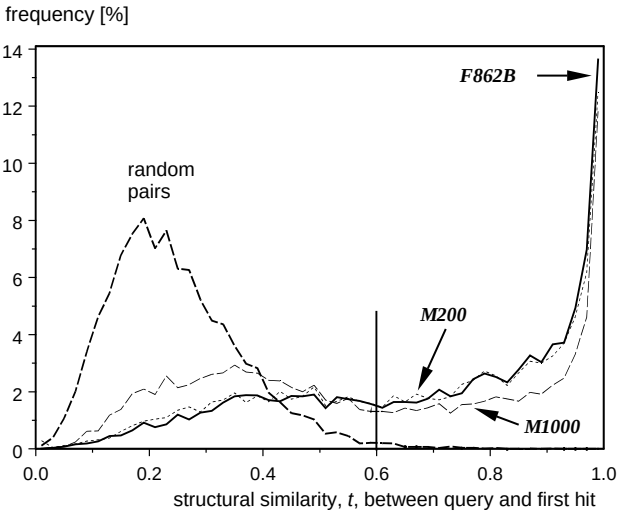


Fig. 5. Frequency distribution of Tanimoto indices, *t*, for 10,000 pairs of randomly selected different structures, and for query and first hit for 10,000 randomly selected queries and the methods M1000, M200, and F862B. The frequency is given in percent for 50 intervals of *t*, each 0.02 units wide. One percent of the random pairs have a structural similarity of more than 0.6.

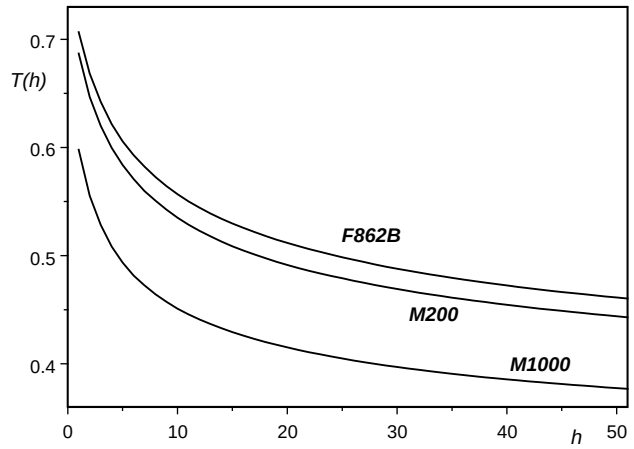


Fig. 6. Similarity $T(h)$ between query structures and the corresponding h hitlist structures for the methods M1000, M200, and F862B, estimated with 10,000 randomly selected query compounds.

3.3.3. Relationship between spectral and structural similarity

An approximate minimum spectral similarity that is generally required to obtain a sufficiently high structural similarity between query and first hit(s) has been estimated as follows. The region 700–1000 of the spectral similarity, S, has been divided into 15 intervals, each 20 units wide. In each interval the Tanimoto indices, $t_{q,1}$, have been averaged and plotted versus the center of the interval. As shown in Fig. 7, the general trend for all three selected methods is an increase of the structural similarity, $t_{q,1}$, with increasing spectral similarity, S. The three diagrams in Fig. 7 also contain cumulative frequency distributions, F, for the spectral similarities. F is the percentage of hitlists reaching a minimum spectral similarity. For instance, using method M1000 the spectra similarity of the first hit is >800 in 79%, and >900 in 51% of the tested 10,000 hitlists.

The plots of $t_{q,1}$, and F can be used to estimate a minimum spectral similarity for obtaining a minimum structural similarity. Arrows in the diagrams guide the interpretation. In Section 3.3.2 has been shown that Tanimoto indices above 0.6 occur in only 1% of randomly selected structure pairs. This critical structural similarity is approximately reached by the first hit at spectral similarities of 920, 830, and 920 applying the methods M1000, M200, and F862B, respectively. From the cumulative frequency distributions, F, the percentage of hitlists can be derived that reach the critical structural similarity. With method M1000 48% of the hitlists have a spectral similarity of >920 between query and first hit; for methods M200 and F862B the part of successful hitlists is 78 and 87%, respectively. These results again show that hitlists with highest structure information are obtained by using spectral features for the spectra similarity calculation (method F862B). The same ranking of the three methods is obtained if not only the first hit is considered but the averaged structural similarity between query and the first five hits.

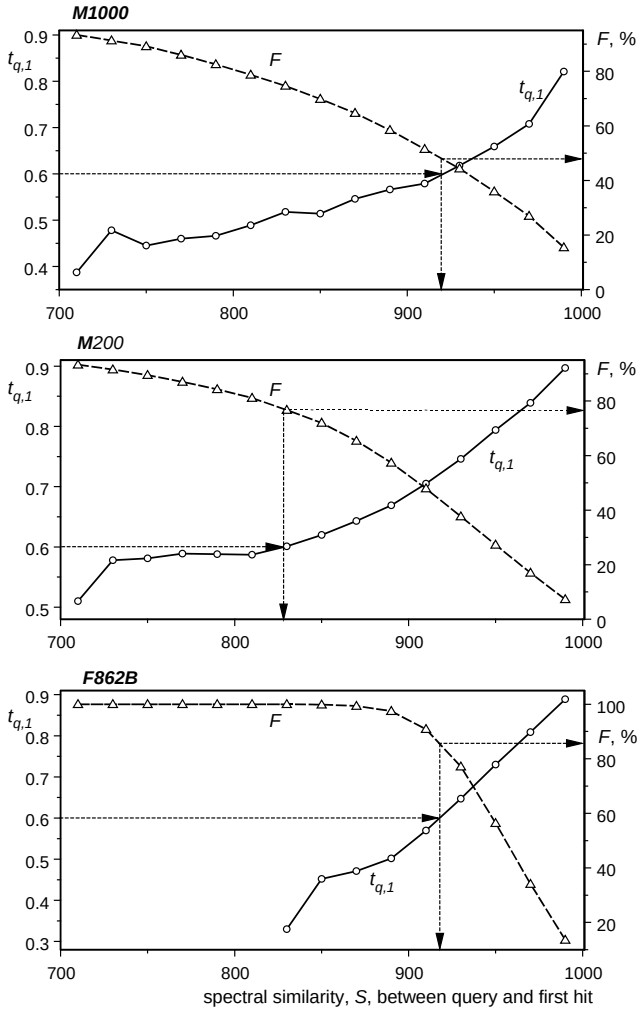


Fig. 7. Relationship between spectral similarity, S, and structural similarity, $t_{q,1}$, between query and first hit for the three methods M1000, M200, and F862B, estimated with 10,000 randomly selected query compounds. F is the cumulative frequency of hitlists with a spectral similarity between query and first hit above a threshold. Data are presented for 15 intervals with spectral similarities between 700 and 1000, each interval 20 units wide.

The found acceptable minimum spectral similarities, however, have to be used cautiously because of the large variation of the structural similarity between query and first hits, even in cases with high spectral similarities. For instance, with method F862B a subset of 4511 hitlists (from the investigated 10,000) gives spectral similarities >950 for the first hit. The structural similarity between query and first hit has a satisfying high mean of 0.819 but shows a skewed, rather broad distribution. In 51% of the 4511 hitlists the structural similarity between query and first hit is between 1 and 0.9, in 33% between 0.9 and 0.6, in 10% between 0.6 and 0.4, in 5% between 0.4 and 0.2, and in 1% below 0.2. That means in 84% the critical structural similarity of 0.6 is reached. Accepting hitlists with spectral similarities >900 for the first hit—actually 91% of the hitlists reach this value—reduces the performance to 70%.

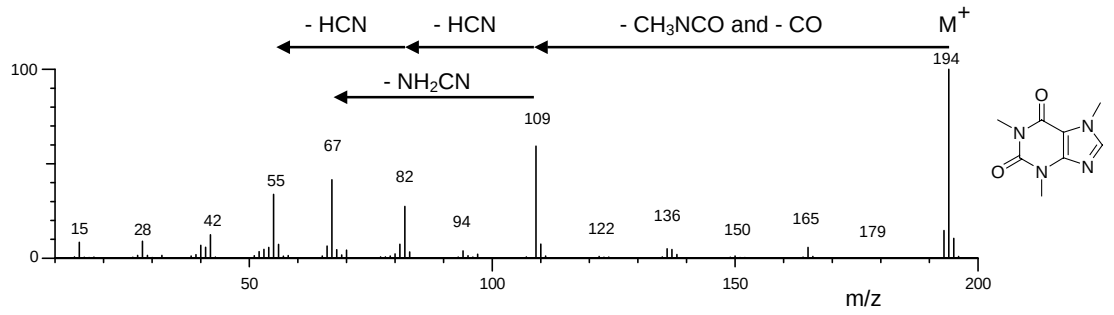


Fig. 8. Mass spectrum of query compound caffeine.

3.4. Hitlist example

A selection of examples for spectra similarity searches is always arbitrary. The compound caffeine (CAS reg. no. 58-08-2, molecular mass 194, Fig. 8) has been chosen because it is well known, has a clear mass spectrum and is typical for the trends found in the statistical investigation of 10,000 hitlists. The NIST Mass Spectral Database [37] contains 741 compounds with molecular mass 194; however, the mass spectrum used as a fingerprint is very selective because only two database spectra have the base peak at *m*/*z* 194 and another peak at *m*/*z* 109 in the intensity interval 50–80% (caffeine and the similar compound proxyphilline). Characteristics in the mass spectrum of caffeine are eliminations of non-radical small molecules, such as methylisocyanate, cyanamide, HCN and CO.

In Fig. 9 hitlists with the first five hits are shown for the more extensively tested methods M1000 and F862B, and

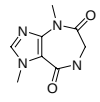
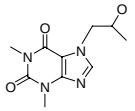
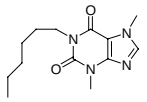
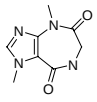
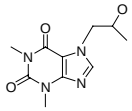
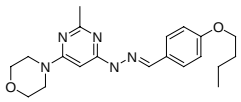
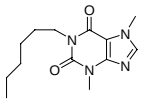
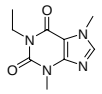
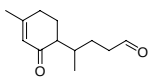
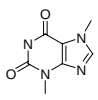
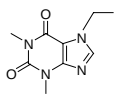
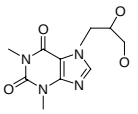
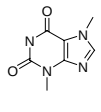
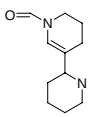
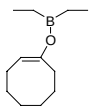
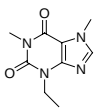
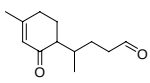
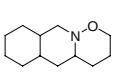
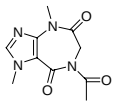
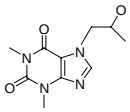
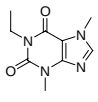
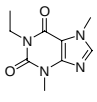
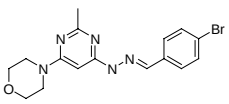
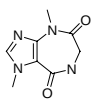
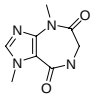
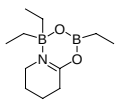
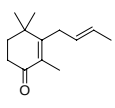
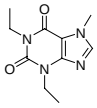
Hit	***M1000*** NIST (106955)	***M200-RFIT*** NIST (106955)	***F862B*** NIST (106955)	***A*** NIST (107886)	***B*** Wiley *(130542)*	***Best*** NIST(106955)
1	 *t* = 0.786 *S* = 843	 *t* = 0.884 *S* = 684	 *t* = 0.884 *S* = 939	 *t* = 0.786 *S*$_A$ = 695	 *t* = 0.884	 *t* = 0.989
2	 *t* = 0.304 *S* = 791	 *t* = 0.884 *S* = 850	 *t* = 0.989 *S* = 939	 *t* = 0.147 *S*$_A$ = 637	 *t* = 0.958	 *t* = 0.989
3	 *t* = 0.400 *S* = 776	 *t* = 0.843 *S* = 842	 *t* = 0.957 *S* = 930	 *t* = 0.308 *S*$_A$ = 615	 *t* = 0.139	 *t* = 0.989
4	 *t* = 0.147 *S* = 764	 *t* = 0.137 *S* = 840	 *t* = 0.786 *S* = 924	 *t* = 0.884 *S*$_A$ = 605	 *t* = 0.989	 *t* = 0.989
5	 *t* = 0.301 *S* = 763	 *t* = 0.786 *S* = 840	 *t* = 0.786 *S* = 923	 *t* = 0.273 *S*$_A$ = 594	 *t* = 0.127	 *t* = 0.989
$t_{q,1\text{-}5}$	0.388	0.706	0.880	0.480	0.619	0.989

Fig. 9. Search results for query test compound caffeine. Spectra similarity search methods used are M1000, M200-RFIT, F862B, and two commercial mass spectra database systems (A, B). The last column shows optimum results obtained by a structure similarity search. *t*, Tanimoto index between query and hit; $t_{q,1\text{–}5}$, averaged structural similarity between query and the first five hits; *S*, spectral similarity; S_A, spectral similarity used in method A; method B does not output an overall spectral similarity.

for M200-RFIT. The latter uses the RFIT mode, mass range 1–200, and the square root of peak intensities which yielded the best results for this mode (Table 2). Furthermore, hitlists are presented in Fig. 9 that have been obtained by applying two commercial database systems (methods A and B). The last column contains the optimum result which is constituted by the compounds from the used database with structures most similar to the query; they all have the same Tanimoto index of 0.989, compared with caffeine, and their sequence is therefore arbitrary.

The hitlist obtained by method M1000 does not contain compounds with the caffeine ring system; only the first hit shows a high structural similarity (0.786); the mean of the structural similarities between query and the first five hits, $t_{q,1–5}$, is only 0.388—which is the poorest result of the compared methods. Method M200-RFIT gives much better results with $t_{q,1–5}$ equal to 0.706; the first three hits contain the caffeine skeleton. Method F862B gives the best hitlist with $t_{q,1–5}$ equal to 0.880. The first three hits are alkyl-substituted caffeine skeletons, one of them (hit 2) is among the best five structures.

A comprehensive test of commercial systems was not possible because of the lack of download facilities of hitlists data; furthermore, different spectral libraries are used. Therefore the example does not evaluate the two methods. With method A the hits 1 and 4 provide useful structure information; the other hits are misleading, and $t_{q,1–5}$, is only 0.480. Method B is better, containing the correct ring system in three hits; however, the structures of two other hits are very different from the query, resulting in 0.619 for $t_{q,1–5}$. In summary, for this example method F862B gave best results.

4. Summary and conclusions

A method for a quantitative evaluation of the *interpretative power* of spectra similarity searches, recently described and applied to infrared spectra [20], has been extensively tested with mass spectra. A high interpretative power yields hitlists that contain compounds with structures very similar to the corresponding query structure; that means the library search method is powerful even in cases the query compound is not contained in the database. The similarities of chemical structures between a query compound and the found hitlist compounds have been characterized by an averaged Tanimoto index, calculated from vectors defined by 1365 binary substructure descriptors. The used similarity measure for mass spectra is based on the correlation coefficient. This measure has been calculated either with peak intensities, weighted peak intensities or spectral features.

The concluding results are based on 10,000 hitlists for randomly selected query compounds. For all applied methods the first hits (corresponding to the most similar spectra) yield highest structural similarity with the query compound. When using peak intensities (in percent of the base peak intensity) in the calculation of the spectral similarity, best results have been obtained with the cubic root of the peak intensities and using a mass range up to 200 (method M200). Significantly poorer results have been obtained, for instance, with peak intensities weighted by the mass of the peak or by considering a larger mass range (for instance, up to 1000, method M1000). The use of spectral features in spectral similarity calculations (method F862B) yielded the best results in terms of structure information. Therefore, this method can be recommended if the aim of a spectral library search is not the identification of an unknown (because it may not be in the database) but gaining chemical structure information from the hitlist compounds.

From the frequency distributions of the spectral and the structural similarities a threshold for a minimum necessary spectral similarity has been estimated. If the spectral similarity between query and the first hit is above this threshold then the structural similarity is very good with a high probability. An extensive comparison of commercial library search systems would require tools for a download of chemical structures and other data from automatically produced hitlists.

Even for a high spectral similarity the structural similarity between query and hits may be low in some cases—and vice versa. For instance, with the best method F862B 45% of the 10,000 hitlists show a high spectral similarity of at least 950 between query and first hit; in this group 84% reach a highly significant structural similarity of at least 0.6 between query and first hit. A careful inspection of the hitlist is therefore advisable, eventually supported by chemometric approaches. Preliminary tests demonstrate the usefulness of methods such as cluster analysis of the hitlist structures [34], a graph representation of the topological hierarchy of the hitlist structures [41], or a joint mapping of spectral and structural data by PLS [25]. Multivariate classification models could be generated from hitlist compounds to predict presence or absence of substructures in the query compound. Such local models are complementary to global approaches [22,42,43] that use random samples from the whole database for the generation of spectral classifiers.

Acknowledgements

The authors thank S. Stein for providing the NIST Mass Spectral Database, A. Kerber and R. Laue for the isomer generator software MOLGEN, as well as H. Scsibrany, and S. Qehaja for collaboration. The work was supported by the Austrian Science Fund, project P14792-CHE.

References

[1] K. Varmuza, in: J.C. Lindon, G.E. Tranter, J.L. Holmes (Eds.), Encyclopedia of Spectroscopy and Spectrometry, Academic Press, London, 2000, pp. 232–243.

[2] A.N. Davies, in: G. Gauglitz, T. Vo-Dinh (Eds.), Handbook of Spectroscopy, vol. 2, Wiley-VCH, Weinheim, 2003, pp. 488–504.
[3] K.S. Kwok, R. Venkataraghavan, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 95 (1973) 4185–4194.
[4] H.E. Dayringer, F.W. McLafferty, R. Venkataraghavan, Org. Mass Spectrom. 11 (1976) 895–900.
[5] H.E. Dayringer, F.W. McLafferty, Org. Mass Spectrom. 11 (1976) 543–551.
[6] H.E. Dayringer, G.M. Pesyna, R. Venkataraghavan, F.W. McLafferty, Org. Mass Spectrom. 11 (1976) 529–542.
[7] H.E. Dayringer, F.W. McLafferty, Org. Mass Spectrom. 12 (1977) 53–54.
[8] K.S. Haraki, R. Venkataraghavan, F.W. McLafferty, Anal. Chem. 53 (1981) 386–392.
[9] F.W. McLafferty, D.B. Stauffer, J. Chem. Inf. Comput. Sci. 25 (1985) 245–252.
[10] F.W. McLafferty, S.Y. Loh, D.B. Stauffer, in: H.L.C. Meuzelaar (Ed.), Computer-Enhanced Analytical Spectroscopy, vol. 2, Plenum Press, New York, 1990, pp. 163–181.
[11] H. Damen, D. Henneberg, B. Weimann, Anal. Chim. Acta 103 (1978) 289–302.
[12] D. Henneberg, in: A. Quayle (Ed.), Advances in Mass Spectrometry, vol. 8B, Heyden, London, 1980, pp. 1511–1531.
[13] L. Domokos, D. Henneberg, B. Weimann, Anal. Chim. Acta 165 (1984) 61–74.
[14] L. Domokos, D. Henneberg, Anal. Chim. Acta 165 (1984) 75–86.
[15] MassLib, Mass spectral database system, MSP Kofel, http://www.msp.ch, Zollikofen, Switzerland, 2003.
[16] D. Henneberg, B. Weimann, U. Zalfen, Org. Mass Spectrom. 28 (1983) 198–206.
[17] S.E. Stein, J. Am. Soc. Mass Spectrom. 6 (1995) 644–655.
[18] S.E. Stein, D.R. Scott, J. Am. Soc. Mass Spectrom. 5 (1994) 859–866.
[19] S.E. Stein, J. Am. Soc. Mass Spectrom. 5 (1994) 316–323.
[20] K. Varmuza, M. Karlovits, W. Demuth, Anal. Chim. Acta 490 (2003) 313–324.
[21] D. Cabrol-Bass, C. Cachet, C. Cleva, A. Eghbaldar, T.P. Forrest, Can. J. Chem. 73 (1995) 1412–1426.
[22] K. Varmuza, W. Werther, J. Chem. Inf. Comput. Sci. 36 (1996) 323–333.
[23] K. Varmuza, Anal. Sci. 17 (2001) i467–i470.
[24] H. Yoshida, R. Leardi, K. Funatsu, K. Varmuza, Anal. Chim. Acta 446 (2001) 483–492.
[25] W. Werther, W. Demuth, F.R. Krueger, J. Kissel, E.R. Schmid, K. Varmuza, J. Chemometrics 16 (2002) 99–110.
[26] F. Erni, J.T. Clerc, Helv. Chim. Acta 55 (1972) 489–500.
[27] P.R. Naegeli, J.T. Clerc, Anal. Chem. 46 (1974) 739A–744A.
[28] F. Drablos, Anal. Chim. Acta 256 (1992) 145–151.
[29] K.S. Lebedew, D. Cabrol-Bass, J. Chem. Inf. Comput. Sci. 38 (1998) 410–419.
[30] K. Varmuza, J. Kissel, F.R. Krueger, E.R. Schmid, in: E. Gelpi (Ed.), Advances in Mass Spectrometry, vol. 15, Wiley, Chichester, 2001, pp. 229–246.
[31] R.J. Anderegg, Anal. Chem. 53 (1981) 2169.
[32] K. Varmuza, Fresenius Z. Anal. Chem. 322 (1985) 170–174.
[33] F.W. McLafferty, Interpretation of Mass Spectra, University Science Books, Mill Valley, CA, 1980.
[34] H. Scsibrany, M. Karlovits, W. Demuth, F. Müller, K. Varmuza, Chemom. Intell. Lab. Syst. 67 (2003) 95–108.
[35] M.A. Farnum, R.L. DesJarlais, D.K. Agrafiotos, in: J. Gasteiger (Ed.), Handbook of Chemoinformatics, vol. 4, Wiley-VCH, Weinheim, 2003, pp. 1640–1686.
[36] A. Kerber, R. Laue, Software MOLGEN, Isomer generator, Institute for Mathematics II, University of Bayreuth, http://www.mathe2.uni-bayreuth.de/molgen4/, Bayreuth, Germany, 2000.
[37] NIST, Mass Spectral Database 98, National Institute of Standards and Technology, http://www.nist.gov/srd/nist1a.htm, Gaithersburg, MD, 1998.
[38] A. Dalby, J.G. Nourse, W.D. Hounshell, A.K.I. Gushurst, D.L. Grier, B.A. Leland, J. Laufer, J. Chem. Inf. Comput. Sci. 32 (1992) 244–255.
[39] MDL-Information-Systems-Inc., CT file format, MDL Information Systems Inc., http://www.mdli.com/downloads/literature/ctfile.pdf, San Leandro, CA, 2002.
[40] 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.
[41] K. Varmuza, H. Scsibrany, J. Chem. Inf. Comput. Sci. 40 (2000) 308–313.
[42] K. Varmuza, P. Penchev, F. Stancl, W. Werther, J. Mol. Struct. 408–409 (1997) 91–96.
[43] K. Varmuza, P. He, K.T. Fang, J. Data Sci. 1 (2003) 391–404.