CASE VIA MS:
RANKING STRUCTURE CANDIDATES BY MASS SPECTRA

Adalbert Kerber[†], Markus Meringer[‡,1], Christoph Rücker[§]

[†]Department of Mathematics, University of Bayreuth,
95440 Bayreuth, Germany

[‡]Department of Medicinal Chemistry, Kiadis B.V., Zernikepark 6–8,
9747 AN Groningen, The Netherlands

[§]Biocenter, University of Basel, Klingelbergstrasse 70,
4056 Basel, Switzerland

April 19, 2006

ABSTRACT. Two important tasks in computer–aided structure elucidation are the generation of candidate structures from a given molecular formula, and the ranking of structure candidates according to compatibility with an experimental spectrum.

Candidate ranking with respect to electron impact mass spectra is based on virtual fragmentation of a candidate structure and comparison of the fragments' isotope distributions against the spectrum of the unknown compound, whence a structure–spectrum compatibility matchvalue is computed. Of special interest is the matchvalue's ability to distinguish between the correct and false constitutional isomers.

Therefore a quality score was computed in the following way: For a (randomly selected) spectrum–structure pair from the NIST MS library all constitutional isomers are generated using the structure generator MOLGEN. For each isomer the matchvalue with respect to the library spectrum is calculated, and isomers are ranked according to their matchvalues. The quality of the ranking can be quantified in terms of the correct structure's relative ranking position (RRP). This procedure was repeated for 100 randomly selected spectrum–structure pairs belonging to small organic compounds. In this first approach the RRP of the correct isomer was 0.27 on average.

Keywords: computer–aided structure elucidation, electron impact mass spectrometry, spectrum–structure compatibility matchvalue, constitutional isomers, structure generation

---

[1]Corresponding author e–mail: m.meringer@kiadis.com; phone: +31 (0)50 5474286; fax: +31 (0)50 5474290

1. Introduction

Computer–aided structure elucidation (CASE) could be of immense importance for present–day drug discovery programs. Thanks to modern screening methods a large number of biologically active compounds can be found in a short time, especially when natural product extracts are considered. Structure elucidation then becomes a serious bottleneck in the drug discovery workflow.

Due to its high sensitivity and selectivity mass spectrometry has the potential to become an analytical key method for elucidation of unknown structures. Mass spectrometers are typically coupled to devices for compound separation, e.g. GC or LC. Two–dimensional separation techniques such as GC×GC became available recently. Allowing separation of complex mixtures with a precision unseen hitherto, such methods produce a plethora of data that clearly requires handling by computer.

In mass spectrometry, soft ionization methods help to preserve the molecular ion, and high resolution techniques allow to determine the molecular formula from the molecular ion's exact mass. In this paper we investigate the ability of low resolution 70eV electron impact mass spectrometry (EI–MS) for distinguishing constitutional isomers.

Typically library–based systems are used for this purpose (e.g. [1]). Hereby a measured spectrum is compared against a large database that stores spectrum–structure pairs. A library search returns the structures belonging to the library spectra that show highest similarity to the measured spectrum.

Obviously for successful library searching the compound under investigation has to be included in the library. However, for a minor fraction only of known chemical compounds a spectrum is deposited in a database, and known compounds themselves are a minority among possible compounds [2]. Therefore library search is destined to failure in most cases, in particular if potentially new chemical entities are to be identified.

An alternative approach is de novo structure elucidation. De novo structure elucidation tries to derive the analyte's structure directly from its spectroscopic data. Following the ideas of [3] such an approach can be divided into three steps:

- Spectra interpretation extracts structural properties from spectral data. In MS this can be done by a set of MS classifiers, e.g. as described in [4, 5].
- Structure generation constructs candidate structures, typically represented by molecular graphs [6] that agree with the structural properties found above.
- Spectra simulation computes virtual spectra from candidate structures. These are finally compared to the experimental
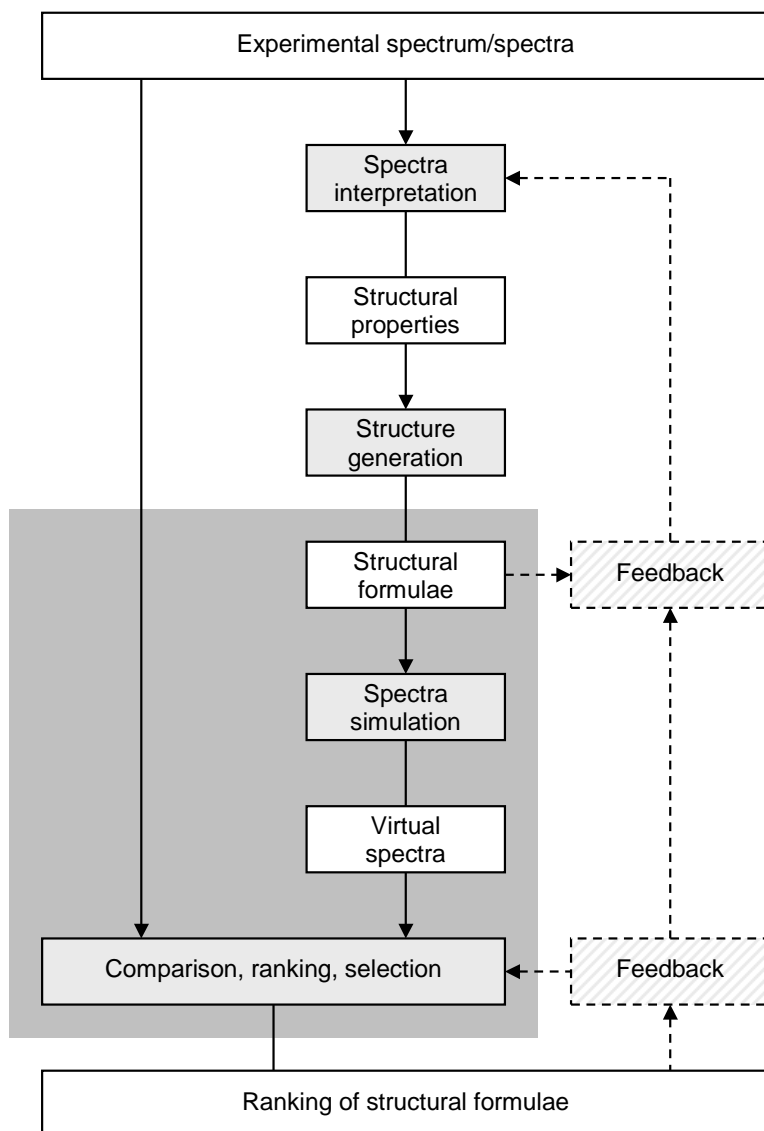
FIGURE 1. General flowchart of CASE

spectrum, and structure candidates are ranked and selected according to the match of experimental and virtual spectrum. We summarize these tasks as spectrum–structure compatibility verification.

Figure 1 illustrates this workflow. Data is always represented by white boxes, algorithmic parts by light grey boxes. Some feedback might be required, represented by dashed arrows and boxes.

A first implementation of all three steps within one computer program has been realized in the software MOLGEN–MS [7]. However, further research is necessary to improve the chemistry-related tasks

spectrum interpretation and spectrum-structure compatibility verification.

For the first step typically methods of supervised statistical learning are used, such as linear discriminant analysis or classification by artificial neural networks, classification trees or support vector classifiers. However, all these methods suffer from classification errors, and erroneous classification will exclude the true structure from those generated. Some new developments [8] were able to slightly improve the accuracy of MS classifiers.

In this approach we used a deterministic structure generator based on methods from combinatorics (orderly generation [9, 10]) and refined by techniques from group theory (fast isomorphism testing). Combination of these techniques results in a highly efficient algorithm. However, even such optimized structure generation algorithms can only compute an approximately constant number of isomers per unit time. Due to the combinatorial explosion of possible structures with increasing molecule size, exhaustive structure generation clearly has its limitations for higher molecular weights

An alternative approach are stochastic structure generators [11], that use spectral information during the structure generation process, in order in order to find the best path through chemical space. Stochastic structure generators based on NMR data seem to work well since chemical shifts are predicted quickly and accurately [12, 13, 14].

In contrast, it is difficult to predict mass spectra or even to decide whether a given MS corresponds to a given structure. For this reason no attempts were made to develop stochastic generators based on MS data. Not even the problem of comparing and ranking structure candidates has yet been examined intensively. In this paper we focus on that particular step, which is enclosed by the dark grey rectangle in Figure 1.

MS basically yields information on the masses of ions occurring in the mass spectrometer. Key to structure elucidation via EI–MS is the fact that there is a large set of fragment ions produced in the mass spectrometer's ionization chamber. Therefore an EI–MS measures a compound's fragment mixture rather than the compound itself, and this is why the mass spectrum of a chemical structure is more difficult to predict than NMR or IR spectra.

Fortunately most fragmentation reactions in an EI–MS follow certain well–known reaction schemes [15], and using these reaction schemes it is possible to generate a set of virtual fragments that will probably appear in an EI–MS.

Concentrations of fragment ions, i.e. peak intensities, depend on reaction dynamics, which are poorly understood due to the extreme

conditions in a mass spectrometer. Therefore prediction of peak intensities, while highly desirable for the structural information contained therein, is out of reach at present.

However, peak positions already allow to exclude unfavorable candidate structures automatically, and to calculate a ranking for a set of candidate structures.
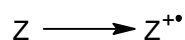
2. Methods

2.1. **Exhaustive Structure Generation.** In order to supply a well–defined set of candidate structures, we used the structure generator MOLGEN [16, 17]. MOLGEN is able to construct constitutional isomers that belong to a given molecular formula. The generation is exhaustive, nonredundant, and efficient. Several thousands of isomers can be generated per second.
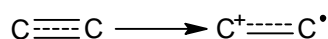
Example 1. The upper part of Figure 2 shows the experimental spectrum of methyl pentanoate $C_6H_{12}O_2$ together with its structural formula. There exist altogether 1313 constitutional isomers of $C_6H_{12}O_2$. These will serve as candidate set for our introductory example. They are generated by MOLGEN 3.5 in less than 0.1 s on a Pentium IV 1.6 GHz CPU.

2.2. **Virtual Fragmentation.** Generation of MS fragments can be divided into two parts. In a first step ions are formed from the uncharged candidate structure. In this paper we allow three types of ionization reactions:
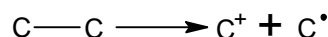
- n–ionization (n–I)

$$Z \longrightarrow Z^{+\bullet}$$

- π–ionization (π–I)

$$C ⩵ C \longrightarrow C^{+} ⩵ C^{\bullet}$$

- σ–ionization (σ–I)

$$C \longrightarrow C \longrightarrow C^{+} + C^{\bullet}$$

Here the following symbols describe generic atoms:
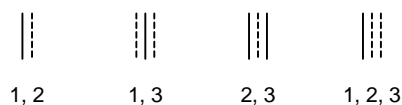
A: any atom
Y: heavy atom (i.e. any element except H)
Z: any atom bearing a free electron pair (N, O, P, S, halogens)

Alternatives for bond multiplicities are coded graphically as follows:

1, 2	1, 3	2, 3	1, 2, 3

After the initial ionization several secondary reactions are executed recursively. These can be either cleavages or rearrangements:
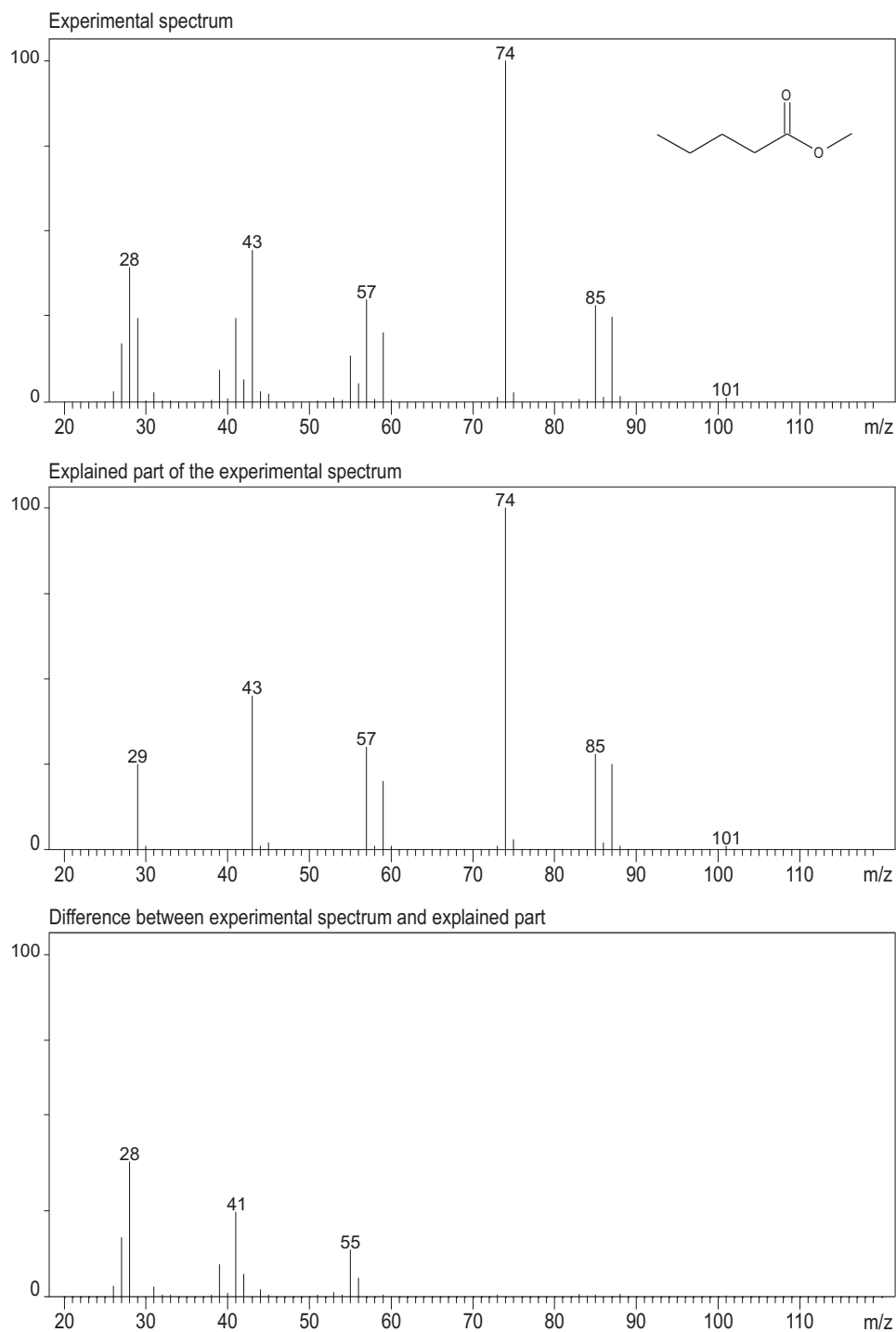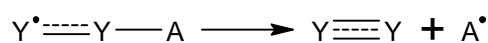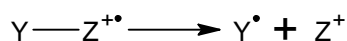
FIGURE 2. Experimental mass spectrum of methyl pentanoate (top), and the parts of the spectrum explained (middle) and unexplained (bottom) by the reactions considered
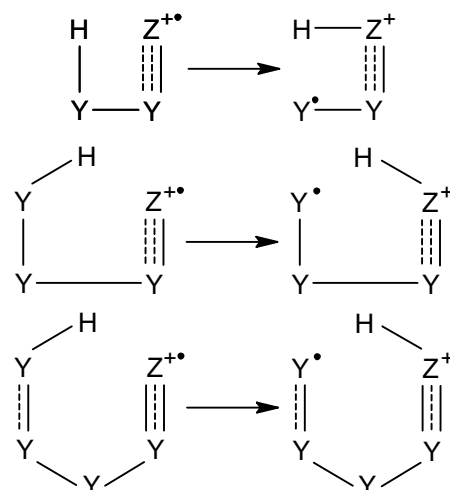
- α–cleavage (α–Cl)

$$Y^{\bullet}\!=\!=\!=Y\!-\!A \longrightarrow Y\!=\!=\!=Y \;+\; A^{\bullet}$$

- σ–cleavage (σ–Cl)

$$Y\!-\!Z^{+\bullet} \longrightarrow Y^{\bullet} \;+\; Z^{+}$$

- H–rearrangements on 4, 5 and 6 atoms (H-R4, H-R5, H-R6)

After each reaction step uncharged fragments are removed. Atoms in ions are labeled canonically [18]. Only ions occurring for the first time in the fragmentation process are considered for further recursive fragmentation. A more detailed description of in silico reactions and the construction of reaction networks is given in [19].

Of course several further reactions can occur in an MS. On the other hand, some of the above generalized reaction schemes may allow specific reactions that are not actually observed in a mass spectrometer. However, this minimalistic set of reaction schemes (extracted partly from [20]) is able to explain several peaks, as seen in the example of methyl pentanoate.

Example 2. Figure 3 shows the MS reaction network for methyl pentanoate obtained by the above reactions schemes. Each square represents an ion; numbers refer to structures in Figure 4. Arrows represent ionization and fragmentation reactions. Labels attached to the arrows denote the reaction scheme applied. Unlabeled arrows represent α–cleavages. π–Ionizations and σ–cleavages do not occur in this example.

Figure 4 lists all 32 ions that are generated from methyl pentanoate by the above reaction schemes. There are 16 different molecular formulae and 15 different integer masses occurring in the set of ions. Structures are ordered by decreasing mass. A structure's mass is given in the center of its header together with the molecular formula (left) and the number referred to in Figure 3.
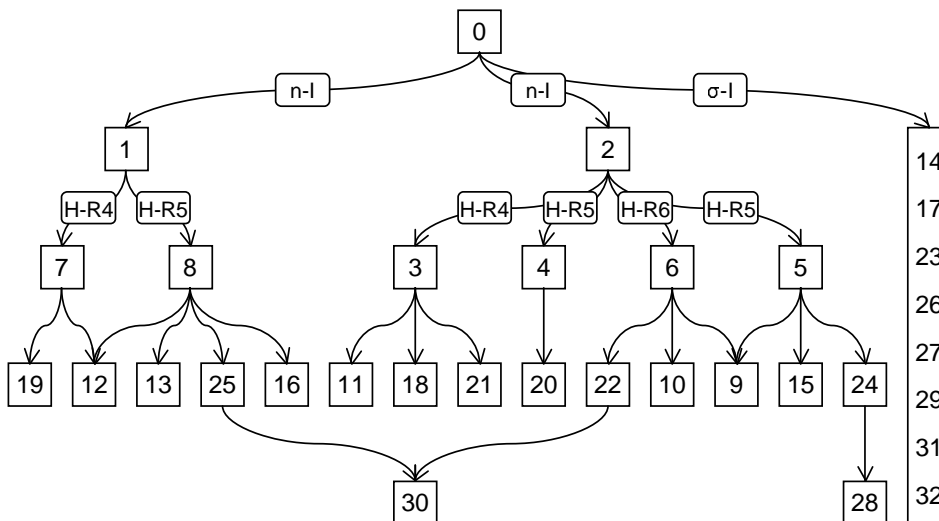
FIGURE 3. MS reactions of methyl pentanoate

However, the experimental spectrum is not completely explained by these fragments. For instance peaks at m/z values 28, 41 and 55 remain unexplained (cf. section 4). Comparison of the fragments obtained by corresponding reactions from competing structure candidates (e.g. structures isomeric to methyl pentanoate) will be discussed in subsection 2.3.

2.3. **Matchvalue Calculation.** As already mentioned, we are not able to calculate intensities for mass spectra. Masses of virtual fragments, however, can be compared to m/z values in an experimental spectrum. Isotopic peak ratios also will be taken into account.

Ideally a spectrum–structure compatibility matchvalue MV should fulfill the following requirements:

(R1) For any spectrum I and any structure S the matchvalue should be between 0 and 1: $\mathrm{MV}(I,S) \in [0,1]$.

(R2) For the correct structure S^T the matchvalue should be exactly 1: $\mathrm{MV}(I,S^T) = 1$.

(R3) For any wrong structure S^F the matchvalue should be less than for the correct structure: $\mathrm{MV}(I,S^F) < \mathrm{MV}(I,S^T)$.

If we had a matchvalue that fulfills the above conditions, the CASE problem would be solved. But of course we have not. In the following we derive a spectrum–structure compatibility matchvalue that at least approximates these requirements. For this purpose some mathematical definitions are useful.

Definition 1. A low resolution mass spectrum I is a mapping

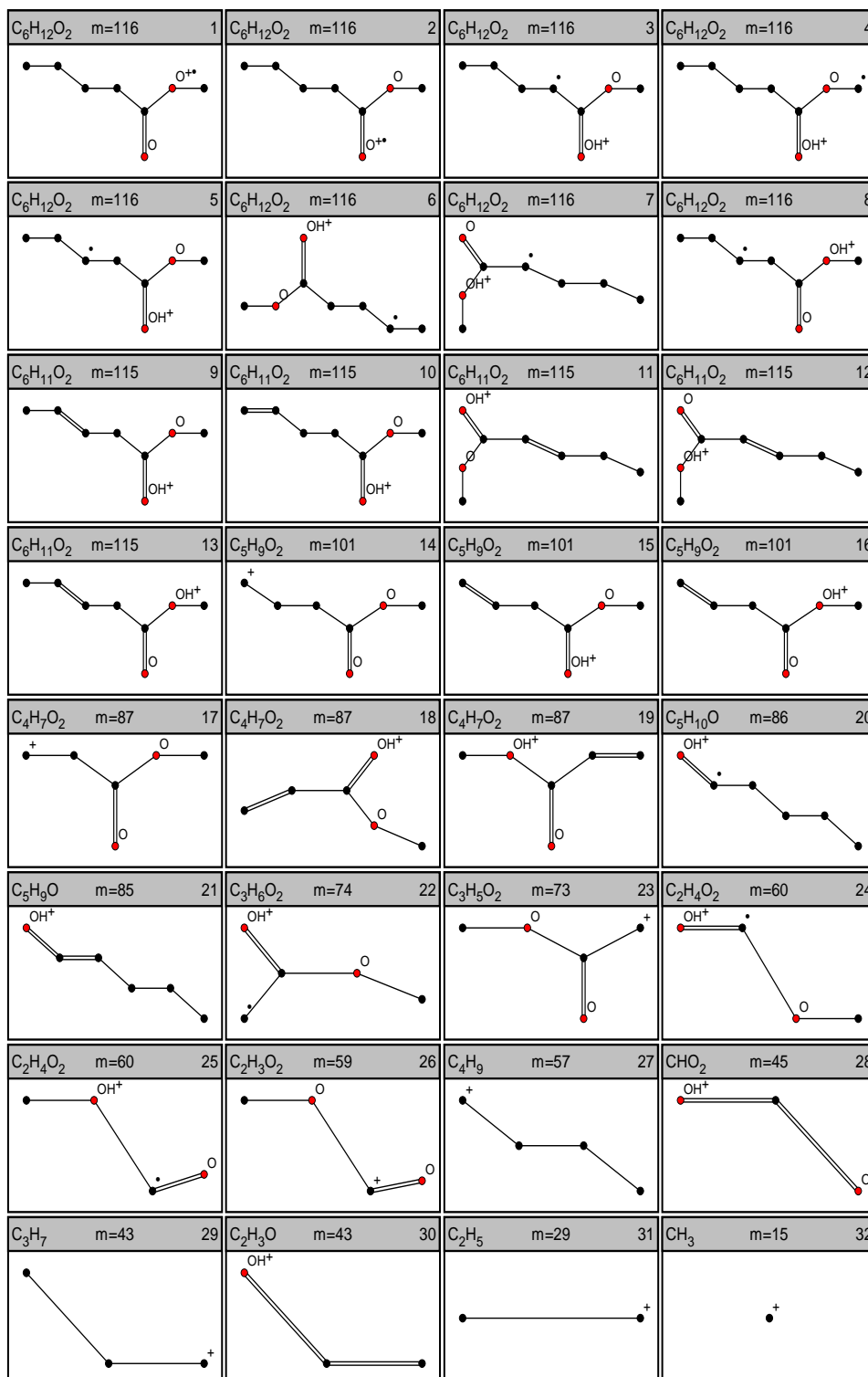$$I : \mathbb{N} \longrightarrow \mathbb{R}_+^0, \qquad m \longmapsto I(m)$$

FIGURE 4. Ions generated from methyl pentanoate

X	$\check{m}_X$	$\hat{m}_X$	$I_X(\check{m}_X)$	$I_X(\check{m}_X+1)$	$I_X(\check{m}_X+2)$
H	1	1	1	0	0
C	12	13	0.989	0.011	0
N	14	15	0.9963	0.0037	0
O	16	18	0.9976	0.0004	0.0020
F	19	19	1	0	0
Si	28	30	0.9223	0.0467	0.0310
P	31	31	1	0	0
S	32	34	0.9504	0.0075	0.0421
Cl	35	37	0.7577	0	0.2423
Br	79	81	0.5069	0	0.4931
I	127	127	1	0	0

TABLE 1. Natural isotope distributions for the elements of $\mathcal{E}$

from the set of natural numbers onto the set of non–negative real numbers. This mapping relates each integer m/z value m with its intensity $I(m)$. There exists a maximum m/z value $\hat{m}$ with $I(\hat{m}) > 0$:

$$\exists \hat{m} : I(\hat{m}) > 0 \wedge \forall m > \hat{m} : I(m) = 0.$$

Analogously a minimal m/z value $\check{m}$ with $I(\check{m}) > 0$ can be assigned. Furthermore a spectrum is typically normalized to a certain maximum intensity. Chemists prefer maximum intensity 100, but in order to simplify mathematical expressions we will claim that the spectrum is normalized to maximum intensity 1:

$$\exists \tilde{m} : I(\tilde{m}) = 1 \wedge \forall m \neq \tilde{m} : I(m) \leq 1.$$

$\tilde{m}$ is typically determined uniquely and called the spectrum's base mass.

In this manner we can describe experimental spectra as well as theoretical isotope distributions and calculated spectra. Every chemical element occurs with its natural isotope distribution. Our experiments will be limited to the 11 elements that are typical for organic chemistry:

$$\mathcal{E} = \{C, H, N, O, Si, P, S, F, Cl, Br, I\}.$$

Table 1 shows the natural isotope distributions I_X of the most common organic elements $X \in \mathcal{E}$ according to [21]. $\check{m}_X$ and $\hat{m}_X$ denote the minimal and maximal (integer) isotope mass of element X; $I_X(m)$ represents the relative natural abundance of isotope ^mX. For all masses $m \notin [\check{m}_X, \hat{m}_X]$ we have $I_X(m) = 0$. Furthermore let m_X denote the isotope mass of maximum abundance, called the monoisotopic mass of X.

Table 1 contains four elements X with $\check{m}_X = \hat{m}_X$. These monoisotopic elements are $\mathsf{H, F, P}$ and I. Hydrogen isotopes Deuterium $^2\mathsf{H}$ and Tritium $^3\mathsf{H}$ are left out for their extremely low abundance.

From the isotope distributions of elements we can compute isotope distributions of molecular formulae.

Definition 2. A molecular formula β is a mapping

$$\beta : \mathcal{E} \longrightarrow \mathbb{N}, \qquad X \longmapsto \beta(X)$$

from the set of chemical elements onto the set of natural numbers. This mapping relates each chemical element X to its multiplicity $\beta(X)$.

Isotope distributions of molecular formulae can be calculated by convolution of element isotope distributions. The convolution of two isotope distributions I_1 and I_2 is defined as

$$(2.3.1) \qquad (I_1 \cdot I_2)(m) := \sum_{i=0}^{m} I_1(i) I_2(m - i).$$

In mathematical terms, the convolution is an associative operation within the set of isotope distributions (for a proof see e.g. [22], pp 184–185). Using definition 2.3.1, the isotope distribution I_β of a molecular formula β can be expressed as

$$I_\beta = \prod_{X \in \mathcal{E}} I_X{}^{\beta(X)}.$$

Analogously to element isotope distributions we denote the minimal isotopomer mass of β by $\check{m}_\beta$ and the maximal isotopomer mass by $\hat{m}_\beta$, respectively. It is obvious that

$$\check{m}_\beta = \sum_{X \in \mathcal{E}} \check{m}_X \beta(X) \ \text{ and } \ \hat{m}_\beta = \sum_{X \in \mathcal{E}} \hat{m}_X \beta(X).$$

The monoisotopic mass of a molecular formula is defined as weighted sum of the monoisotopic masses of its elements:

$$m_\beta = \sum_{X \in \mathcal{E}} m_X \beta(X).$$

The monoisotopic mass of a molecular formula is not necessarily equal to the base mass $\tilde{m}_\beta$ of the formula's isotope distribution, as demonstrated by the following example.

Example 3. Consider the simple example of bromine monochloride, i.e. molecular formula BrCl. We have $\check{m}_{\mathsf{BrCl}} = \check{m}_{\mathsf{Cl}} + \check{m}_{\mathsf{Br}} = 114$ and $\hat{m}_{\mathsf{BrCl}} = \hat{m}_{\mathsf{Cl}} + \hat{m}_{\mathsf{Br}} = 118$. The isotope distribution I_β of BrCl is

computed as follows:

$$
\begin{aligned}
I_{\mathsf{BrCl}}(114) &= I_{\mathsf{Cl}}(35)I_{\mathsf{Br}}(79) = 0.3841 \\
I_{\mathsf{BrCl}}(115) &= 0 \\
I_{\mathsf{BrCl}}(116) &= I_{\mathsf{Cl}}(35)I_{\mathsf{Br}}(81) + I_{\mathsf{Cl}}(37)I_{\mathsf{Br}}(79) = 0.4964 \\
I_{\mathsf{BrCl}}(117) &= 0 \\
I_{\mathsf{BrCl}}(118) &= I_{\mathsf{Cl}}(37)I_{\mathsf{Br}}(81) = 0.1195
\end{aligned}
$$

Wee see that the base mass $\tilde{m}_{\mathsf{BrCl}} = 116$, whereas the monoisotopic mass $m_{\mathsf{BrCl}} = 114$.

Note that most summands in equation 2.3.1 are equal to zero (omitted in the above example). The convolution is quite cheap an operation in terms of CPU time: Summands with at least one factor zero need not be computed and accumulated.

Now let $\beta_1, ..., \beta_n$ denote the different molecular formulae that were found among the ions generated by virtual fragmentation. Assuming

(A1) $\beta_1, ..., \beta_n$ enclose all real fragment ions' molecular formulae, and

(A2) the experimental spectrum I was recorded without any errors in measurement,

then I can be written as a linear combination of the isotope distributions $I_{\beta_1}, ..., I_{\beta_n}$:

$$
(2.3.2) \qquad I = \sum_{i=1}^{n} x_i I_{\beta_i}, \quad \mathbf{x} \geq 0,
$$

where the linear combination of isotope distributions is defined in the following natural way:

$$
\left(\sum_{i=1}^{n} x_i I_{\beta_i} \right)(m) = \sum_{i=1}^{n} x_i I_{\beta_i}(m)
$$

As already mentioned, it is not feasible to compute the concentrations x_i. The idea of the method presented here is to treat concentrations as unknowns in a quadratic optimization problem

$$
(2.3.3) \qquad \min_{\mathbf{x} \geq 0} \sum_m \left(I(m) - \sum_{i=1}^{n} x_i I_{\beta_i}(m) \right)^2.
$$

Due to equation 2.3.2 this term becomes 0 for the true structure, and it is at most $\sum_m (I(m))^2$. Accordingly, we define a matchvalue

$$
\mathrm{MV}(I, S) = 1 - \left(\sum_m (I(m))^2 \right)^{-1} \min_{\mathbf{x} \geq 0} \sum_m \left(I(m) - \sum_{i \in n} x_i I_{\beta_i}(m) \right)^2
$$

that fulfills requirement R1, and due to equation 2.3.2 requirement R2 holds. Whether requirement R3 will be fulfilled, however, depends on how much the virtual fragment ions of false structures differ from

β_i	m_{β_i}	x_i	β_i	m_{β_i}	x_i
C_2H_5	29	0.2515	$C_3H_5O_2$	73	0.0156
C_2H_3O	43	0.0000	$C_3H_6O_2$	74	1.0379
C_3H_7	43	0.4606	C_5H_9O	85	0.3008
CHO_2	45	0.0242	$C_5H_{10}O$	86	0.0000
C_4H_9	57	0.3134	$C_4H_7O_2$	87	0.2619
$C_2H_3O_2$	59	0.2093	$C_5H_9O_2$	101	0.0138
$C_2H_4O_2$	60	0.0013			

TABLE 2. Calculation of the matchvalue for methyl pentanoate and the experimental spectrum from Figure 2

those of the true structure. For instance a false structure may cause the same set of fragment ions as the true structure. Then of course also the matchvalues for the true and the false structure will be equal.

Furthermore assumptions A1 and A2 are typically not fulfilled. However they were useful for modeling our matchvalue. Even with some deviations from these assumptions good ranking results can be obtained, as we will see in the following example.

Example 4. Table 2 lists molecular formulas β_i of fragment ions produced by virtual fragmentation of methyl pentanoate together with their monoisotopic masses m_{β_i}. When comparing this list carefully with Figure 4 we see that several molecular formulae are missing: CH_3 ($m=15$), $C_6H_{11}O_2$ ($m=115$), $C_6H_{12}O_2$ ($m=116$). These need not be considered for the matchvalue calculation as their masses do not occur in the experimental MS.

Column x_i shows solutions for the unknowns in the optimization problem 2.3.3. The calculated matchvalue is $MV(I, S^T) = 0.84421$. We can use the calculated x_i in order to represent the explained amount of intensity of the experimental spectrum. In Figure 2, middle, we see the explained part $I' = \sum_i x_i I_{\beta_i}$ of the experimental spectrum, and the residual peaks are shown in Figure 2, bottom.

2.4. **Candidate Ranking.** Next we examine whether our matchvalue is useful to distinguish the true structure from false candidate structures with the same molecular formula. For that purpose we calculate matchvalues for all isomers and sort them in descending order.

Example 5. For each of the 1313 isomers $C_6H_{12}O_2$ we obtain between 7 and 162 ions represented by 3 to 26 molecular formulae. The minimal matchvalue calculated is 0.00009, the maximal matchvalue 0.93488.

Figure 5 shows the 24 isomers with highest matchvalues, arranged in decreasing order of MV. The true structure is located at position 16. The first 13 positions are occupied by cyclic structures. This is surprising, as the ratio between cyclic and acyclic structures among
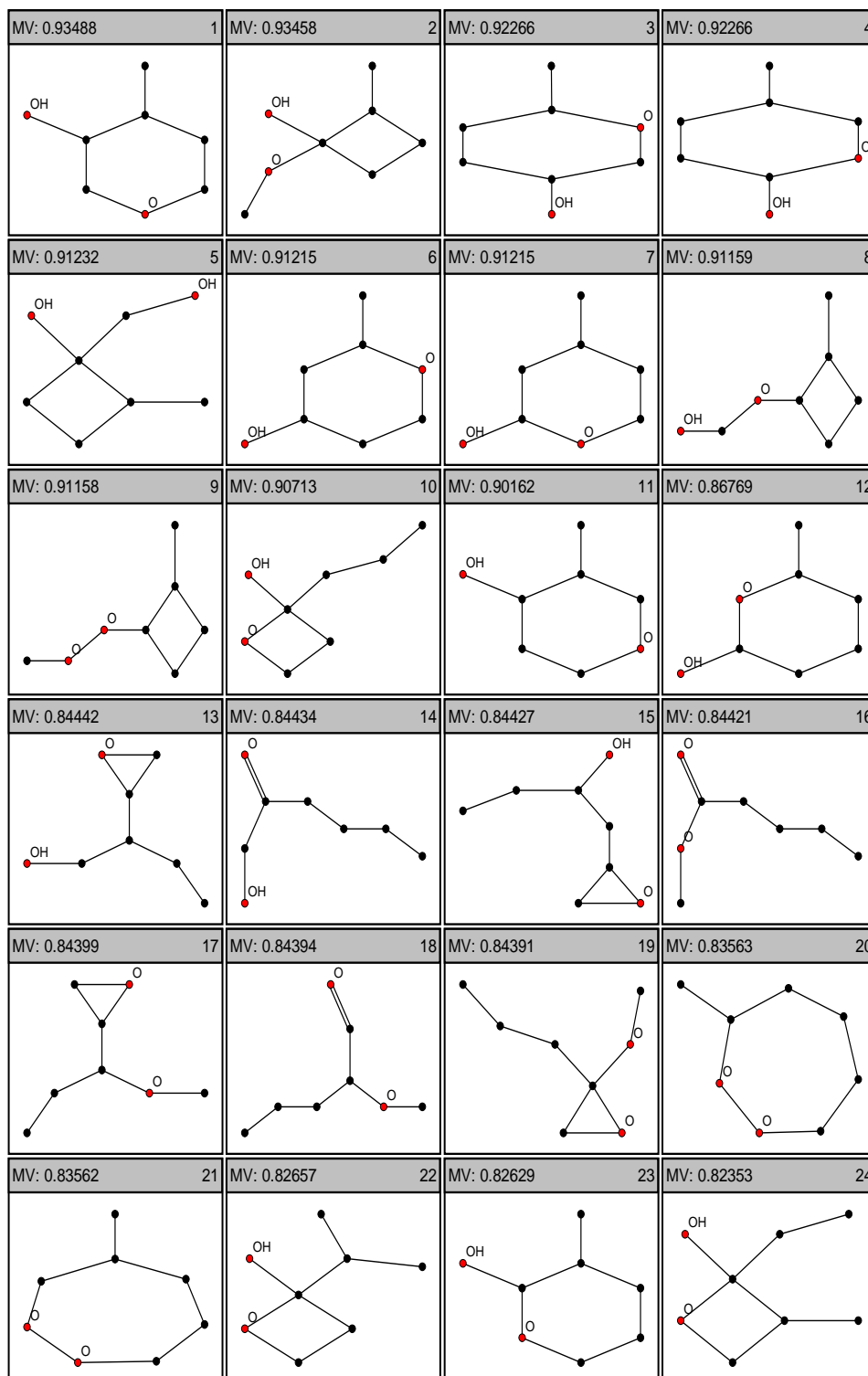
FIGURE 5. Ranking of $C_6H_{12}O_2$ isomers by compatibility with the experimental spectrum of methyl pentanoate

the $C_6H_{12}O_2$ isomers is close to 1 (641 acyclic, 672 cyclic structures). If there existed a possibility to distinguish cyclic and acyclic structures by means of the MS, the correct structure would advance to position 2.

Figures 6 and 7 show a histogram and a bar chart of the matchvalues. In this example the matchvalue seems to be well suited for excluding the major part of candidate structures. One could make a candidate selection according to the distribution of matchvalues and for instance refuse all candidates with matchvalues less than 0.5. The problem of candidate selection will be discussed in more detail in subsection 2.5.

In the histogram we clearly see a valley from matchvalue 0.4 to 0.55. Indeed there are no structures with matchvalues between 0.38423 and 0.55016. Structures on the right side of this valley produce a fragment ion of mass 74 and therefore are able to explain the experimental spectrum's base peak, while structures on the left have no fragment ion of that mass. Correspondingly, the bar chart exhibits a steep descent between structures 264 and 265. There are 264 structures with $MV \geq 0.55016$ and 1049 structures with $MV \leq 0.38423$.

In order to evaluate the quality of a ranking we can either use the absolute or the relative position of the true structure among structure candidates. We define the absolute ranking position (ARP) simply by the number of better candidates (BC, the number of candidates having higher MV than the true structure) plus 1.

When ranking samples of different numbers of candidates, it is more useful to consider a relative ranking position than the absolute ranking position. We want the relative ranking position to be a value between 0 and 1. Lower values should reflect better rankings. The relative ranking position should be 0 if the true structure is ranked first and 1 if the true structure is ranked last.

Let WC denote the number of worse candidates, i.e. candidates having lower MV than the true structure, and let TC be the (total) number of candidates. There are two possibilities to define a relative ranking position:

$$\text{RRP}_0 := \frac{\text{BC}}{\text{TC} - 1} \quad \text{and} \quad \text{RRP}_1 := 1 - \frac{\text{WC}}{\text{TC} - 1}.$$

Of course RRP_0 and RRP_1 are defined only if there exist at least two candidates. Both definitions fulfill the above requirements, but in the case of false candidates having the same MV as the true structure, RRP_0 and RRP_1 will differ. In order to take such situations into account, we finally define the relative ranking position as mean of RRP_0 and RRP_1:

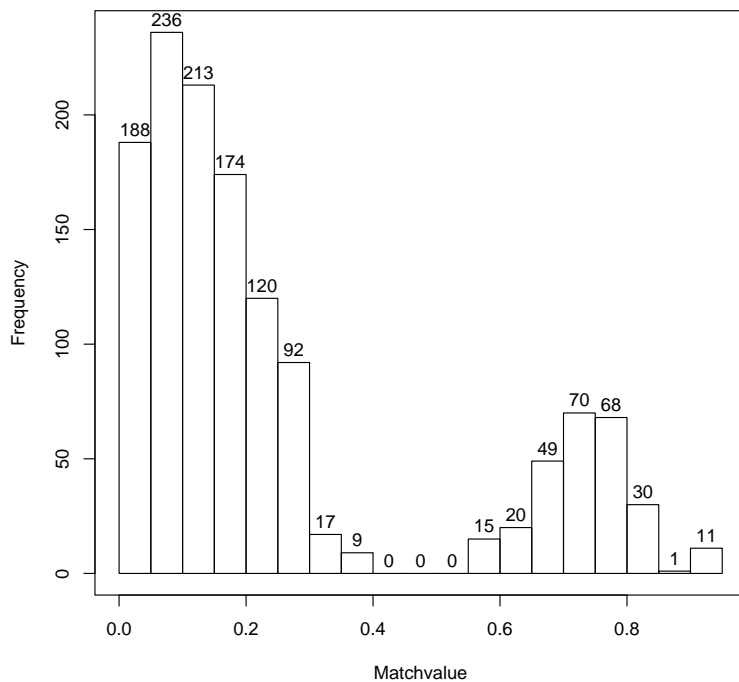$$\text{RRP} := \frac{1}{2} \left(1 + \frac{\text{BC} - \text{WC}}{\text{TC} - 1} \right).$$

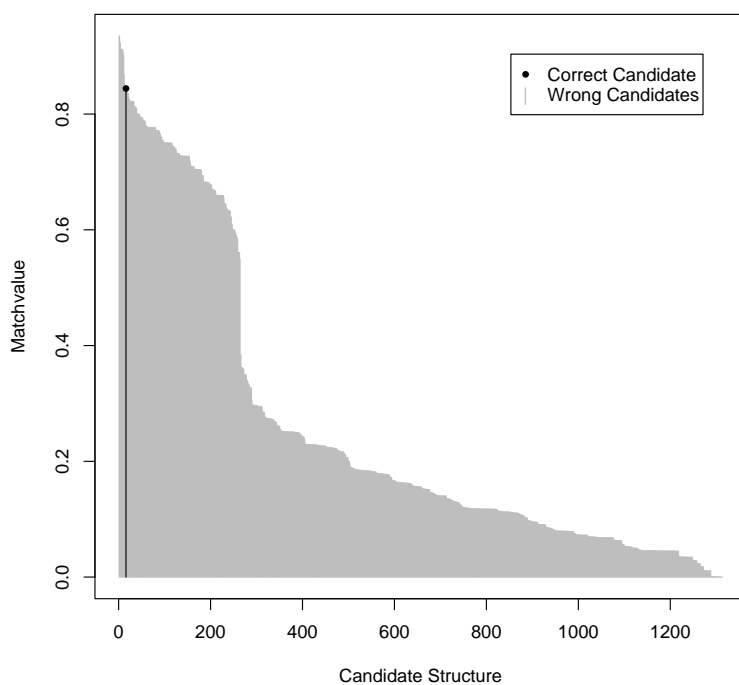FIGURE 6. Histogram of matchvalues for the constitutional isomers $C_6H_{12}O_2$

FIGURE 7. Bar chart of matchvalues for the constitutional isomers $C_6H_{12}O_2$

p	q_p	p	q_p	p	q_p
0.01	0.00912	0.10	0.14949	0.91	0.96549
0.02	0.02823	0.20	0.32390	0.92	0.96846
0.03	0.03613	0.30	0.46425	0.93	0.97382
0.04	0.05678	0.40	0.56902	0.94	0.97638
0.05	0.06846	0.50	0.68699	0.95	0.98216
0.06	0.09068	0.60	0.78238	0.96	0.98667
0.07	0.10605	0.70	0.85278	0.97	0.98950
0.08	0.11938	0.80	0.91589	0.98	0.99198
0.09	0.13128	0.90	0.96290	0.99	0.99547

TABLE 3. Quantiles q_p at several probabilities p for the matchvalues of the random sample of 1000 mass spectra

For instance, if all candidates have the same MV, then $RRP_0 = 1$, $RRP_1 = 1$, and $RRP = 0.5$.

Example 6. For our example methyl pentanoate, ranking by MV as described results in $RRP = 0.0114$, which appears to be quite good. Since the matchvalue of the true structure is unique, RRP is equal to RRP_0 and RRP_1.

2.5. **Candidate Selection.** A possibility for candidate selection by their matchvalues is based on simple statistics. To gather experience on the behavior of matchvalues from spectrum/structure pairs, we take a random sample of $n = 1000$ such pairs from the NIST MS library [1] and compute their matchvalues (i.e. for each spectrum the matchvalue of the true structure).

Figures 8 and 9 show a histogram and a bar chart of these matchvalues. As expected, matchvalues of the true structures tend to be rather high. More than 30% of the matchvalues are above 0.85. The mean is 0.62189, the median 0.68699. Unfortunately, there are low matchvalues also, which might be due to the insufficient set of reactions taken into consideration.

Next we calculate quantiles of these 1000 matchvalues. A p–quantile, $0 < p < 1$, is a number q_p where $p \cdot 1000$ of the 1000 matchvalues are less or equal than q_p, and $(1 - p) \cdot 1000$ of the 1000 values are greater or equal than q_p. In Figure 9 the 0.1, 0.3, 0.5, 0.7 and 0.9–quantiles are indicated. Table 3 shows several calculated quantiles.

The quantiles can be used in the following way: If we want to make a selection of candidate structures that contains the true candidate with a certain reliability r, we would have to choose all candidates with matchvalues at least q_{1-r}. As long as we consider spectra within the above random sample, the correct candidate will be among the chosen
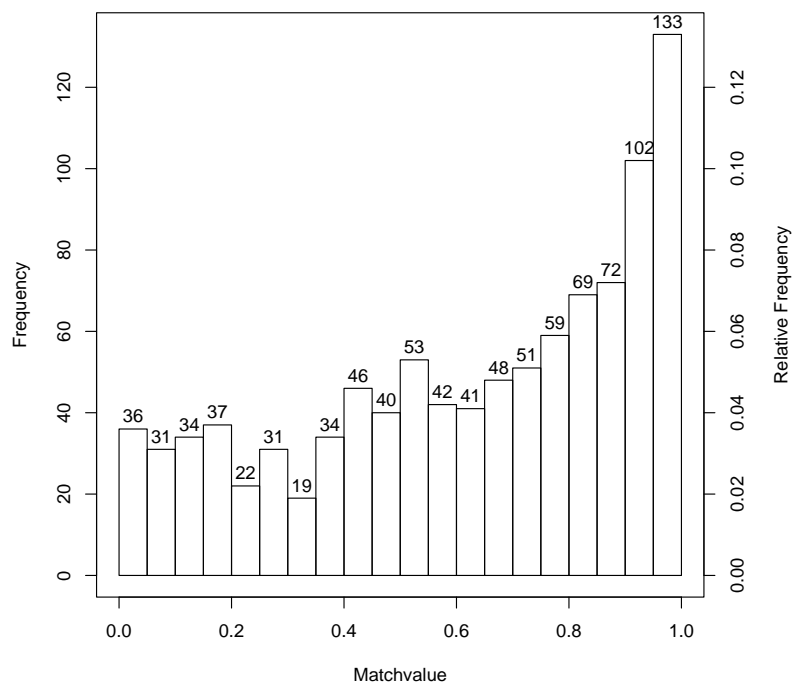
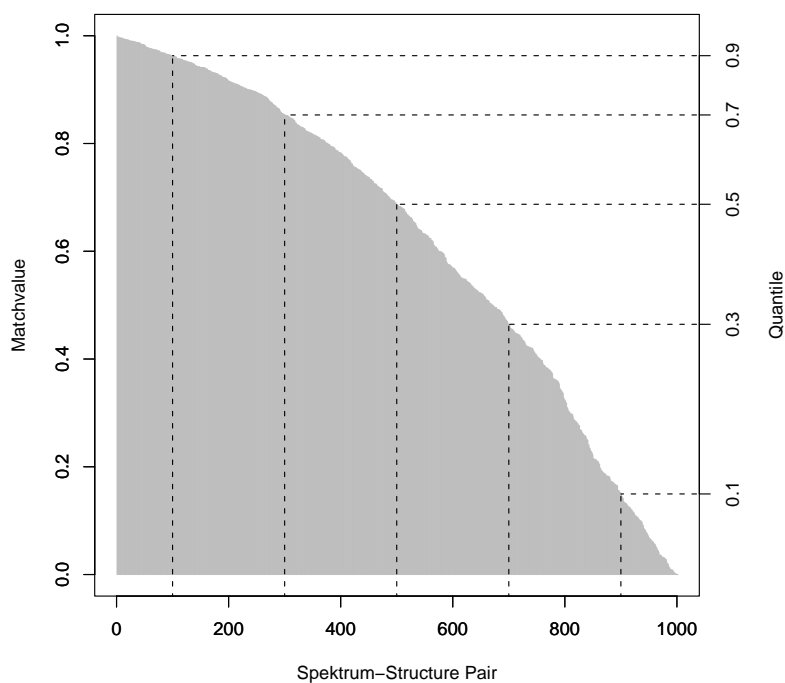FIGURE 8. Histogram of matchvalues of true structures for a random sample of 1000 mass spectra

FIGURE 9. Bar chart of matchvalues of true structures for a random sample of 1000 mass spectra

candidates with probability r. The large size of the random sample allows us to use these quantiles also for spectra outside the sample.

Example 7. We apply these statistics to the 1313 candidate structures for the spectrum of methyl pentanoate. If we want to have the correct structure within our selection with a reliability of 0.9, we have to select all isomers with matchvalues at least $q_{0.1} = 0.14949$. We would have to consider 676 structures. At a reliability of 0.5 the selection would comprise 184 structures, and the true candidate would still be included. Going down with the reliability decreases the size of the selection, but increases the risk of losing the correct candidate. If we choose reliability 0.3 there will remain only 12 candidates in the selection (those with matchvalue at least 0.85278), but the true candidate will be excluded.

The lowest reliability that still results in the true structure S^T to be selected is 0.32. This is based on the fact that $q_{0.68} = 0.83777 < \mathrm{MV}(I, S^T) < 0.84723 = q_{0.69}$.

3. Experimental

Obviously, the performance of MV in ranking structure candidates should be tested in a larger set of structure elucidation problems. Therefore we picked a random sample of 100 spectra from the NIST MS library. In order to keep computational costs moderate and to focus on standard organic chemistry we only chose spectrum–structure pairs which fulfilled the following restrictions:

- The molecular formula consists of elements from $\mathcal{E}$ exclusively.
- All atoms must have standard valencies, i.e. 1 for H and halogens, 2 for O, S, 3 for N, P and 4 for C, Si.
- Multi–component structures, isotopically labeled compounds, radicals and ions were excluded.
- The molecular mass of the structure is at most 200 amu.
- There exist more than 1 and at most 10000 constitutional isomers for the molecular formula.

As above, we generated for each spectrum–structure pair the set of constitutional isomers, performed for each isomer a virtual fragmentation and calculated the spectrum–structure compatibility matchvalues. We obtained 100 rankings and computed the relative ranking positions.

Table 4 shows the results of this experiment. The columns contain the following information:

Nr: An ID. In the Appendix for each ID a structure–descriptive chemical name is listed.

NIST: The spectrum's NIST–ID. This is useful for readers in order to reproduce the results.

β**:** The structure's molecular formula.

m**:** The structure's monoisotopic mass.

TC: The number of candidate structures, i.e. the number of constitutional isomers with the same molecular formula β.

MV: The matchvalue for the true structure.

BC: The number of false candidates with better matchvalue than the true structure.

EC: The number of false candidates with matchvalue equal to that of the true structure.

RRP: The relative ranking position.

C90: The number of candidates at reliability 0.9.

The total computation time was 13 h 30 min on a 1.6 GHz PC; the average number of candidates was 1839.12. Figure 10 shows a plot of absolute ranking positions vs. numbers of candidates. Of course no points are located above the diagonal. In 78 of the 100 cases the absolute ranking position is less or equal to half the number of candidates. These cases are represented by points lying on or below the broken line.

Figure 11 is a plot of relative ranking positions vs. number of candidates. There are 5 cases of RRP = 0 (Nr. 50, 74, 81, 85, 96), but also 1 case of RRP = 1 (Nr. 66). The average RRP is 0.2736 (standard deviation 0.2642), the median lies at 0.1806. Note that if we ranked candidates just by random, the expected average and median RRP would be 0.5. In 77 cases RRP is smaller than 0.5, represented by points below the solid line. In two cases (Nr. 10 and 13) all candidates share the same matchvalue, and accordingly RRP = 0.5. Figure 12 shows a histogram of the RRPs. We see that more than half of the cases have RRP $\leq$ 0.2.

Finally we applied the candidate selection as introduced in subsection 2.5. Figure 13 shows the results as a scatterplot. Each point represents one case in our random sample of 100 spectrum–structure pairs. The y–axis represents the absolute ranking position (of the true structure), the x–axis shows the number of selected candidates at reliability 0.9. Points above the diagonal represent cases where the true structure would be excluded from the candidate selection. There are 13 points above the diagonal (Nr. 10, 13, 15, 36, 42, 54, 60, 62, 64, 65, 76, 77 and 97), i.e. for 87% of the cases the true structure would be included in the selection.

Another important characteristic of this experiment is the ratio selected/total candidates. For reliability of 90% this quotient has a mean of 0.5973, i.e. on average more than 40% of all isomers are rejected at that reliability. However, values of this quotient much closer to 0 would be desirable.

Nr	NIST	β	m	TC	MV	BC	EC	RRP	C90
1	61627	C_9H_{16}	124	1902	0.97144	392	32	0.2146	1247
2	26708	$C_8H_{17}N$	127	2258	0.77435	1125	5	0.4996	1141
3	113790	$C_9H_{20}O$	144	405	0.33455	82	1	0.2042	243
4	158384	C_7H_{14}	98	56	0.45663	31	7	0.6273	50
5	38909	$C_{10}H_{18}$	138	5568	0.92117	684	0	0.1229	4236
6	61924	$C_{10}H_{20}$	140	852	0.19394	484	25	0.5834	575
7	60708	C_8H_{12}	108	2082	0.89620	318	0	0.1528	518
8	1911	$C_6H_{12}O_2$	116	1313	0.80581	16	0	0.0122	603
9	61640	$C_{13}H_{28}$	184	802	0.88881	0	208	0.1298	781
10	4617	CN_3F_5	149	11	0.00000	0	10	0.5000	0
11	194167	$C_4H_8N_2O$	100	6754	0.66949	172	0	0.0255	3149
12	186524	C_6H_9OBr	176	3703	0.30099	816	0	0.2204	1427
13	38120	CH_5SiBr	124	2	0.07170	0	1	0.5000	0
14	146109	$C_4H_2N_2FCl$	132	6393	0.76109	1160	0	0.1815	6393
15	73456	$C_5H_{11}Br$	150	8	0.11532	4	0	0.5714	3
16	61694	C_9H_{14}	122	7244	0.55448	1891	16	0.2622	6394
17	42198	$C_6H_{11}OBr$	178	1115	0.96765	27	0	0.0242	262
18	109982	$C_4H_7SiCl_3$	188	729	0.76491	16	20	0.0357	476
19	120	C_2H_3NO	57	26	0.26965	2	0	0.0800	4
20	154091	C_8H_{14}	110	654	0.51045	508	7	0.7833	654
21	71109	$C_6H_{14}N_2$	114	2338	0.91410	65	0	0.0278	1353
22	162833	$C_{10}H_{18}$	138	5568	0.85516	580	0	0.1042	5200
23	249757	C_5H_9N	83	313	0.51743	160	0	0.5128	313
24	3238	$C_5H_{10}O_2S$	134	4560	0.21210	794	1	0.1743	1473
25	113090	C_8H_{14}	110	654	0.91435	122	9	0.1937	361
26	63698	$C_3H_4N_2O$	84	1371	0.36161	191	0	0.1394	1371
27	74975	$C_6H_{12}O_3$	132	6171	0.79195	820	3	0.1331	3063
28	185578	$C_5H_{10}O_4$	134	5841	0.97237	875	0	0.1498	1721
29	61113	$C_{10}H_{20}$	140	852	0.97943	45	3	0.0546	805
30	160559	$C_4H_{13}NP_2$	137	396	0.24629	151	0	0.3823	185
31	46389	$C_5H_{10}O_3$	118	1656	0.96950	80	0	0.0483	824
32	46612	$C_9H_{18}O$	142	4745	0.94694	223	0	0.0470	3396
33	105465	$C_7H_{16}Si$	128	889	0.96954	1	3	0.0028	594
34	61433	$C_{11}H_{24}$	156	159	0.80741	97	14	0.6582	122
35	113438	C_8H_{16}	112	139	0.26305	96	0	0.6957	126
36	215368	$C_6H_{10}O$	98	747	0.12264	654	2	0.8780	613
37	20664	C_9H_{20}	128	35	0.80888	15	3	0.4853	26
38	62859	C_8H_{14}	110	654	0.68888	106	2	0.1639	536
39	69684	$C_{11}H_{24}O$	172	2426	0.73615	21	1	0.0089	1353
40	629	$C_5H_{13}N$	87	17	0.97332	1	0	0.0625	4
41	152851	$C_4H_7O_2Cl$	122	487	0.38246	6	0	0.0123	225
42	114082	$C_6H_{14}O$	102	32	0.10306	17	1	0.5645	16
43	196609	$C_5H_{11}NO_2$	117	6418	0.78537	1372	0	0.2138	1853
44	204405	C_9H_{14}	122	7244	0.83933	2327	10	0.3220	4708
45	28546	$C_5H_{12}O_2$	104	69	0.45592	1	0	0.0147	28
46	113901	C_9H_{16}	124	1902	0.69541	362	4	0.1915	1799
47	193841	$C_6H_{16}OSi$	132	425	0.99558	101	0	0.2382	102
48	604	$C_4H_6O_2$	86	263	0.73741	15	0	0.0573	263
49	73972	$C_9H_{21}NO$	159	7769	0.99527	316	6	0.0411	1939
50	63639	$C_2H_6O_2$	62	5	0.87246	0	0	0.0000	1

TABLE 4. Random selection of 100 spectrum–structure pairs

| Nr	NIST	β	m	TC	MV	BC	EC	RRP	C90
51	135135	C_4H_8NOCl	121	1371	0.62012	25	0	0.0182	499
52	63008	C_5H_6	66	40	0.71431	23	1	0.6026	40
53	61471	$C_{13}H_{28}$	184	802	0.87646	209	133	0.3439	800
54	60569	$C_8H_{17}Cl$	148	89	0.11500	13	2	0.1591	0
55	41785	$C_8H_{16}O$	128	1684	0.93602	115	9	0.0710	862
56	66064	C_9H_{14}	122	7244	0.37132	4759	7	0.6575	6977
57	160476	$C_6H_{10}O$	98	747	0.98744	129	3	0.1749	286
58	73870	C_8H_{12}	108	2082	0.69297	667	24	0.3263	2082
59	108516	$C_4H_{12}N_2$	88	38	0.93928	4	1	0.1216	12
60	4169	$C_3H_3Cl_3$	144	8	0.00389	4	0	0.5714	0
61	46224	$C_5H_{13}N$	87	17	0.76497	12	2	0.8125	17
62	158830	C_7H_9Br	172	2732	0.01160	1682	2	0.6163	593
63	61715	C_8H_{14}	110	654	0.77029	172	3	0.2657	651
64	1123	$C_4H_4O_3$	100	1073	0.13159	252	2	0.2360	186
65	156613	$C_9H_{22}NP$	175	9663	0.00081	7546	1	0.7810	2386
66	176	C_2H_7P	62	2	0.29376	1	0	1.0000	2
67	114550	$C_7H_{14}O$	114	596	0.80029	71	2	0.1210	185
68	214253	$C_5H_{13}NO$	103	149	0.87563	7	1	0.0507	33
69	70751	$C_7H_{19}N_3$	145	4238	0.84251	328	1	0.0775	1623
70	62909	$C_6H_{12}O$	100	211	0.72500	66	0	0.3143	150
71	37206	$C_7H_{13}N$	111	3809	0.58466	1271	0	0.3338	3189
72	229049	$C_4H_{11}NO$	89	56	0.94641	13	0	0.2364	14
73	19272	C_6H_{10}	82	77	0.17119	71	0	0.9342	73
74	831	C_2NF_3	95	5	0.74769	0	0	0.0000	1
75	114407	C_7H_{12}	96	222	0.95768	5	0	0.0226	132
76	5393	$C_4H_6O_2Cl_2$	156	1131	0.05743	203	0	0.1796	135
77	30409	$C_5H_{18}Si_3$	162	521	0.00000	498	22	0.9788	479
78	60785	$C_9H_{20}O$	144	405	0.72746	95	5	0.2413	387
79	72642	$C_9H_{22}N_2$	158	4994	0.93936	1614	382	0.3615	1997
80	118272	C_3H_7NO	73	84	0.85375	4	0	0.0482	84
81	108346	$C_3H_7O_2Br$	154	38	0.18857	0	0	0.0000	8
82	26687	C_8H_{14}	110	654	0.53326	547	2	0.8392	654
83	113772	$C_7H_{14}O$	114	596	0.28305	402	3	0.6782	456
84	1614	C_8H_{16}	112	139	0.85901	5	0	0.0362	131
85	107506	$C_9H_{19}F$	146	211	0.50982	0	0	0.0000	147
86	98625	$C_6H_{14}Si$	114	314	0.93385	26	0	0.0831	29
87	1908	$C_6H_{12}O_2$	116	1313	0.41749	515	0	0.3925	809
88	134724	C_3H_4NSBr	165	480	0.26994	111	12	0.2443	480
89	50930	C_9H_{18}	126	338	0.61212	57	7	0.1795	308
90	64555	$C_5H_{10}N_2$	98	2668	0.84749	416	0	0.1560	2521
91	113750	$C_9H_{20}O$	144	405	0.23624	22	10	0.0668	242
92	114530	$C_8H_{16}O$	128	1684	0.37670	143	0	0.0850	1092
93	61453	$C_{12}H_{24}$	168	5513	0.31383	978	2	0.1776	3085
94	37233	C_9H_{16}	124	1902	0.31667	582	0	0.3062	1402
95	60877	$C_{12}H_{24}$	168	5513	0.94596	411	0	0.0746	1695
96	63617	C_3H_4O	56	13	0.88094	0	0	0.0000	12
97	72945	C_4H_5OCl	104	175	0.05026	32	0	0.1839	0
98	113601	$C_{12}H_{24}$	168	5513	0.87997	70	0	0.0127	4439
99	52322	$C_5H_{13}N_3$	115	4054	0.28507	1154	0	0.2847	3107
100	215367	C_6H_8O	96	1623	0.53769	955	21	0.5953	1623

TABLE 4, continued
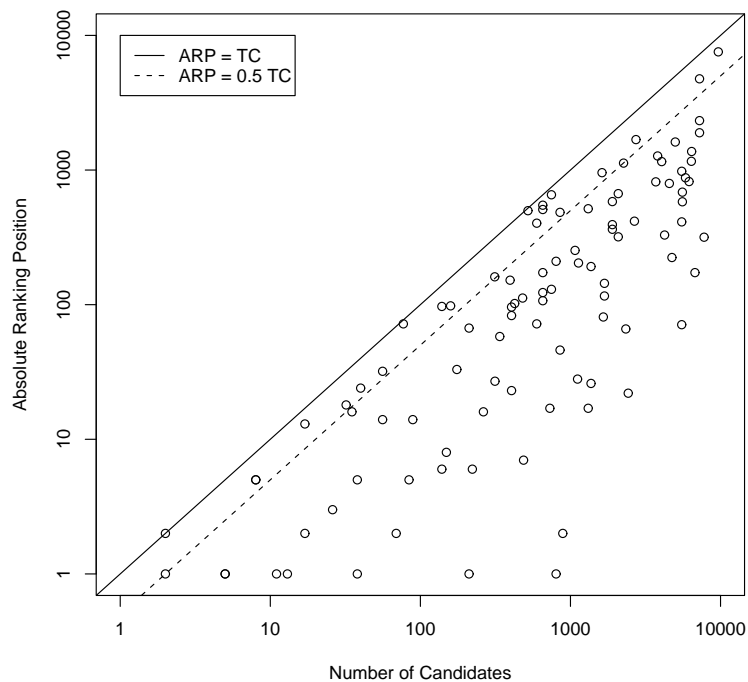
FIGURE 10. Absolute ranking positions and numbers of candidates for a random sample of 100 mass spectra
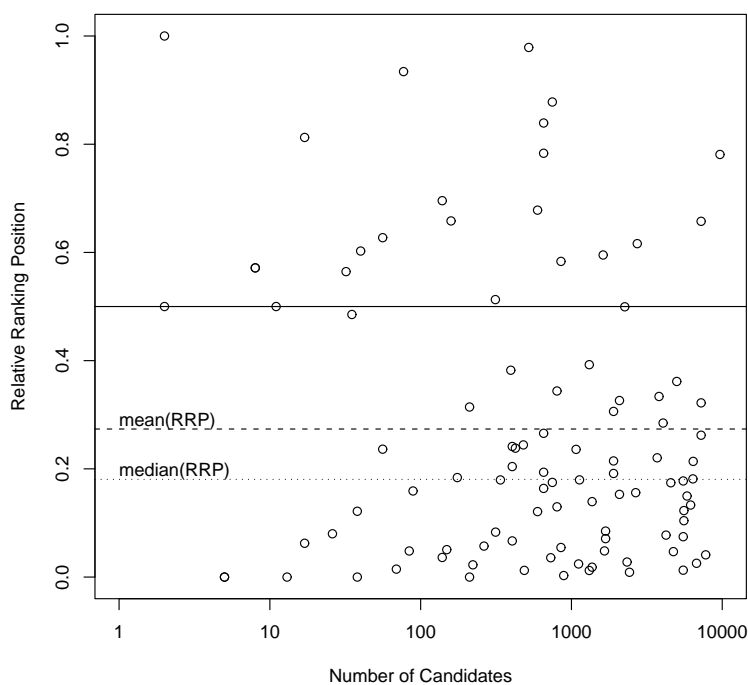

FIGURE 11. Relative ranking positions and numbers of candidates for a random sample of 100 mass spectra
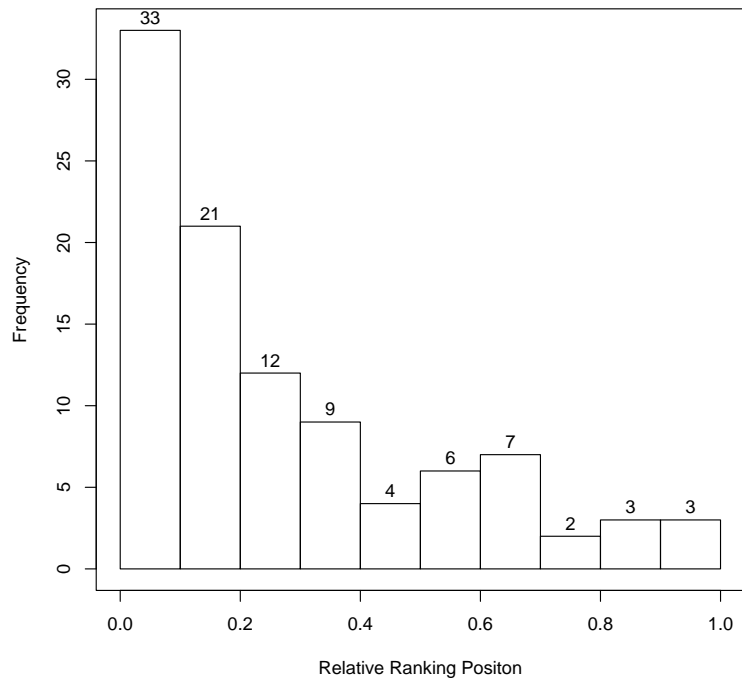
FIGURE 12. Histogram of relative ranking positions for a random sample of 100 mass spectra
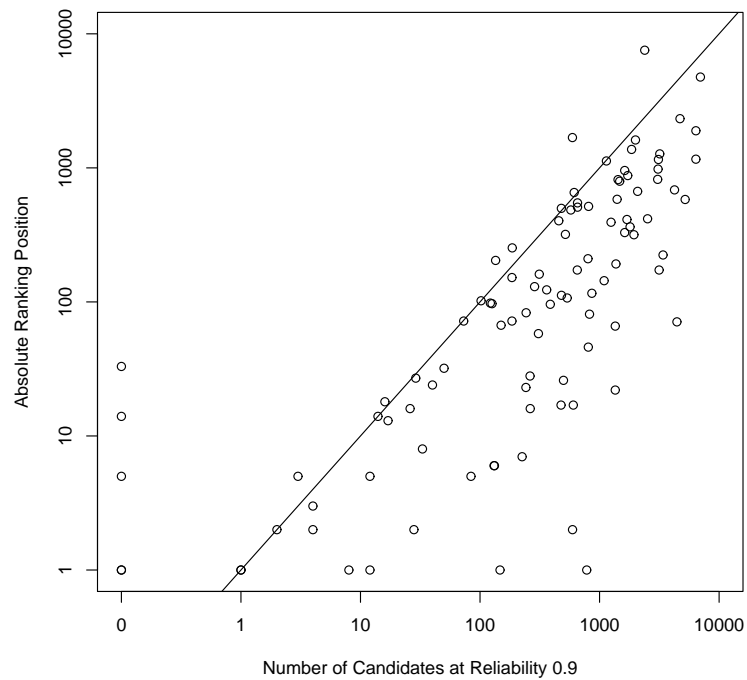

FIGURE 13. Absolute ranking positions and numbers of selected candidates at reliability 0.9

4. Discussion

Although the sample data was limited to small molecules and small candidate spaces, the results obtained are not yet sufficient for automated structure elucidation. It seems, however, worthwhile to develop the approach, given the continuously improving analytical and IT methods.

When revising subsection 2.2 we found that most unexplained peaks in Figure 2 can be explained by inductive cleavage reactions and loss of hydrogen. Thereby we obtain fragment ions of m/z 27, 28, 41, 42, 55 and 56. After formulating reaction schemes that realize these fragmentations and adding them to the catalogue of MS reaction schemes for virtual fragmentation, we obtained a far better result for this particular example, methyl pentanoate. For the true structure now a matchvalue of 0.99367 was obtained, and it is now ranked second (see Figure 14). Also, in this new ranking the matchvalues of the three leading structures differ clearly from the others. However, when applying these additional reaction schemes to the 100 randomly selected spectrum–structure pairs, no improvement in the average RRP was observed.

Several improvements are possible regarding subsections 2.2 and 2.3. There exist more sophisticated computer programs for virtual fragmentation [23, 24] that raise hope for better ranking results. First experiments with MassFrontier on very small sample sets resulted in a lower average RRP, but it has not yet been applied to the 100 sample spectra as described in section 3.

One should keep in mind that by adding further reaction schemes one will generally be able to explain more observed peaks. This, however, as seen above, will not necessarily lead to improved ranking results, as wrong structure candidates also will enjoy higher matchvalues.

Even for the matchvalue calculation alternatives have to be tested. Solving the optimization problem 2.3.3 is extremely time consuming if large sets of theoretical isotope distributions and densely populated spectra are to be processed. Instead, fuzzy isotope distributions as in [25] promise similar results with far less computational effort. One should also think about methods that penalize predicted virtual fragments that do not appear in the experimental spectrum.

Of course progress in predicting intensities of fragments would be most important to CASE via MS. If we were able to compute intensities, we could simply compare virtual and measured spectra by algorithms known from MS library search programs, such as the normalized dot product. For early attempts to quantitatively model the reactions occurring in a mass spectrometer see references [26, 27]. Regrettably, these programs were never tested in the manner shown above. A recent approach [28, 29] is currently about to be evaluated with the above protocol.
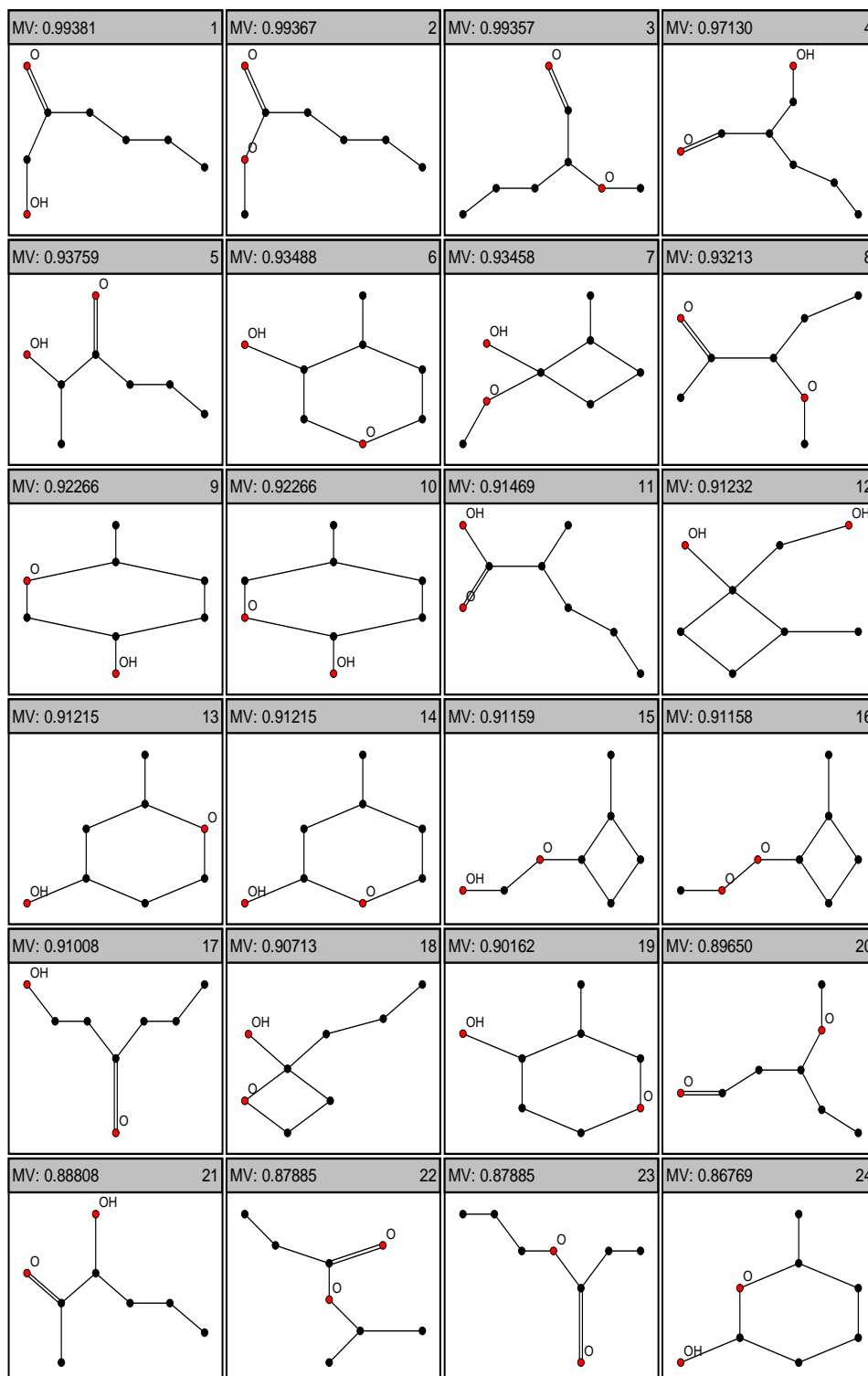
FIGURE 14. Ranking of $C_6H_{12}O_2$ isomers by compatibility with the experimental spectrum of methyl pentanoate using additional fragmentation schemes

For candidate selection one could think about more sophisticated methods. These should take into account the distribution of false candidate's matchvalues.

However, the methods described herein, especially for evaluating the quality of ranking procedures, could be important tools for future developments, and they are not restricted to mass spectrometry.

Finally, beyond mass spectrometry, one could also use retention time prediction in order to improve the ranking of candidate structures. Several studies on the prediction of GC retention times appeared in the past (e.g. [30]), and an application in combination with CASE via MS seems to be promising.

A possible scenario for the application of structure ranking by MS could be in the context of combinatorial chemistry. Then the set of candidate structures would not comprise all constitutional isomers, but only a small subset that lies inside the combinatorial library under investigation. In combination with more accurate high–resolution MS/MS techniques the approach described here could pave the way towards automated structure elucidation via mass spectrometry.

5. Appendix

1: 1,5-Heptadiene, 3,3-dimethyl-, (E)-; 2: Aziridine, 1-(1,1-dimethylethyl)-2,3-dimethyl-, trans-; 3: 4-Heptanol, 3-ethyl-; 4: 3-Methyl-2-hexene; 5: Cyclohexane, 1-methyl-3-(1-methylethylidene)-; 6: 3-Nonene, 3-methyl-, (E)-; 7: Cyclobutane, 1,2-diethenyl-, trans-; 8: Hexanoic acid; 9: Decane, 2,5,6-trimethyl-; 10: 3-Diaziridinamine, N,N,1,2,3-pentafluoro-; 11: Formic acid N'-ethylidene-N-methyl-hydrazide; 12: 2-Bromomethyl-3,4-dihydro-2H-pyran; 13: Silane, (bromomethyl)-; 14: 4-Chloro-6-fluoro-pyrimidine; 15: Butane, 1-bromo-2-methyl-, (.+/-.)-; 16: 3-Nonen-1-yne, (Z)-; 17: Cyclopentane, 1-bromo-2-methoxy-, trans-; 18: (2,2-Dichlorovinyl)dimethylchlorosilane; 19: Acetonitrile, hydroxy-; 20: 2,4-Hexadiene, 2,3-dimethyl-; 21: 1-Pyrrolidineethanamine; 22: Bicyclo[4.1.0]heptane, 3,7,7-trimethyl-, [1S-(1.alpha.,3.alpha.,6.alpha.)]-; 23: 1H-Pyrrole, 2,3-dihydro-1-methyl-; 24: Propanoic acid, 3-(ethylthio)-; 25: 6-Methyl-1,5-heptadiene; 26: Formamide, N-(cyanomethyl)-; 27: Butanoic acid, 2-hydroxy-3,3-dimethyl-; 28: Propanoic acid, 2-(methoxymethoxy)-; 29: Cyclohexane, 1-ethyl-2,4-dimethyl-; 30: Amine, bis(2-phosphinoethyl)-; 31: Butanoic acid, 4-methoxy-; 32: 2-Pentanone, 3,3,4,4-tetramethyl-; 33: trans-1,2-Dimethylsilacyclohexane; 34: Octane, 2,2,6-trimethyl-; 35: 1-Pentene, 3-ethyl-3-methyl-; 36: 3-Methylpenta-1,3-diene-5-ol, (E)-; 37: Hexane, 2,2,5-trimethyl-; 38: 1,1'-Bicyclopropyl, 1,1'-dimethyl-; 39: 2-Undecanol; 40: 1-Butanamine, 3-methyl-; 41: Propanoic acid, 3-chloro-, methyl ester; 42: 1-Butanol, 2-ethyl-; 43: Propanamide, 2-hydroxy-2,N-dimethyl-; 44: 3-Allylcyclohexene; 45: Hydroperoxide, pentyl; 46: 2-Nonyne; 47: tert-Butyldimethylsilanol; 48: 2-Propenoic acid, 2-methyl-; 49: N,N-Dimethyl-3-butoxypropylamine; 50: 1,2-Ethanediol; 51: N-(2-Chloroethyl)acetamide; 52: 3-Penten-1-yne, (E)-; 53: Decane, 5-propyl-; 54: Octane, 4-chloro-; 55: Cycloheptane, methoxy-; 56: Bicyclo[6.1.0]non-1-ene; 57:

4-Penten-2-one, 4-methyl-; 58: Octatriene, 1,3-trans-5-trans-; 59: 2-Methyl-1,2-propanediamine; 60: 1-Propene, 3,3,3-trichloro-; 61: Ethanamine, N-ethyl-N-methyl-; 62: Bicyclo[3.2.0]hept-2-ene, 4-bromo-; 63: 1,3-Hexadiene, 2,5-dimethyl-; 64: 2,4(3H,5H)-Furandione; 65: Dimethylamine, N-(diisopropylphosphino)methyl-; 66: Dimethylphosphine; 67: 2,4-Dimethyl-4-penten-2-ol; 68: 4-Amino-1-pentanol; 69: 1,3-Propanediamine, N-(3-aminopropyl)-N-methyl-; 70: 1-Penten-3-ol, 3-methyl-; 71: 8-Azabicyclo[3.2.1]octane; 72: N,N-Dimethylaminoethanol; 73: Cyclopentane, methylene-; 74: Acetonitrile, trifluoro-; 75: Cyclopentene, 1-ethyl-; 76: Butanoic acid, 2,3-dichloro-; 77: Silane, (silylmethyl)[(trimethylsilyl)methyl]-; 78: Pentane, 1-butoxy-; 79: N,N,N',N'-Tetramethyl-1,5-pentanediamine; 80: N-Ethylformamide; 81: 3-Bromo-1,2-propanediol; 82: 2,4-Hexadiene, 3,4-dimethyl-, (Z,Z)-; 83: 3-Methyl-3-hexen-2-ol; 84: Cyclopentane, 1-ethyl-3-methyl-, cis-; 85: 1-Fluorononane; 86: 1,1,3-Trimethyl-1-silacyclobutane; 87: Butanoic acid, 2-ethyl-; 88: 2-Bromo-ethyl isothiocyanate; 89: 1-Octene, 7-methyl-; 90: 1H-Pyrazole, 4,5-dihydro-4,5-dimethyl-; 91: 4-Octanol, 2-methyl-; 92: 2-Methyl-6-hepten-3-ol; 93: 1-Undecene, 2-methyl-; 94: Cycloheptane, 1-methyl-4-methylene-; 95: Cyclobutane, 1-hexyl-2,3-dimethyl-; 96: 2-Propyn-1-ol; 97: 2-Butenoyl chloride; 98: 1-Methyl-2-(4-methyl-pentyl)cyclopentane; 99: 1-Piperazinamine, 4-methyl-; 100: 2-Penten-4-yn-1-ol, 3-methyl-, (E)-.

References

[1] *NIST/EPA/NIH Mass Spectral Library, NIST '98 version.* U.S. Department of Commerce, National Institute of Standards and Technology. Gaithersburg, U.S.

[2] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *Molecules in Silico: Potential versus Known Organic Compounds.* MATCH Commun. Math. Comput. Chem., 54:301–312, 2005.

[3] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project.* McGraw–Hill Book Company, New York, St. Louis, San Francisco, 1980.

[4] W. Werther, H. Lohninger, F. Stancl, and K. Varmuza. *Classification of Mass Spectra: A Comparison of Yes/No Classification Methods for the Recognition of Simple Structural Properties.* Chemom. Intel. Lab. Syst., 22:63–76, 1994.

[5] K. Varmuza and W. Werther. *Mass Spectral Classifiers for Supporting Systematic Structure Elucidation.* J. Chem. Inf. Comput. Sci., 36:323–333, 1996.

[6] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *Molecules in Silico: The Generation of Structural Formulae and its Applications.* J. Comput. Chem. Jpn., 3:85–96, 2004.

[7] A. Kerber, R. Laue, M. Meringer, and K. Varmuza. *MOLGEN–MS: Evaluation of Low Resolution Electron Impact Mass Spectra with MS Classification and Exhaustive Structure Generation*, volume 15 of *Advances in Mass Spectrometry*, pages 939–940. Wiley, 2001.

[8] K. Varmuza, P. He, and K.-T. Fang. *Boosting Applied to Classification of Mass Spectral Data.* J. Data Sci., 1:391–404, 2003.

[9] R. C. Read. *Everyone a Winner.* Annals of Discrete Mathematics, 2:107–120, 1978.

[10] C. J. Colborn and R. C. Read. *Orderly Algorithms for Generating Restricted Classes of Graphs.* J. Graph Theory, 3:187–195, 1979.

[11] J.-L. Faulon. *Stochastic Generator of Chemical Structure. 3. 2. Using Simulated Annealing to Search the Space of Constitutional Isomers*. J. Chem. Inf. Comput. Sci., 36:731–740, 2001.

[12] J. Meiler and M. Will. *Automated Structure Elucidation of Organic Molecules from ^{13}C NMR Spectra using Genetic Algorithms and Neural Networks*. J. Chem. Inf. Comput. Sci., 41:1535–1546, 2001.

[13] J. Meiler and M. Köck. *Novel Methods of Automated Structure Elucidation Based on ^{13}C NMR Spectroscopy*. Magn. Reson. Chem., 42:1042–1045, 2004.

[14] C. Steinbeck. *SENECA: A Platform-Independent, Distributed, and Parallel System for Computer–Assisted Structure Elucidation in Organic Chemistry*. J. Chem. Inf. Comput. Sci., 41:1500–1507, 2001.

[15] F. W. McLafferty and F. Turecek. *Interpretation of Mass Spectra*. University Science Books, Mill Valley, California, 4. edition, 1993.

[16] C. Benecke, T. Grüner, A. Kerber, R. Laue, and T. Wieland. *Molecular Structure Generation with MOLGEN, new Features and Future Developments*. Fresenius J. Anal. Chem., 358:23–32, 1997.

[17] T. Grüner, A. Kerber, R. Laue, and M. Meringer. *MOLGEN 4.0*. MATCH Commun. Math. Comput. Chem., 37:205–208, 1998.

[18] J. Braun, R. Gugisch, A. Kerber, R. Laue, M. Meringer, and C. Rücker. *MOLGEN–CID, A Canonizer for Molecules and Graphs Accessible through the Internet*. J. Chem. Inf. Comput. Sci., 44:642–548, 2004.

[19] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *Molecules in Silico: A Graph Description of Chemical Reactions*. Adv. Quantum Chem. In press.

[20] W. Werther. *Versuch einer Systematik der Reaktionsmöglichkeiten in der Elektronenstoß–Massenspektrometrie (EI–MS)*. Unpublished, 1996.

[21] *Exact Masses and Isotopic Abundances of the Elements*. Mass Spectrometry and Chromatography — Scientific Instrument Services. Inc. www.sisweb.com/referenc/source/exactmaa.htm.

[22] M. Meringer. *Mathematical Models for Combinatorial Chemistry and Molecular Structure Elucidation*. Logos–Verlag Berlin, 2004. In German.

[23] *MassFrontier 4.0*. HighChem, Ltd. Bratislava, Slovakia.

[24] *ACD/MS Manager*. Advanced Chemistry Development, Inc. Toronto, Canada.

[25] B. Seebass and E. Pretsch. *Automated Compatibility Tests of the Molecular Formulas or Structures of Organic Compounds with their Mass Spectra*. J. Chem. Inf. Comput. Sci., 39:713–717, 1999.

[26] J. Gasteiger, W. Hanebeck, and K.-P. Schulz. *Prediction of Mass Spectra from Structural Information*. J. Chem. Inf. Comput. Sci., 32:264–271, 1992.

[27] J. Gasteiger, W. Hanebeck, K.-P. Schulz, S. Bauerschmidt, and R. Höllering. *Automatic Analysis and Simulation of Mass Spectra*. volume 4 of *Computer–Enhanced Analytical Spectroscopy*, pages 97–133. Kluwer Academic Publishers, 1993.

[28] H. Chen, B. Fan, M. Petitjean, A. Panaye, J. P. Doucet, H. Xia, and S. Yuan. *MASSIS: A Mass Spectrum Simulation System. 1. Principle and Method*. Eur. J. Mass Spectrom., 9:175–186, 2003.

[29] H. Chen, B. Fan, M. Petitjean, A. Panaye, J. P. Doucet, F. Li, H. Xia, and S. Yuan. *MASSIS: A Mass Spectrum Simulation System. 2: Procedures and Performance*. Eur. J. Mass Spectrom., 9:445–457, 2003.

[30] Z. Garkani-Nejad, M. Karlovits, W. Demuth, T. Stimpfl, W. Vycudilik, M. Jalali-Heravi, and K. Varmuza. *Prediction of Gas Chromatographic Retention Indices of a Diverse Set of Toxicologically Relevant Compounds*. J. Chromatogr. A, 1028:287–295, 2004.