Anal. Chem. **2009,** *81,* 3608–3617

Matching Structures to Mass Spectra Using Fragmentation Patterns: Are the Results As Good As They Look?

Emma L. Schymanski,*,† Markus Meringer,‡ and Werner Brack†

UFZ, Helmholtz Centre for Environmental Research-UFZ, Department of Effect-Directed Analysis, Permoserstrasse 15, D-04318 Leipzig, Germany, and DLR, German Aerospace Centre-DLR, Remote Sensing Technology Institute, Münchener Strasse 20, D-82234 Oberpfaffenhofen-Wessling, Germany

Three programs were assessed for their ability to predict mass spectral fragmentation patterns for all constitutional isomers of an experimental low-resolution electron impact mass spectrum (EI-MS), given the molecular formula, and use this information to identify the "correct structure". MOLGEN 3.5 was used to generate the structures, while all spectra were extracted from the NIST database. The commercial programs Mass Frontier and ACD MS Manager, as well as MOLGEN-MSF (developed by the University of Bayreuth) were used to generate mass spectral fragments. MOLGEN-MSF was used to generate "match values" to compare the different programs and their ability to identify the "correct structure". Although high match values could be achieved with certain settings, the ranking of the correct structure relative to other constitutional isomers was not significantly better than the results published previously and in some cases significantly worse. Furthermore, all programs showed bias toward specific structures, which changed significantly with minor changes to the program settings. Thus, advances in mass spectral fragment prediction have not necessarily improved computer aided structure elucidation (CASE) from EI-MS and indicate that caution must be used when confirming the identity of a compound only based on the match between its predicted fragments and the mass spectrum.

Analysts dealing with unknown compounds have many questions to answer in order to prove that their "tentative identification" is in fact the correct compound. In many cases, for example, complex environmental samples, the initial information can be limited to a mass spectrum extracted from the output of gas or liquid chromatography coupled with mass spectrometric detection (GC/MS or LC/MS), where sample amount and/or purity prevent further analysis such as nuclear magnetic resonance spectroscopy (NMR) from being undertaken. Tentative identification of a compound from a mass spectrum can be made by conducting a library search (e.g., using the NIST database[1]), by hand or by using mass spectral "classifiers" to identify substructures and then building the matching molecule(s), either by hand or using structural generators such as MOLGEN.[2,3] The program MOLGEN-MS already combines the use of classifiers and structure generation[4–6] in the interpretation of low-resolution electron impact mass spectra (EI-MS) and has recently been extended to accept input of NIST classifiers.

Programs for Calculation of Fragments. One way of assessing whether a proposed structure could match a mass spectrum is to calculate the possible mass spectral fragments resulting from the structure and match these to the fragment masses that appear in the experimental mass spectrum (see Figure S-1 in the Supporting Information). Many general mass spectrometric fragmentation rules have been developed and published over the years for EI-MS.[7] These rules have been implemented, to various degrees, in a number of programs, including MOLGEN-MS,[5] MOLGEN-MSF[8] (a command-line version of part of MOLGEN-MS), Mass Frontier from HighChem,[9] and ACD MS Fragmenter. ACD MS Fragmenter is incorporated into ACD MS Manager,[10] which was used in this study; the functionality of the two programs, in terms of fragment prediction, is equivalent.

MOLGEN-MSF uses general mass spectral fragmentation rules[6] but can also accept additional fragmentation mechanisms as an optional input file when calculating the fragments. Mass Frontier, developed by HighChem and marketed by Thermo

* Corresponding author. Phone: +49 341 235 1490. E-mail: emma.schymanski@ufz.de.

† UFZ, Helmholtz Centre for Environmental Research-UFZ.

‡ DLR, German Aerospace Centre-DLR.

(1) NIST/EPA/NIH. NIST Mass Spectral Library; National Institute of Standards and Technology, U.S. Secretary of Commerce, U.S. Government Printing Office: Washington, DC, 2005.

(2) Benecke, C.; Grüner, T.; Kerber, A.; Laue, R.; Wieland, T. *Fresenius J. Anal. Chem.* **1997**, *359*, 23–32.

(3) Kerber, A.; Laue, R.; Grüner, T.; Meringer, M. *MATCH* **1998**, 205–208.

(4) Schymanski, E. L.; Meinert, C.; Meringer, M.; Brack, W. *Anal. Chim. Acta* **2008**, *615*, 136–147.

(5) Kerber, A.; Laue, R.; Meringer, M.; Varmuza, K. *Adv. Mass Spectrom.* **2001**, *15*, 939–940.

(6) Kerber, A.; Meringer, M.; Rücker, C. *Croat. Chem. Acta* **2006**, *79*, 449–464.

(7) McLafferty, F. W.; Turecek, F. *Interpretation of Mass Spectra*; University Science Books: Mill Valley, CA, 1993.

(8) Meringer, M. *MOLGEN-MSF*; Leonrodstr. 55, 80636 Munich, Germany, 2009. User manual available from www.molgen.de.

(9) HighChem. *Mass Frontier*, 5.0 trial version; HighChem Ltd./Thermo Scientific, 2007.

(10) ACD. *MS Manager*, version 11.01; Advanced Chemistry Development, Inc.: Toronto, Ontario, Canada, 2007.

10.1021/ac802715e CCC: $40.75
Published on Web 03/26/2009

Scientific Inc., generates the predicted mass spectral fragments according to general (basic) fragmentation rules, to specific library rules (either from a user library or the library provided by HighChem), or both. The library provided with the software contains 19 000 mechanisms taken from the literature.[11] MS Manager, from Advanced Chemistry Development Inc. (ACD) assigns generated fragments for a given structure to the given spectrum via the AutoAssignment option.[12] The output is a table of fragments and the "assignment quality index" (AQI), which summarizes the percent of the spectrum assigned by the calculated fragments in terms of the total ion chromatogram, TIC. The settings used for Mass Frontier and ACD are included in the Supporting Information (section S-3).

Ranking Structural Candidates. MOLGEN-MSF is able to rank structure candidates by generating "match values". Instead of attempting to predict the magnitude of predicted fragments, the magnitude of fragments in the experimental spectrum is assigned to the predicted fragments, as shown in the simplified equation presented below:

$$\mathrm{MV} = 1 - \sqrt{\frac{\sum_m (I(m) - x(m)I(m)^2}{\sum_m (I(m))^2}} \qquad (1)$$

where MV is the match value, m is the mass to charge (m/z) ratio of the fragment, $I(m)$ is the intensity of the experimental mass spectral peak at m (scaled to the base peak to a value between 0 and 1), and $x(m)$ indicates the presence/absence of predicted fragments such that $x(m) = 0$ if there is no predicted fragment for m and $x(m) = 1$ if there is a predicted fragment for m. Further discussion and an example are given in the Supporting Information (section S-1) and other references.[6,13]

The performance of MOLGEN-MSF was assessed previously[6] using 100 randomly selected spectra from the NIST database[1] (1998 version) and generating all constitutional isomers matching the molecular formula of the spectrum. An earlier version of MOLGEN-MSF was used to calculate match values for each constitutional isomer, and the structures were then sorted according to their match values, to rank the candidates and determine the position of the correct structure with respect to the other constitutional isomers. The rank of the candidates was expressed in terms of the "relative ranking position" (RRP), given in eq 2:

$$\mathrm{RRP} = \frac{1}{2}\left(1 + \frac{\mathrm{BC} - \mathrm{WC}}{\mathrm{TC} - 1}\right) \qquad (2)$$

where BC denotes the number of better candidates, i.e., those with a higher match value than the correct structure, WC denotes the number of worse candidates, and TC denotes the total number of candidate structures. The RRP ranges from 0 to 1, where RRP = 0 if the correct structure is ranked first (i.e., BC = 0), RRP = 0.5 if BC = WC, and RRP = 1 if the correct structure is ranked last (WC = 0).

The results[6] indicated that the use of general fragmentation rules alone was insufficient (in terms of accuracy) for automatic structure elucidation (i.e., to enable identification of the "correct structure"), but the authors suggested that the use of more sophisticated programs for virtual fragmentation may improve the ranking results.[6] Although some recent studies focused on high-resolution or tandem mass spectroscopic methods refer to both Mass Frontier and ACD MS Fragmenter for use in structure elucidation,[14,15] these were based on a limited set of candidates such as matching database entries[14] or a given set of precursor ions,[15] rather than all possible structures. Although alternative approaches exist to match structure to spectrum,[15] software packages such as MASSIS,[16] MASSIMO,[17] and EPIC[18] were not available to us for this study. The software FiD[15] for tandem MS data shows promising first results compared with Mass Frontier and it may be interesting to investigate this approach further in future studies where high-resolution data is available.

The present study compares MOLGEN-MSF with the two relatively prominent commercial programs Mass Frontier and ACD MS Manager. Using the previous study[6] as a baseline, we assess the ability of these programs to identify the "correct structure" (in this paper, the compound that was analyzed to produce the mass spectrum) for a given mass spectrum, from all constitutional isomers for a given molecular formula. Furthermore, the influence of different program settings for these commercial programs (inclusion of library reactions provided with Mass Frontier and the use of additional fragmentation steps compared with the default settings) was also assessed.

METHODS

The electron impact mass spectra considered in this study were retrieved from the NIST05 Mass Spectral Database[1] by spectrum number and saved in the mass spectral transfer (MSP) format.[19] All programs were assessed, where possible, using the 100 randomly selected spectra from the previous study,[6] to ensure comparability. Spectra no longer available in the 2005 NIST database were recovered from the archive of the Kerber et al. study for consistency. Minor adjustments to the MSP format were made, where necessary, for import into the different programs (see section S-2, Supporting Information). The much longer calculation times for ACD MS Manager and Mass Frontier with library reactions required a reduction in the data set to spectra with less than 500 constitutional isomers (41 spectra) and 200 constitutional isomers (27 spectra), respectively.

(11) HighChem, Mass Frontier User Information, http://www.highchem.com/massfrontier/mass-frontier.html, 2007.

(12) *ACD ACD/MS Manager and Processor Reference Manual*, version 11.0; Advanced Chemistry Development, Inc., 2007.

(13) Meringer, M. *Mathematical Models for Combinatorial Chemistry and Molecular Structure Elucidation*; Logos-Verlag: Berlin, Germany, 2004 (in German).

(14) Hill, D. W.; Kertesz, T. M.; Fontaine, D.; Friedman, R.; Grant, D. F. *Anal. Chem.* **2008**, *80*, 5574–5582.

(15) Heinonen, M.; Rantanen, A.; Mielikainen, T.; Kokkonen, J.; Kiuru, J.; Ketola, R. A.; Rousu, J. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3043–3052.

(16) Fan, B. T.; Chen, H. F.; Petitjean, M.; Panaye, A.; Doucet, J. P.; Xia, H. R.; Yuan, S. G. *Spectrosc. Lett.* **2005**, *38*, 145–170.

(17) Gasteiger, J.; Hanebeck, W.; Schulz, K. P. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 264–271.

(18) Hill, A. W.; Mortishire-Smith, R. J. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111–3118.

(19) D'Arcy, P.; Mallard, W. G. *AMDIS-User Manual*; National Institute of Standards and Technology (NIST), U.S. Department of Commerce, U.S. Government Printing Office: Washington, DC, 2004. Available from http://chemdata.nist.gov/mass-spc/amdis/.

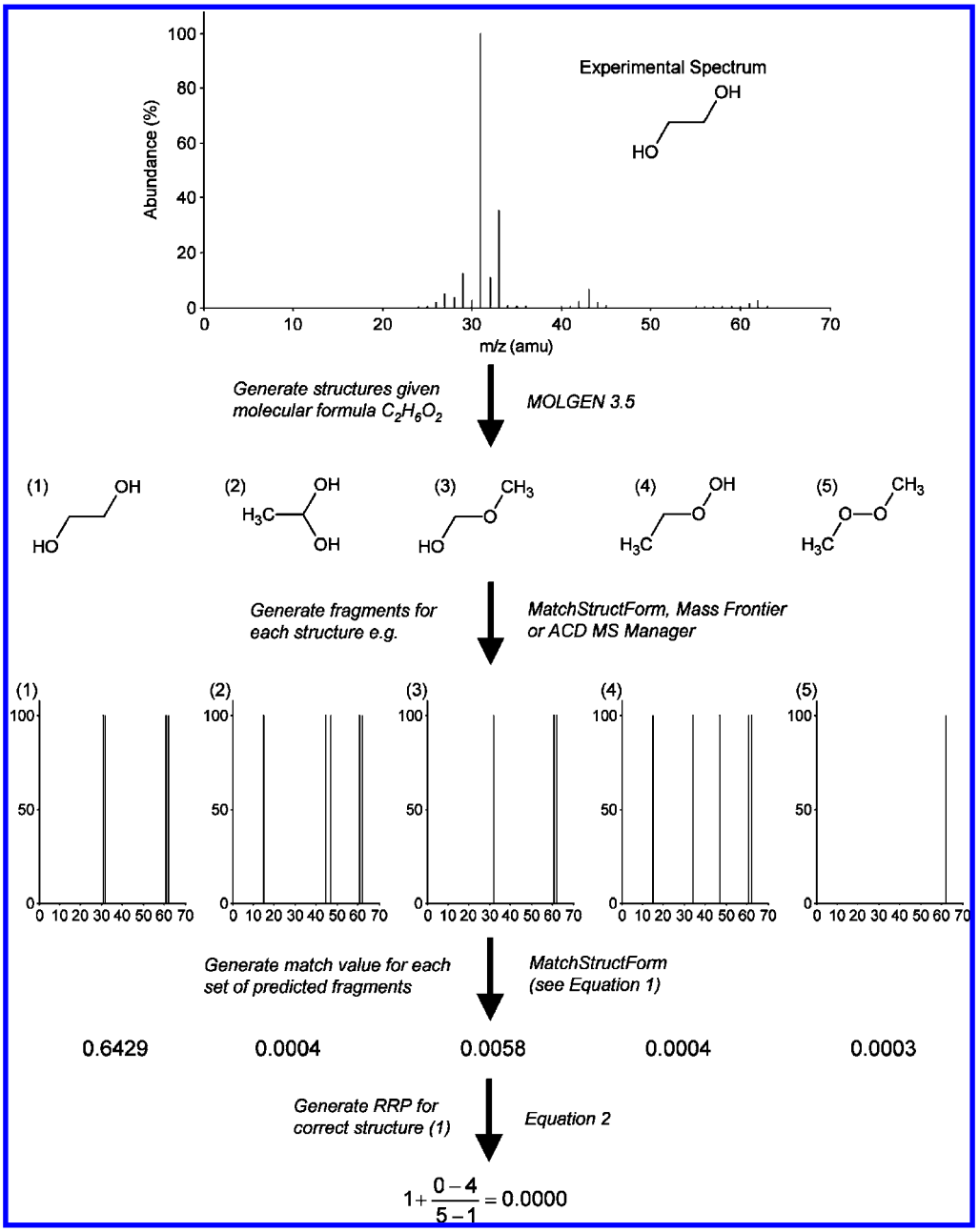

Figure 1. Schematic for matching candidate structures to an experimental spectrum using fragmentation patterns. All possible structures are generated for the formula from the experimental spectrum, fragments are predicted for each structure, the match value is calculated to match the fragments to the experimental spectrum, and finally the match values are used to determine the number of "better" and "worse" candidates for calculation of the relative ranking position (RRP) of the correct structure.

The candidate structures (constitutional isomers) for each spectrum and for the three specific examples used in this paper were generated using MOLGEN 3.5[2], with no restrictions unless indicated otherwise. The molecules were saved in the MDL SDF format,[20,21] hereafter referred to as "SDF format". Specific details regarding the generation of fragments by each program are given in the Supporting Information.

The match value (see eq 1) was used in this study to compare the results generated by all programs, as it requires only the input of the fragments generated by each program and uses the peaks from the experimental spectrum to assign magnitudes. Once match values are calculated for each structure, the relative ranking position can be calculated, to assess the accuracy or selectivity of the different programs. The basic concept, from generation of structures through the calculation of the relative ranking position is presented in Figure 1.

The code of MOLGEN-MSF was extended to read the Mass Frontier and ACD outputs to enable a consistent match value calculation for all programs. The input for match value calculations also included the experimental spectrum, typically in the MSP format. The "assignment quality index" (ACD MS Manager) was calculated for all ACD runs, to compare with the match value.

(20) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.

(21) Symyx Technologies, Inc., MDL. *CTFile Formats*; Symyx Technologies, Inc.: San Ramon, CA, 2007. Available from http://www.symyx.com/downloads/public/ctfile/ctfile.pdf.

Table 1. Abbreviations to Describe the Programs and Settings Used[a]

abbreviation	description
MSF	MOLGEN-MSF fragmentation
MF_3st	Mass Frontier, 3 step fragmentation, general fragmentation rules only
MF_3st_wLib	Mass Frontier, 3 step fragmentation, general and library fragmentation rules
MF_5st	Mass Frontier, 5 step fragmentation, general fragmentation rules only
MF_5st_wLib	Mass Frontier, 5 step fragmentation, general and library fragmentation rules
ACD_3st	ACD MS Manager AutoAssignment, 3 step fragmentation
ACD_5st	ACD MS Manager AutoAssignment, 5 step fragmentation
ACD_3st_AQI	ACD MS Manager AutoAssignment, 3 step fragmentation, results expressed in terms of the "assignment quality index"
ACD_5st_AQI	ACD MS Manager AutoAssignment, 5 step fragmentation, results expressed in terms of the "assignment quality index"

[a] Unless stated otherwise, results obtained from each program are presented in terms of the match value given in eq 1, while those with the "AQI" suffix are expressed in the ACD assignment quality index.

The definition of the assignment quality index was not sufficient for us to attempt to reproduce this calculation for fragments generated using Mass Frontier and MOLGEN-MSF. Further discussion is provided in the Supporting Information, section S-1.

Kerber et al.[6] also used simple statistics to assess the results of structure fragmentation and ranking. They defined confidence intervals to indicate how many structures need to be considered for a given spectrum to ensure inclusion of the correct structure with a certain probability, using an independent random selection of 1000 structure−spectrum pairs. For these spectra, the match values were calculated only for the correct structure, not all constitutional isomers. The same 1000 spectra from the previous study were used to calculate confidence intervals in this study, for all program and settings combinations, except Mass Frontier with library fragmentation (due to the long calculation time required for the large structures included in the 1000 spectra). The definition and explanation of the confidence interval concept can be found in Kerber et al.[6]

RESULTS

To simplify presentation of the results, each program and the corresponding settings have been given a short name. The abbreviations and explanations are shown in Table 1.

The results for MOLGEN-MSF and Mass Frontier 3 and 5 step fragmentations using the general fragmentation rules only for the 100 randomly selected spectra are shown in the Supporting Information, Table S-2. The match values for the correct structure calculated for all program and settings combinations for the reduced data set of 27 spectra (those spectra with less than 200 structures) are given in Table 2. Relative ranking positions for the same data set are presented in the Supporting Information, Table S-3. The quantiles calculated for the different programs and settings are presented in the Supporting Information, Table S-4, excluding Mass Frontier with library fragmentation due to the long computation times for large structures.

Several other parameters calculated for each spectrum that were presented by Kerber et al.[6] as well as the results for the ACD MS Manager calculations for spectra with 200−500 possible structures have been excluded due to space considerations. Instead, average values of these parameters are presented in the Supporting Information (Table S-5). The average relative ranking positions calculated for the different programs and settings, for spectra with 0−200, 0−500, and 0−10 000 structures are plotted in Figure 2. This shows that the average relative ranking position is larger (i.e., worse) for spectra with few possible structures, compared with all spectra. This trend, which was apparent in all programs, indicates that the ranking success of the match value (and AQI) is generally worse for spectra with few candidate structures.

The selection of specific examples, in addition to the randomly selected spectra presented above, provides additional insight into the performance of each program by allowing the consideration of phenomena specific to certain structures. Three examples are used in this paper to evaluate the use of fragmentation to match structures to their mass spectra. The formulas were selected based on the presence of several spectra for the given formula in the NIST database, where some of the spectra were clearly different from others (specifically containing different peak groups, not just different magnitudes of peaks). The examples were also chosen for the low number of possible structures (<100), to aid in interpretation and presentation of results.

Specific Example 1: $C_3H_5O_2Cl$. The first formula, $C_3H_5O_2Cl$, contains two oxygens and one ring or double bond equivalent (RDBE), consistent with molecules such as carboxylic acids, keto ethers, esters or cyclic ethers, and alcohols, with significant differences in fragmentation possibilities. The number of possible molecules generated using MOLGEN is 84 (excluding those structures with O−Cl bonds; including O−Cl bonds results in 110 structures). The six NIST spectra with this formula (excluding stereoisomers) are shown in the Supporting Information, together with the structures, Figure S-2.

The match values for the correct structure for a given spectrum are listed in Table 3 and the ranking results of the six spectra in figure format (Figure 3). The idea of this figure is to compare the ranking of the six "known" molecules with each other, to see how well each program (with different settings) matches the structure and spectrum. If the programs match the correct structure and spectrum pair, the pattern of the top ranked structure for each spectrum, indicated by a cross, should be in a diagonal from top left to bottom right, indicated by the bold outlined squares in each matrix.

The matrices show a couple of interesting features for these small structures. With the exception of MF_5st_wLib, the results for Mass Frontier and MOLGEN-MSF are relatively accurate and comparable, picking the correct structure of the six spectra three to five times. Although MF_5st (with general reactions only) was the most accurate for these six structures, including the library fragmentation changed the situation dramatically, selecting the correct molecule only twice and additionally giving structure 1 the highest match value for all runs. In comparison with the Mass Frontier and MOLGEN-MSF results, the distribution of results from ACD is far more chaotic, with often several structures selected as the best match. The ACD match values and assignment quality indices were much higher than for Mass Frontier

Table 2. Match Values of the Correct Structure Calculated for 27 Spectra for All Programs and Settings[a]

			match value of correct structure						
no.	formula	possible structures	MSF	MF_3st	MF_5st	MF_3st _wLib	MF_5st _wLib	ACD_3st	ACD_5st
4	C_7H_{14}	56	0.2631	0.2668	0.7077	0.2899	0.7762	0.9375	0.9644
10	CN_3F_5	11	0.0000	0.0551	0.2280	0.0551	0.2280	0.0551	0.0551
13	CH_5SiBr	2	0.0366	0.0293	0.0293	0.0293	0.0293	0.5205	0.5358
15	$C_5H_{11}Br$	8	0.0595	0.1389	0.1539	0.5014	0.6043	0.9450	0.9729
19	C_2H_3NO	26	0.1454	0.0010	0.0010	0.0010	0.0942	0.4821	0.4821
34	$C_{11}H_{24}$	159	0.5614	0.5511	0.5511	0.8990	0.8995	0.9530	0.9763
35	C_8H_{16}	139	0.1416	0.1367	0.1382	0.1367	0.6799	0.8560	0.9527
37	C_9H_{20}	35	0.5628	0.5628	0.5628	0.8330	0.8332	0.9252	0.9435
40	$C_5H_{13}N$	17	0.8367	0.8350	0.8407	0.8644	0.8648	0.9812	0.9836
42	$C_6H_{14}O$	32	0.0528	0.0125	0.0601	0.1775	0.7176	0.9623	0.9826
45	$C_5H_{12}O_2$	69	0.2624	0.0256	0.0326	0.1453	0.1617	0.8386	0.8413
50	$C_2H_6O_2$	5	0.6429	0.6307	0.6307	0.8634	0.8658	0.6755	0.6755
52	C_5H_6	40	0.4656	0.3690	0.3690	0.5321	0.5321	0.6303	0.6303
54	$C_8H_{17}Cl$	89	0.0592	0.0363	0.2249	0.3877	0.4864	0.9151	0.9639
59	$C_4H_{12}N_2$	38	0.7545	0.7566	0.7566	0.7733	0.8614	0.9944	0.9945
60	$C_3H_3Cl_3$	8	0.0019	0.6502	0.6502	0.6521	0.6521	0.7820	0.7980
61	$C_5H_{13}N$	17	0.5151	0.7369	0.7369	0.8148	0.8285	0.9554	0.9593
66	C_2H_7P	2	0.1597	0.1597	0.1597	0.1597	0.1597	0.5337	0.5337
68	$C_5H_{13}NO$	149	0.6480	0.6499	0.8028	0.9135	0.9149	0.9148	0.9408
72	$C_4H_{11}NO$	56	0.7706	0.7712	0.7724	0.8502	0.9241	0.9929	0.9929
73	C_6H_{10}	77	0.0896	0.6213	0.6213	0.6383	0.6383	0.8586	0.8680
74	C_2NF_3	5	0.4977	0.6830	0.6830	0.6830	0.6830	0.7857	0.7857
80	C_3H_7NO	84	0.6177	0.1824	0.1824	0.6206	0.6313	0.9766	0.9803
81	$C_3H_7O_2Br$	38	0.0992	0.1001	0.4454	0.2888	0.8528	0.9804	0.9804
84	C_8H_{16}	139	0.6245	0.6023	0.6091	0.7566	0.7679	0.8287	0.9804
96	C_3H_4O	13	0.6550	0.0270	0.0270	0.0270	0.0270	0.7350	0.7350
97	C_4H_5OCl	175	0.0255	0.6993	0.7014	0.9056	0.9094	0.9371	0.9848
	averages		0.354	0.381	0.433	0.511	0.616	0.813	0.833

[a] Abbreviations set out in Table 1. The number in the first column corresponds with the spectrum number in Table S-2 in the Supporting Information.

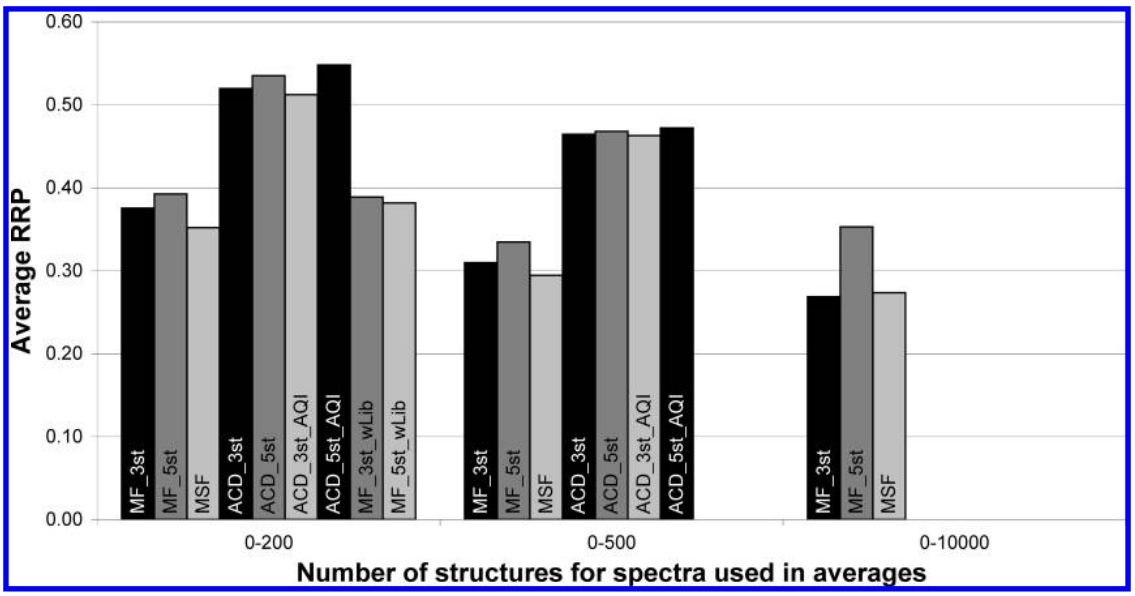

Figure 2. Average relative ranking positions (RRPs) for the different programs and settings, taken over spectra with 0−200, 0−500, and 0−10 000 structures.

and MOLGEN-MSF (shown by the shading in Figure 3), but all structures had high match values, not just the correct ones, a fact that does not reflect the differences in the structures and spectra shown in Figure S-2 in the Supporting Information.

These results were reflected in the relative ranking positions calculated for the programs for all 84 possible structures. The average relative ranking positions calculated for the six spectra for MOLGEN-MSF and Mass Frontier ranged between 0.0412 (MF_3st), meaning the correct structure is in the top 4%, to 0.2279 (MF_5st_wLib), with MOLGEN-MSF in the middle (0.1486). In contrast, the relative ranking positions for ACD ranged between 0.3404 (ACD_3st) and 0.3936 (ACD_5st), such that the correct structure is only in the top 34 and 40% of all structures, reflecting the lack of specificity demonstrated in the matrices in Figure 3.

Comparing the matrices with the data included in Table 3 also indicates some counterintuitive rankings for the match values

Table 3. Match Values and Assignment Quality Indices of the Actual Molecule to the Spectrum from NIST, Predicted by the Various Programs with Different Settings[a]

	match values							assignment quality indices (%)	
spectrum	MSF	MF_3st	MF_3st _wLib	MF_5st	MF_5st _wLib	ACD _3st	ACD _5st	ACD_3st _AQI (%)	ACD_5st _AQI (%)
1	0.1776	0.0938	0.1832	0.0938	0.1895	0.9640	0.9669	97.8	98.4
2	0.0018	0.0472	0.0479	0.0492	0.0508	0.9550	0.9554	96.5	96.8
3	0.6754	0.8119	0.9670	0.8633	0.9670	0.9884	0.9884	100.0	100.0
4	0.0264	0.5493	0.5513	0.5494	0.5648	0.9877	0.9883	98.9	99.1
5	0.6206	0.6206	0.6211	0.6336	0.6372	0.6336	0.6372	97.1	97.5
6	0.7556	0.7559	0.7559	0.7625	0.7627	0.9786	0.9791	96.9	97.4
average	0.376	0.480	0.521	0.492	0.529	0.918	0.919	97.9	98.2

[a] Abbreviations are given in Table 1.

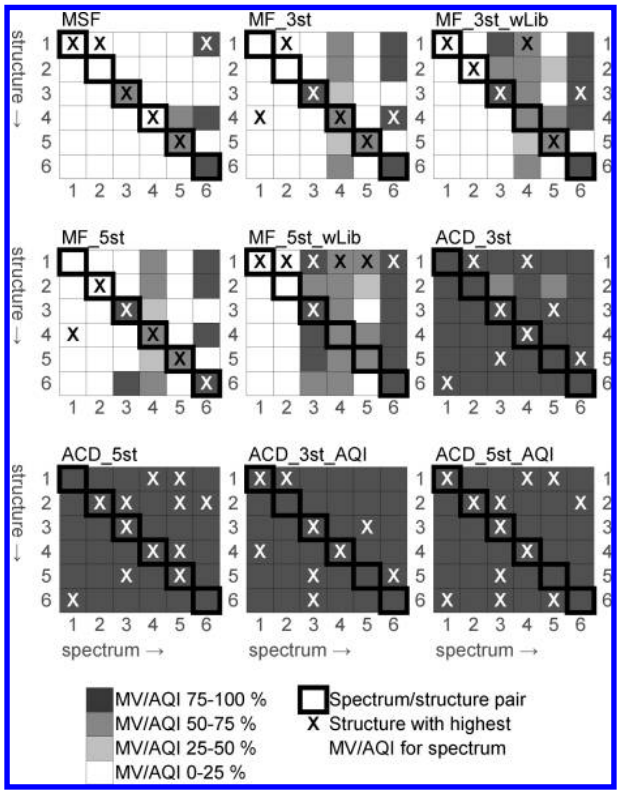

Figure 3. Matrices of the six $C_3H_5O_2Cl$ structures (rows) and spectra (columns). The bolded squares indicate the structure−spectrum pair (i.e., structure 1 matches spectrum 1). The crosses indicate the structure with the highest match value of the six structures, for a given spectrum, such that each column has at least one cross. More than one cross for a spectrum (column) indicates two or more structures with the same match value or assignment quality index. The shading indicates the approximate match value, as shown in the legend. The program abbreviations are given in Table 1.

calculated for the actual molecules. Although MOLGEN-MSF only predicts a match value of 0.0264 for spectrum 4, the MSF matrix indicates that this structure was correctly identified as the correct match for this spectrum, i.e., this match value was higher than the match value for the other five spectra. In contrast, although structure 3 has an assignment quality index of 100% for spectrum 3, the ACD_5st_AQI matrix in Figure 3 shows that at least three other spectra also had an assignment quality index of 100%, so that this is less selective than the much lower match value generated by MOLGEN-MSF for spectrum 4. Additionally, although the correct structure for spectrum 6 has a match value close to 0.98 for the 3 and 5 step ACD calculations, this structure is not identified correctly for this spectrum. This shows that the match value of the correct structure gives little information about the ranking position of this structure in relation to all other possible structures.

Specific Example 2: $C_5H_{12}S_2$. The second formula, $C_5H_{12}S_2$, includes dithiols, alkyl thiols, or disulfides, with the main differences in the mass spectra resulting from differences in alkyl substitutions and symmetry of the structures. The number of possible structures generated in MOLGEN using 2-valent sulfur is 69. There are 11 NIST spectra with this formula, shown in the Supporting Information along with the structures (Figure S-3). The matrices presenting the ranking of the correct structure compared with the other "known" structures (from NIST spectra) are presented in the Supporting Information (Figure S-4). The match values were again generally much higher for the ACD calculations than for MOLGEN-MSF or Mass Frontier.

The matrices demonstrate an interesting trend, with several programs showing a bias toward certain structures, including structure 2 (MSF, MF_3st, ACD_5st), structure 6 (MSF, MF_3st), and structure 11 (MF_3st, MF_5st, ACD, especially in the assignment quality indices), indicated by many crosses in each row. What is also interesting with these matrices is the structures that are not selected (no crosses in a row), structures 1 and 3 are never selected (either correctly or falsely), whereas structure 10 is only selected twice. The favoring of structures 2 and 6 could be because these are the least symmetrical chain molecules (i.e., with the most possible fragments), while structures 1 and 3 are branched structures with few possible fragments. The average number of fragments predicted for each structure (averaged over the 11 spectra) for each program and settings combination are presented in Table 4. The average number of fragments generated for structures 2 and 6 (46.8 and 45.4, respectively) are much greater than the average (39.9), while the number of fragments generated for structures 1 and 3 (32.5 and 31.0, respectively) are much lower.

Specific Example 3: $C_7H_6Cl_2O$. The third formula, $C_7H_6Cl_2O$, has 155 987 possible structures, although considering only those with a benzene ring present reduces this to 49 structures. There are 12 spectra in NIST with this formula, shown in the Supporting Information, Figure S-5. This formula was chosen to assess the ability of the different programs to discern between aromatic substitution isomers, as we had encountered difficulties identifying unknown spectra with this formula.

The matrices from this example are in the Supporting Information (Figure S-6) and show the influences that the choice of a

Table 4. Number of Fragments Predicted for Structures 1−11, Averaged over All Spectra, For Each Program and Settings Combination[a]

structure	MSF	MF_3st	MF_3st _wLib	MF_5st	MF_5st _wLib	ACD_3st	ACD_5st	average[b]
1	21.6	29.0	33.0	29.0	35.0	37.4	42.5	32.5
2	38.8	40.6	48.6	48.2	64.3	41.5	45.5	46.8
3	22.5	20.6	24.3	20.6	49.6	37.2	42.5	31.0
4	28.1	39.5	45.4	48.2	54.5	38.9	42.6	42.4
5	29.7	39.7	50.8	43.4	58.1	41.0	44.9	43.9
6	36.1	43.0	53.0	46.8	63.3	35.0	40.6	45.4
7	30.9	29.3	49.3	37.3	64.0	39.0	41.5	41.6
8	24.0	37.4	39.4	37.4	41.4	38.7	42.8	37.3
9	21.5	28.3	45.8	38.1	59.4	37.0	43.3	39.0
10	31.7	20.7	41.2	20.7	60.1	36.8	44.5	36.5
11	31.5	35.4	42.5	44.3	55.4	42.3	46.0	42.5
average[b]	28.8	33.0	43.0	37.6	55.0	38.6	43.3	39.9

[a] Program abbreviations are given in Table 1. [b] The bottom row contains the average number of fragments generated for all structures and spectra for each program, whereas the final column contains the average number of fragments generated for that structure, over all programs.

program and/or settings within the program can have on the outcome of structure ranking. Almost all of these cases show a bias in the program toward one structure above the others, but the actual structure ranked highest changes with minor modifications to program settings, especially for Mass Frontier.

The average match values and assignment quality indices calculated for the correct structure for ACD results in this example (shown in Table S-7 in the Supporting Information) were much lower than the averages over the randomly selected spectra (shown in Figure 2 and in Table S-5 in the Supporting Information), for example, an average match value of 0.694 compared with 0.860 over 41 spectra for ACD_3st. The relative ranking positions for MOLGEN-MSF and Mass Frontier were also much worse than the average relative ranking positions presented in Table S-5 in the Supporting Information, which partially explains our problems with identifying aromatic structures using match values. MF_3st was the only calculation to successfully group any of the positional isomers together with the highest match values (for the dichloromethoxybenzene isomers), while including the library reactions split the isomer grouping to the detriment of the overall results.

Structures 7, 8, and 11 feature very strongly in the results, with structure 8 being the only structure picked correctly in all runs except MF_5st_wLib (where no molecule was picked correctly). Structure 8, 4-chloro-1-chloromethoxybenzene, is very different from all the other molecules but does not show an above-average number of fragments (Table S-6 in the Supporting Information). The selection of structure 11 in many cases but not the positional isomers 9 and 10 can be clarified, at least partially, by the number of fragments generated (again see Table S-6 in the Supporting Information). We suspect again that this is in part due to the molecular symmetry, structures 9 and 10 have a degree of symmetry in them, whereas all substituents on structure 11 are on one side, leaving more fragmentation possibilities open. This is discussed further in the Supporting Information.

Using Classifiers to Eliminate Structure Candidates. As alluded to in the introduction, there are a number of additional strategies available for limiting the number of candidate structures for EI-MS, prior to calculation of the fragments and match values. One of these is the use of mass spectral classifiers to identify substructures present or absent in the mass spectrum, based on the fragmentation patterns. This has already been implemented in the program MOLGEN-MS and is discussed further by Schymanski et al.[4] MOLGEN-MS has since been extended to read in NIST classifier information to assist in incorporation of this prior to structure generation.

Table 5. Number of Constitutional Isomers and the Relative Ranking Both without and with the Consideration of Mass Spectral Classifiers (Varmuza and NIST Classifiers) for Spectra from Specific Example 1 ($C_3H_5O_2Cl$)[a]

	without classifiers				with classifiers (95% probability)			
spectrum	TC	BC	EC	RRP	TC	BC	EC	RRP
1	84	4	1	0.0482	2	0	1	0.0000
2	84	34	1	0.4096	1	0	1	
3	84	16	5	0.2169	1	0	1	
4	84	5	2	0.0663	19	3	2	0.1944
5	84	2	1	0.0241	1	0	1	
6	84	12	1	0.1446	1	0	1	

[a] MOLGEN-MSF was used to calculate the match values. TC, total number of candidates; BC, number of candidates with a higher match value; EC, number of candidates with match value equal to that of the correct structure (EC = 1 if only the correct structure has that match value); RRP, relative ranking position (see eq 2).

To demonstrate the use of both NIST and Varmuza[22,23] classifiers in reducing the number of candidate structures for a spectrum, we have taken specific example 1, $C_3H_5O_2Cl$. A comparison of the results with and without classifiers is presented in Table 5, using MOLGEN-MSF to calculate the match values. This table shows clearly, even for this small example, that the use of classifiers is instrumental in reducing the number of candidate structures prior to fragment generation and hence improving the chance of identifying the correct structure and limiting the number of other structures with higher match values. In four of the six spectra, the use of classifiers reduces the data set from 84 molecules to 1, in each case the correct structure. This means a greatly reduced data set for identification purposes and in some cases a relatively robust tentative identification, when no (or few) other molecules are possible for the given classifiers. There are cases where the classifiers can be wrong, hence the probabilities associated with classifier assignment. Discussion on these cases can be found for Varmuza classifiers[22] or more

(22) Varmuza, K.; Stancl, F.; Lohninger, H.; Werther, W. *Lab. Autom. Inf. Manage.* **1996**, *31*, 225–230.
(23) Varmuza, K.; Werther, W. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 323–333.

generally.[4] Additional compound properties can also be used to eliminate candidate structures that do not match the experimental data, in addition to spectral classifiers.[4,14]

DISCUSSION

Does a High Match Value Mean a Better Ranking? It is a feature of human nature to automatically react positively to a structure with a high match value and negatively to a structure with a low match value. We hope that the data presented in this article enables readers to question this automatic response, as it is clear from the results presented here that a positive reaction to a high match value can lead to a false sense of expectation regarding identification of a structure. Taking spectrum 5 from Table S-2 in the Supporting Information, although MF_5st has the highest match value (0.726), it has the worst RRP (0.501), compared with MF_3st (MV = 0.637, RRP = 0.185) and MSF (MV = 0.719, RRP = 0.123). This also demonstrates another important fact, that the match values between the programs and settings are not directly comparable. In this case a very small difference in the match value between MF_5st and MSF masks a large difference in the structure ranking, where for MSF only 12% of the other possible structures have higher match values, while for MF_5st, 50% of the possible structures have higher match values. This makes selection of the "correct" structure based on match values alone (i.e., setting a match value "threshold") challenging.

As the variation in individual examples is huge, we calculated the averages for the different program, setting, and spectrum combinations (see Table 2 for the averages over 27 spectra and Figure 2 for the RRPs). The average match value for the correct structure for MSF (0.354) and MF_3st (0.381) are lower than most people's "positive response limits", while that for ACD_5st (0.833) is significantly higher. However, despite the high match values assigned by ACD_5st, this program setting combination demonstrated the least selectivity, with the worst average relative ranking position over the 27 spectra, at 0.535. This relative ranking position means that, on average, over 53% of the constitutional isomers have a greater match value than the correct structure. If the match values had been assigned randomly to all structures, the average relative ranking position for the correct structure would be 0.50, meaning that ACD_5st is actually, on average over these 27 spectra, slightly worse than randomly assigning match values to each structure. In contrast, MSF and MF_3st, despite having low average match values, have the best relative ranking positions at 0.352 and 0.375, respectively, over the 27 spectra. MF_5st, although producing higher match values than MF_3st (due to the calculation of more fragments), experiences a corresponding loss of selectivity, with the average relative ranking position increasing relative to MF_3st for all averages (Table S-5 in the Supporting Information). The Mass Frontier calculations with library reactions were also less selective than MF_3st. Thus, although the use of additional settings increases the match values in all cases (as one would expect when the number of fragments increases) this is accompanied in all cases with a loss of predictive selectivity, demonstrated by the increasing relative ranking position. Although the simpler settings may miss many specific fragmentation pathways for the correct structure, it is clear that on average the additional fragmentation pathways increase the specific fragmentations for the "incorrect" molecules more than for the correct molecules.

Figure 2 shows that this trend is consistent also for comparison over the smaller data sets; for instance, including spectra with up to 500 structures improves the ACD_3st and ACD_5st relative ranking positions to below 0.5 (0.465 and 0.468, respectively), but this is still significantly higher than the comparable average relative ranking positions for MF_3st and MSF (0.29 and 0.31, respectively). The increase in the average relative ranking position with decreasing number of structures can be explained by considering the variation in the structures. For a small data set (e.g., spectrum 61, $C_5H_{13}N$ with 17 possible structures), there are few variations in the combination of atoms in generating the structures, which corresponds to a decrease in the different fragmentation possibilities between the structures and thus decreases the probability of being able to use fragmentation to distinguish the structures successfully. Contrarily, large sets of structures have, generally, more combinational possibilities, greater numbers of possible fragmentation pathways, and hence greater differences between the match values predicted for the structures.

Match Value vs Assignment Quality Index. Another issue we wish to draw the reader's attention to is the use of "black-box" indicators. The prediction of energies and barriers in the creation of fragments is difficult and is at this stage not sufficiently investigated to allow for incorporation into a spectrum−structure match.[11,24] The match value calculation, by taking the magnitude of the peaks from the experimental mass spectrum (see eq 1), does not attempt in any way to predict the abundance of fragments but instead uses the only information available (experimental) and provides a compromise solution while the prediction of fragment intensity remains in its infancy. The match value can also be calculated for any set of fragments, as long as this can be exported from the program generating the fragments in some way.

In contrast, while the ACD assignment quality index attempts to incorporate the magnitude as well as presence of fragmentation peaks in the mass spectrum, the results included here, as well as several not included, left us regarding this value with some skepticism. A general demonstration of the assignment quality index distribution is in the Supporting Information, Table S-4, which contains the quantiles for the 1000 randomly selected spectra. This shows that 95% of the structures given an assignment quality index will have a value above 46% (3 step) or 48.9% (5 step), compared with the corresponding quantile for match values calculated from the ACD results of 0.23 and 0.26, indicating that only 23−26% of the experimental spectrum (abundance) is covered by the ACD fragments counted in the assignment quality index. Likewise the 99% quantile is 100%, implying that for a spectrum with 10 000 constitutional isomers, on average 1% or 100 structures will have an assignment quality index of 100%, which makes selection between these candidates impossible and, given the average relative ranking position for the ACD calculations, the top 1% of structures is extremely unlikely to include the correct candidate structure anyway.

We also draw the reader's attention to the discrepancies between the ACD results processed with the match values in comparison with the assignment quality index results in specific

(24) Lehotay, S. J.; Mastovska, K.; Amirav, A.; Fialkov, A. B.; Martos, P. A.; Kok, A. d.; Fernández-Alba, A. R. *TrAC, Trends Anal. Chem.* **2008**, *27*, 1070–1090.

examples 1 and 2 (see Figure 3 and Figure S-4 in the Supporting Information) and, to a lesser degree, in specific example 3 (Figure S-5 in the Supporting Information). In these figures it is apparent that the assignment quality index changes the relative ranking of the structures compared with the match value, in some cases with significant differences in the top candidate selection (see especially the increased bias toward structure 11 in Figure S-4 in the Supporting Information). As the assignment quality index is not clearly defined,[12] we are not able to shed light on the nature of the differences. The developers themselves also offer a few words of caution regarding this index, adjacent to its description.[12]

Another reason to exercise caution when interpreting the results of ACD is the presentation of only the fragments present in the experimental spectrum, not all fragments calculated. The absence of predicted fragments that are not in the experimental spectrum results in the loss of a crucial additional interpretation tool. Both MOLGEN-MSF and Mass Frontier perform fragmentation calculations independent of the mass spectrum, fragmenting each structure according to the given rules/settings and then comparing these results with the experimental spectrum. Although the match value is only calculated on those fragments present in the experimental spectrum, the alternative outputs in MOLGEN-MSF include the export of all fragments, such that this information is still recoverable to the user, both for MOLGEN-MSF and Mass Frontier inputs. As the ACD calculation requires input of the experimental spectrum from the beginning, it appears that the "additional" fragments (i.e., those not present in the experimental spectrum) are filtered out of the results before these are presented to the user. We were not able to find any possible adjustment to the settings to export all fragments generated, rather than just those present in the experimental spectrum. Given that the ACD match values are significantly higher than those for MOLGEN-MSF and Mass Frontier, it was our suspicion that the ACD program calculated many more fragments, both present and absent, than either MOLGEN-MSF or Mass Frontier, but we are unable to confirm this at this stage.

Candidate Inclusion/Exclusion. The results shown in this paper highlight the problems associated with considering a limited subset of constitutional isomers and using the assigned fragments to prove (or disprove) the match of the structure to spectrum. Several examples above have shown that in many cases the "correct" molecule can have a very low match values or that several other molecules can have much higher match values, such that distinguishing "correct" from "incorrect" is very difficult based on the match value or fragmentation patterns alone. Even the best program and setting combinations (Mass Frontier with 3 step fragmentation and MOLGEN-MSF) can only reduce the number of possible candidates (on average) to 27% of all possible molecules for that spectrum's molecular formula, although to ensure inclusion of the correct structure with 90% certainty, many more molecules have to be included in most cases (expressed by the quantiles in Table S-4 in the Supporting Information). While consideration of all possible structures allows at least an objective overview of the match value range, consideration of a limited subset (e.g., only those structures in a database) is unlikely to give the full distribution of match values and could result in incorrect selection of an apparently good match. Even for spectra with a small number of possible structures (say 100), the inclusion of a significant percentage of the total possible candidates can result in consideration of over 30 candidate structures, which is already impractical for rapid identification/confirmation purposes. MS classifiers can be used to reduce the number of candidate structures prior to calculation of match values, shown above in Table 5.

Which Program/Which Settings? A detailed discussion on the time for calculation, the relative cost of and restrictions associated with each program is provided in the Supporting Information, Section S-5. Although ACD MS Manager is significantly cheaper and more accessible than Mass Frontier, it should be used with caution to assess proposed structures based on its predicted fragments, as the ranking results are very close to that of a random number generator. Further discussion is in the Supporting Information, but it should be noted that both the ACD MS Manager and Mass Frontier contain many other settings and features that we have not considered here, which may be useful for other purposes.

CONCLUSIONS

Despite the positivity expressed in 2006[6] in relation to the improvement in fragmentation and match value calculation by using more sophisticated computer programs, we have not been able to produce the desired improvements so far. The results presented here show convincingly that the simplest and quickest of the program and settings combinations (Mass Frontier with 3 step fragmentation and MOLGEN-MSF) are still the most effective in terms of ranking the correct structure relative to other constitutional isomers based on electron impact mass spectra, despite the lower match values. Longer calculation times with more fragmentation steps or including library reactions to produce higher match values generally resulted in a decreased selectivity.

The specific examples used in this study demonstrate the bias of all programs toward certain structures in many cases, even for different mass spectra with the same molecular formula. This bias can change significantly with minor changes in program settings and is often related to the number of fragments predicted in total, not just fragments present in the mass spectrum. Specific examples 2 and 3 show that the bias is often toward the more asymmetrical molecules (with a correspondingly greater number of possible fragments) and away from the symmetrical molecules (with fewer possible fragments resulting from the symmetry), such that asymmetrical molecules will be selected more often, whether correct or incorrect.

At this stage the match value, which uses the experimental spectrum to assign magnitudes to the predicted fragments, appears to be the only well-defined way to match fragments of a structure to the mass spectrum. We were not able to fully understand the trends shown by the assignment quality index generated by the ACD MS Manager and are a little skeptical of its ability to take fragment intensity into account. We would therefore caution users against using a "black-box" indicator without fully understanding the calculations behind it. Although it is apparent that the number of fragments produced by the programs has some influence on the program bias toward some structures, we have not yet been able to incorporate this information in any meaningful way to result in a positive impact on the relative ranking position. Otherwise the way to improve the assessment of the fragment−spectrum relationship would be to

focus efforts on the prediction of fragment intensity, despite the inherent problems involved.

As we were unable to identify a combination of program settings to improve the relative ranking position significantly above the 0.27 reported previously,[6] we are unable to improve on their conclusion that the generation of fragments and match values alone are not sufficient, at this stage, to allow for computer-aided structure elucidation (CASE) via electron impact mass spectra at this stage. This leaves CASE via MS significantly behind that of other analytical techniques such as NMR (see, for example, the recent review,[25] which details much better success rates using another software developed by ACD). However, incorporating additional information in candidate selection and developments in high resolution and tandem MS techniques opens up new windows to improve CASE via MS.[26] The alternative strategy for matching structures and spectra based on combinatorial structures rather than fragmentation prediction, implemented in the software FiD,[15] shows promising results for tandem MS data and should be investigated further.

Most of all we wish to caution the users of various programs for CASE to not just limit their calculations to one candidate molecule but to consider other possibilities, maintain their skepticism, avoid the use of poorly defined indicators, be aware of the large effect minor program settings can have, and finally evaluate their results carefully.

ACKNOWLEDGMENT

This manuscript was supported by the European Commission through the Integrated Project MODELKEY (Contract No. 511237-GOCE). We would like to thank Stan Schymanski, Peter Carsten von der Ohe, and the anonymous reviewers for reading and improving the manuscript. Further information about MOLGEN-MSF and contact details are available at www.molgen.de.

SUPPORTING INFORMATION AVAILABLE

Match value derivation, file format information, program settings and abbreviations, and additional results and discussion. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review December 22, 2008. Accepted March 6, 2009.

AC802715E

(25) Elyashberg, M. E.; Williams, A.; Martin, G. E. *Prog. Nucl. Magn. Reson. Spectrosc.* **2008**, *53*, 1–104.

(26) Schymanski, E. L.; Bataineh, M.; Goss, K.-U.; Brack, W. *TrAC, Trends Anal. Chem.* **2009**, doi:10.1016/j.trac.2009.03.001.