STRUCTURE VERIFICATION OF SMALL MOLECULES

H. THIELE,[1] G. MCLEOD,[2] I. KREBS[1]

[1]**Bruker Daltonik GmbH**
(28359 Bremen, Germany; e-mail: ht@bdal.de)

[2]**Bruker Daltonics Ltd.**
(Coventry, England)

PACS 31.10.Bi, 07.75.+h

The high mass accuracy of the MS and MS/MS data provided by the new generation of ESI-TOF- and ESI-Q-TOF mass spectrometers provides an accurate determination of molecular weight, which is used specifically for the structural verification and purity determination of substances. The high separation of the isotope profile for both MS and MS/MS spectra affords further dimensions of information to achieve precise molecular formula determination.

1. Introduction

In the service laboratories in the pharmacy industry today, chemists synthesize huge numbers of different compounds on the basis of parallel synthesis and combinatorial chemistry, which must be verified. One main aspect is the confirmation of chemical identity and information about molecular formulae to identify possibly present impurities. Most often, a screening approach is applied in order to obtain a rough estimation of the purity, concentration, and identity of the synthesized products. This points out samples below a certain purity or with too many impurities and samples that do not have the intended contents. It is often impossible to use all samples at the same time; therefore, most of the samples are stored in libraries for later use. Undertaking the quality control on a regular basis is mandatory in order to check for possible degradation reactions.

The main area of application for quality control and confirmation of molecular ID is synthetic chemistry (medicinal chemistry, core facility, organic chemistry, and pharmacy NCE). Another broad field of applications is the identification of small molecules in life sciences, such as metabolite ID and food and drug contaminations [1–8].

2. Methods

2.1. From mass spectrum to molecular formula

With high-resolution instruments, molecular formulae can be calculated directly from the MS spectrum [9]. Because mass spectra do not automatically convey elemental information, data analysis tools are necessary to extract the information inherent in MS spectra to provide molecular formula candidates. Software programs typically produce a list of potential candidates near the measured mass, calculate the expected exact isotope masses and isotopic intensity distribution, and compare them to the measured values. Mass accuracy is an essential parameter to limit the number of potential candidates.

Due to the high number of possible combinations, additional constraints for formula generation are used to restrict the number of solutions. Basic chemical knowledge can supply boundary constraints for formula generation [10–11]. Some of the constraints can be derived directly from 1D-NMR measurements. The integral of non-overlapping signals can give information about the number (min/max) of aliphatic hydrogen atoms and the number of olefinic and/or aromatic double bonds.

2.2. The unique ESI-TOF technology

Ideally, an instrument that can provide both high mass accuracy and stable, true isotopic pattern (TIP) information could provide greater information content. This technical concept allows for a two-dimensional analytical method: a combination of accurate mass determination with the analysis of the isotopic distribution. Combining complementary information is essential to find the correct formula for the elemental composition. Time-of-flight instruments are most often the best choice for molecular formula determination. This is especially true for Electrospray-Q-TOF-MS instruments (micrOTOF/-Q, maXis [Bruker Daltonik, Germany]) which use the linear ion counting to determine the known natural isotopic ratios.

Maximum certainty in small molecule identification requires cutting-edge performance from the MS instrument: A resolution of, typically, 15.000–20.000 at a high acquisition speed of 20 spectra/s is mandatory to cope with ultra fast chromatography systems. Mass resolution and mass accuracy have to be maintained

in all scan modes and speeds; in MS as well as in MS/MS. The outstanding dynamic range is related to the fast repetition rate of the TOF (5.000–20.000 Hz) and the adequate analogue-to-digital-conversion (ADC) technique. This allows the exact mass determination over the whole dynamic range as it is not compromised by dead time effects found in the more common time-to-digital-conversion (TDC).

The ADC technique combined with high-resolution TOF-MS is also important for the accuracy of the relative intensities of isotopic peaks. This accuracy is required for the success of the true isotopic pattern (TIP) matching strategy. This can be a severe problem for TDC-based TOF-MS, because the intensity of isotopes following high abundant isotopes is often reduced by the dead time of a detector. In MS/MS spectra generated with the micrOTOF-Q II [Bruker Daltonik, Germany], the isotopic pattern information and the accuracy are also retained in the fragment ions. Sum formula proposals can be made for the fragment ions in the same way as for MS spectra, which add a third level of confidence [11].

2.3. Isotopic pattern analysis – scoring of formula candidates

Mass accuracy is not enough to reduce the number of possible hits in molecular formula generation [10]. The SmartFormula^{TM} [SW delivered by Bruker Daltonik, Germany] approach considers the isotopic pattern distribution for MS spectra. After the generation of a list of all possible formulae for a window around a selected mass of an LC-MS peak, the measured isotopic pattern is compared with the theoretical isotopic pattern – resulting in a similarity measure, sigma (σ). This measure is simply the root-mean-square deviation between the normalized measured and theoretical isotopic intensity distributions. This comparison is done for all the generated molecular formulae. Then, the sigma value is used to rank the formula candidates.

2.4. Enhanced probability-based concept

In many cases, there are several molecular formulae with well-matched properties. It is not sufficient to consider only one hit with the best scoring factor, there are a few hits satisfying the overall criteria. The overall ranking of the formulae candidates has to be extended into a probability-based scoring concept, modeling the distributions of mass accuracy and true isotopic pattern matching for a number of possible candidates.

An isotopic pattern is described by three characteristic properties: the mass position of the peaks in the pattern; the peak intensities; and the peak distances within the pattern. It is possible to combine those values into a more informative score for the individual hits. This boils down to finding weights modeling the relative informative value of the different properties, which would depend on the accuracies, precisions, and resolutions of masses and intensities of the MS instrument. Such a score can be used to rank the list of hypothetical formulas based on a meaningful quality criterion. Considering all of the generated formula candidates using a Bayesian statistical modeling of the deviations, a score value with a range of 0 to 100 can be derived.

2.5. Precision in formula generation: true isotopic pattern of fragments

Examination of accurate mass and the isotopic pattern is not sufficient to point out the correct formula. Further fragmentation of the ions of interest produces valuable structural information which can be exploited to narrow down the list of sum formulae to a few ones and to elucidate the underlying structure of the neutral species. Therefore, several techniques have been developed, which use the information from MS/MS spectra to create a more confident possible formulae for the precursor ion. These methods sum up the potential formulae for the product ion and the neutral loss to establish the identity of the precursor ion.

In contrast to these algorithms, a new approach – SmartFormula3D^{TM} – [SW delivered by Bruker Daltonik, Germany] utilizes accurate measured mass and additionally accurate measured isotopic pattern. It generates a confident list of formulae simultaneously for the precursor ion and all fragment ions. All formulae for fragment ions are assigned to precursor ion formulae if these formulae are a true subset of the precursor ion formula candidate and the difference formula, which reflects the neutral loss molecule, is chemically meaningful and the deviation of its mass and the observed difference of the signals in the spectrum for the precursor ion and the fragment ion are within a predefined window. Fragment ion formulae which cannot be related to any precursor ion formula or precursor ion formulae which cannot be related to any fragment ion candidate will be eliminated.

In contrast to the already known algorithms, both candidate lists are already drastically pruned, thus reducing the time to evaluate all possible relationships between the potential precursor formulae and the related product ions.

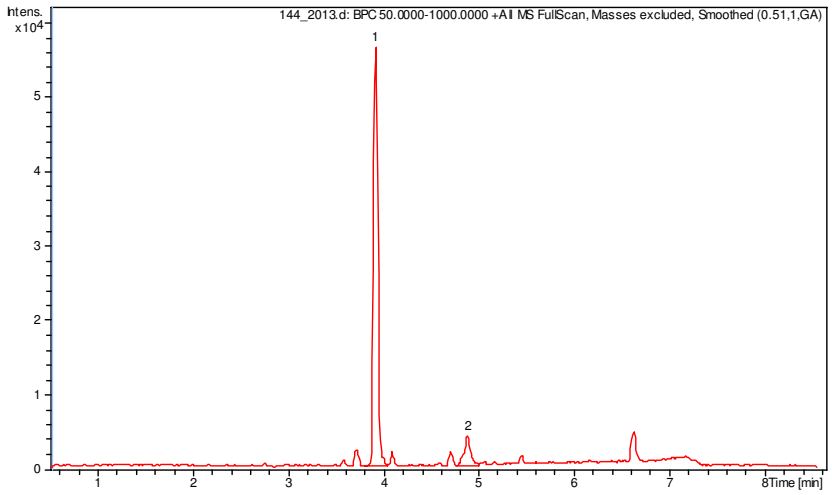


Fig. 1. The LC-MS base peak chromatogram of sample 144 shows a single major peak

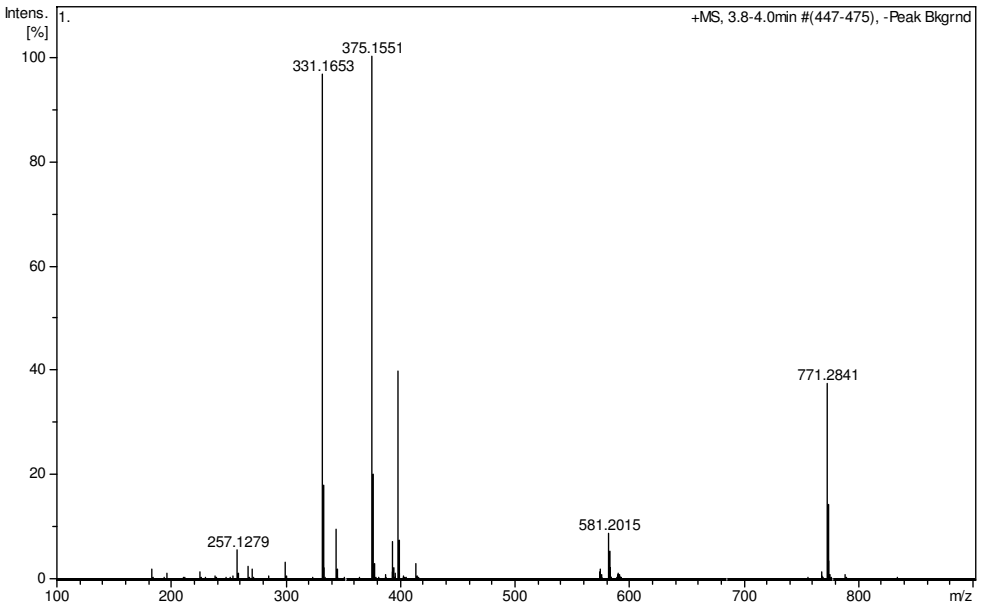


Fig. 2. Full scan TOF-MS spectrum shows two major $[M+H]^+$ ions, sodium adducts, and sodiated dimer

2.6. Results: molecular formula determination

2.6.1. LC-MS Results

For one of the analyzed samples – Ethyl 3.5 – dimethyl-4-(1-benzyloxy-3-oxo-propyl)-1H-pyrole-2-carboxylate (referred as sample 144), an example of an LC-MS chromatogram is shown in Fig. 1. For sample ID144, a single component accounted for approximately 90% of the detected material. The mass spectrum in Fig. 2 shows the mass measurement of the component of interest. Table 1 shows the MS information derived from the spectrum.

For each ion – the pseudomolecular ion, dimer, and adducts – all of the possible formulae within 1 mDa error and acceptable isotopic pattern fit score ("Sigma" value) were returned in table format. The theoretical mass of the proposed ions is shown, along with the measurement errors (in ppm and mDa) and the isotopic fit value (mSigma). A score value is also returned, based on combined isotopic fit and mass, normalized to 100. For the majority of the ions, there is only one possibility for each ion, so the score is simply 100. For the higher mass sodiated dimer at m/z

T a b l e 1. Molecular formula results generated by SmartFormula

Meas. m/z	N	Formula	Score	m/z	err [mDa]	err [ppm]	mSigma	rdb	ei Conf	N-Rule
331.1653	1	$C_{18}H_{23}N_2O_4$	100	331.1652	−0.1	−0.2	10.4	8.5	even	ok
343.1284	1	$C_{17}H_{16}N_6NaO$	100	343.1278	−0.6	−1.8	3.7	12.5	even	ok
375.1551	1	$C_{19}H_{23}N_2O_6$	100	375.1551	0.0	0.0	8.1	9.5	even	ok
397.1368	1	$C_{19}H_{22}N_2NaO_6$	100	397.1370	0.2	0.6	14.6	9.5	even	ok
771.2841	1	$C_{37}H_{48}$ NaO_{16}	100	771.2835	−0.7	−0.9	20.3	13.5	even	ok
	2	$C_{38}H_{44}N_4$ NaO_{12}	79.6	771.2848	0.7	0.9	30.6	18.5	even	ok

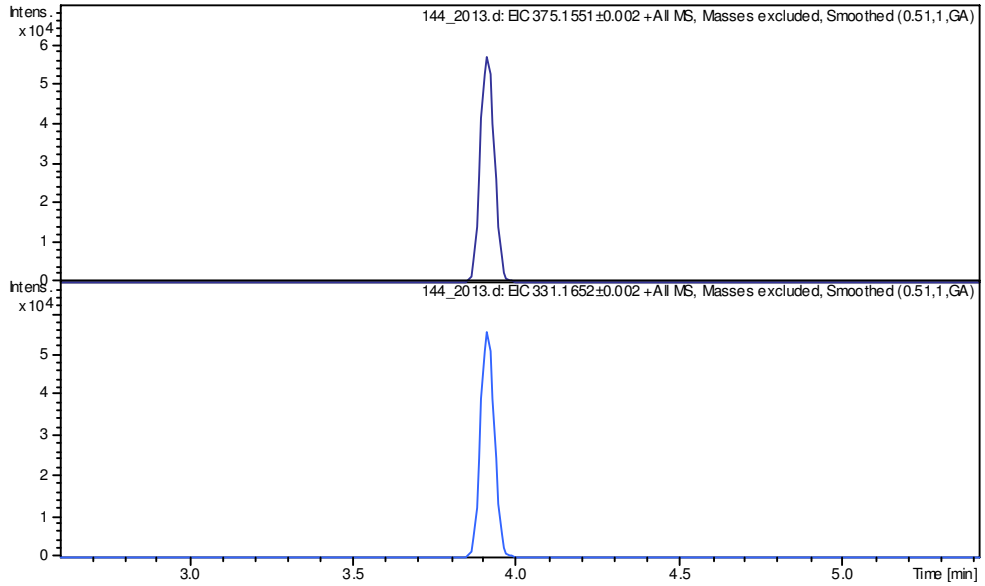


Fig. 3. Extracted ion chromatograms of the two peaks (using window +/− 0.002 Da) show perfect co-elution, suggesting that the lighter ion is an in-source fragment formed by loss of CO_2

771.2841, there are two possibilities with scores 100 and 79.6. The score of 79.6 is also very good because both of these ions have acceptable mass measurement and mSigma fits, so it is not possible to determine on that basis alone which is correct. However, it is clear that $C_{38}H_{44}N_4NO_{12}$ is a dimer of one of the other ions, so that assignment can be made with confidence.

This particular spectrum is slightly complicated by the presence of two apparent $[M+H]^+$ ions. Due to a very high stability of the accurate mass measurement of the sample, it is possible to create extracted ion chromatograms of each ion with very fine tolerance. Figure 3 shows how high-resolution extracted ion chromatogram (hrEIC) traces using a window of +/− 0.002 Da indicate perfect co-elution of the ions. Coupled with a difference in the formula of CO_2, this shows that the lighter ion was created by elimination of neutral carbon dioxide in source. Therefore, the formula of the actual molecule of interest must be $C_{19}H_{22}N_2O_6$.

2.6.2. LC-MS/MS Results

Further structural information on sample ID144 was derived from the MS/MS results. Accurate-mass spectra of the compound of interest were obtained from the Auto-MS/MS experiment. The isotopic pattern information can also be obtained for the fragment ions by setting an appropriate isolation window. Figure 4 shows the MS/MS spectrum of the precursor m/z 331 with accurate mass measurements. Figure 5 shows the MS/MS spectrum of m/z 375. Tables 2 and 3 show the assignment of the masses of the ions of interest from the MS/MS spectra of m/z 331 and 375, respectively, again using a 1-mDa window and isotopic pattern fit threshold.

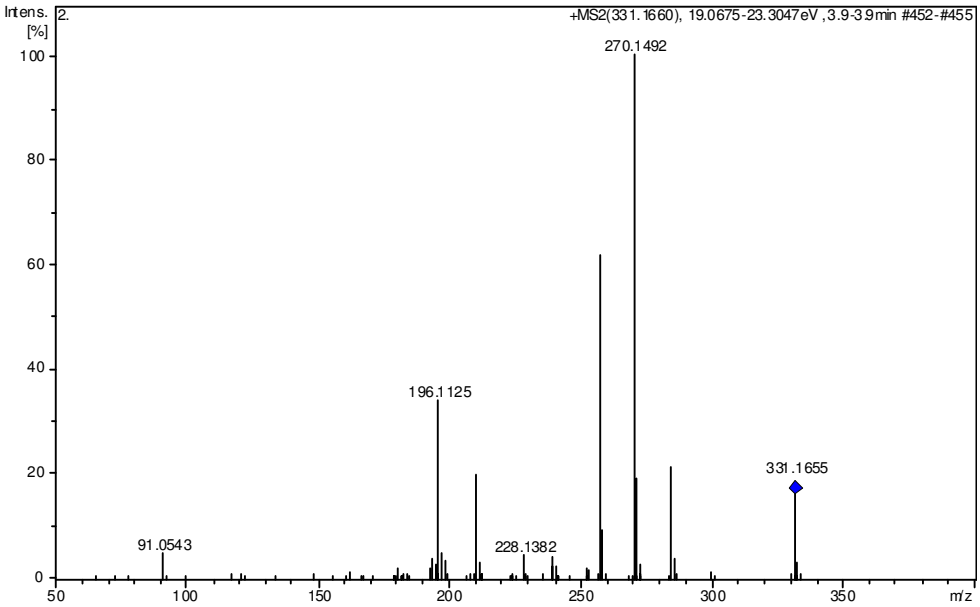


Fig. 4. MS/MS spectrum of m/z 331. Formulae of fragments below

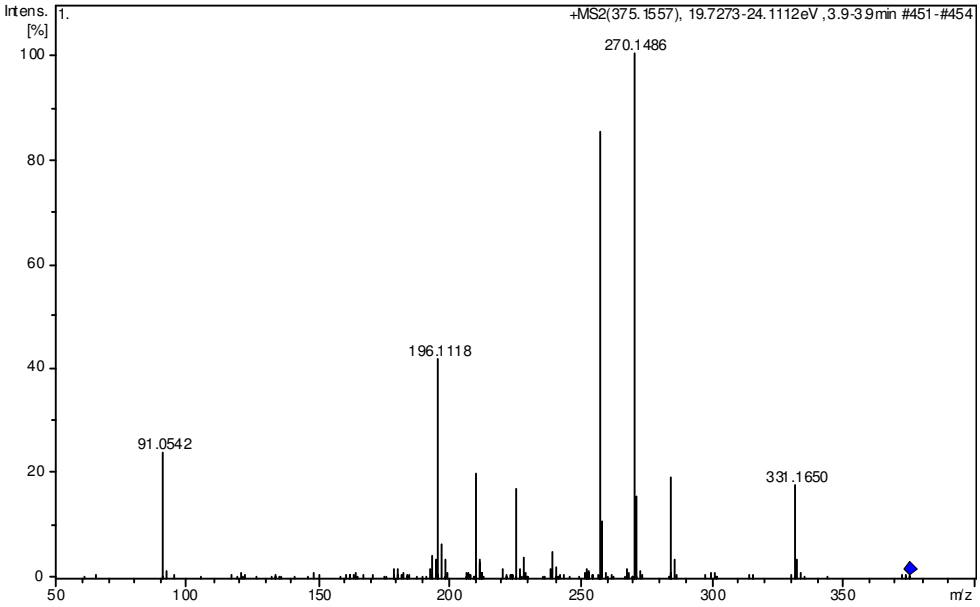


Fig. 5. MS/MS spectrum of m/z 375 produces the same fragments as the previous spectrum via the m/z 331

The two product ion spectra are essentially from the same compound, via the elimination of carbon dioxide as explained above. Figure 6 shows the central region of the spectrum side-by-side for the two experiments, demonstrating that the results are, in fact, the same.

All of the product ions in the spectrum were assigned to individual formulae based on the same mass measurement and isotopic pattern thresholds, again using the SmartFormula^{TM} software tool. For compounds where there is ambiguity about the empirical formula, that is, where there is more than one reasonable possibility proposed for the precursor ion by SmartFormula^{TM}, the SmartFormula3D^{TM} software tool would be employed, in which those precursor and product ion formulae that could not logically combine to form one another would be filtered out. SmartFormula3D uses accurate mass anal-

T a b l e 2. Formulae of fragments for MS/MS spectrum of m/z = 331

Meas. m/z fragments of m/z 331	N	Formula	Score	m/z	err [mDa]	err [ppm]	mSigma	rdb	ei Conf	N-Rule
91.0543	1	C_7H_7	100	91.0542	−0.1	−0.7	0.6	4.5	even	ok
196.1125	1	$C_{14}H_{14}N$	100	196.1121	−0.4	−1.9	9.0	8.5	even	ok
210.1281	1	$C_{15}H_{16}N$	100	210.1277	−0.3	−1.6	9.1	8.5	even	ok
228.1382	1	$C_{15}H_{18}NO$	100	228.1383	0.1	0.4	13.0	7.5	even	ok
257.1289	1	$C_{15}H_{17}N_2O_2$	100	257.1285	−0.4	−1.6	12.5	8.5	even	ok
270.1492	1	$C_{17}H_{20}NO_2$	100	270.1489	−0.4	−1.3	2.6	8.5	even	ok
284.1651	1	$C_{18}H_{22}NO_2$	100	284.1645	−0.6	−2.1	14.2	8.5	even	ok
331.1655	1	$C_{18}H_{23}N_2O_4$	100	331.1652	−0.2	−0.7	20.8	8.5	even	ok

T a b l e 3. Formulae of fragments

Meas. m/z fragments of m/z 375	N	Formula	Score	m/z	err [mDa]	err [ppm]	mSigma	rdb	ei Conf	N-Rule
91.0542	1	C_7H_7	100	91.0542	0.0	0.3	9.9	4.5	even	ok
196.1118	1	$C_{14}H_{14}N$	100	196.1121	0.3	1.6	9.6	8.5	even	ok
210.1272	1	$C_{15}H_{16}N$	100	210.1277	0.5	2.5	23.4	8.5	even	ok
225.0865	1	$C_{10}H_{13}N_2O_4$	100	225.0870	0.5	2.4	11.2	5.5	even	ok
257.1283	1	$C_{15}H_{17}N_2O_2$	100	257.1285	0.1	0.4	24.4	8.5	even	ok
270.1486	1	$C_{17}H_{20}NO_2$	100	270.1489	0.2	0.8	18.8	8.5	even	ok
284.1640	1	$C_{18}H_{22}NO_2$	100	284.1645	0.5	1.9	19.8	8.5	even	ok
331.1650	1	$C_{18}H_{23}N_2O_4$	100	331.1652	0.2	0.7	16.9	8.5	even	ok

ysis, fragmentation data, and isotope pattern (in both MS and MS/MS) to identify the molecular formula of the parent compound.

In this example, to start with, a generous window of 3 ppm was applied, and the mSigma (isotope pattern score) threshold is set to 1000 (max) so it is ignored (all results will be shown by mass accuracy alone). Note these initial parameters were selected for example purposes only. Normally, the mSigma value would be set to exclude poor isotopic fits. Filtering was applied using the H/C ratio and estimated carbon number, and a RDB limit was applied (restricts to plausible organic ions). Even-electron species only were allowed for the precursor, but the possibility of radical product ions was allowed.

To show the power of using MS/MS data, a wide range of elements was allowed: unlimited C, H, N, O plus up to 3S, 2P, 2Br, 2I, 2Cl and 3F, in any combination. So no assumptions are being made about the chemistry: results will be entirely objective.

2.6.3. SmartFormula3D™ Result window

The software assigns ions (highlighted purple) and their isotopes (highlighted green) automatically. Based on MS data and matching the minimum *one* fragment, there are 20 possibilities for the formula of the precursor ion (Fig. 7). Normally, mSigma filtering alone would remove the species with Br, Cl or S from this list, as they have poor isotopic fit, leaving ten possibilities based on mass and isotope pattern. Note the "Peaks explained" column. The highlighted precursor formula can make a logical fit with 10 product ions; many of the others can explain only a handful of the fragments. Clicking on that promising candidate formula shows the proposed fragment formulae and neutral losses.

If another proposal is selected, C16H23F3N4OP, some proposed fragment formulae for that ion are displayed. However, the ion 331 in MS/MS has changed to a peach colour. This means that it is a fragment ion *not* explained by the selected precursor – i.e., there is no formula that fits that product ion which could be created from that proposed precursor ion formula – but it *is* explained by another formula in the list.

The results should be calculated using the full MS/MS filtering tools that are available: Set a sensible mSigma threshold; state that the top 8 MS/MS peaks must be explained; limit the mass error permitted for neutral losses to 2 mDa. The result shown in Fig. 9 is that

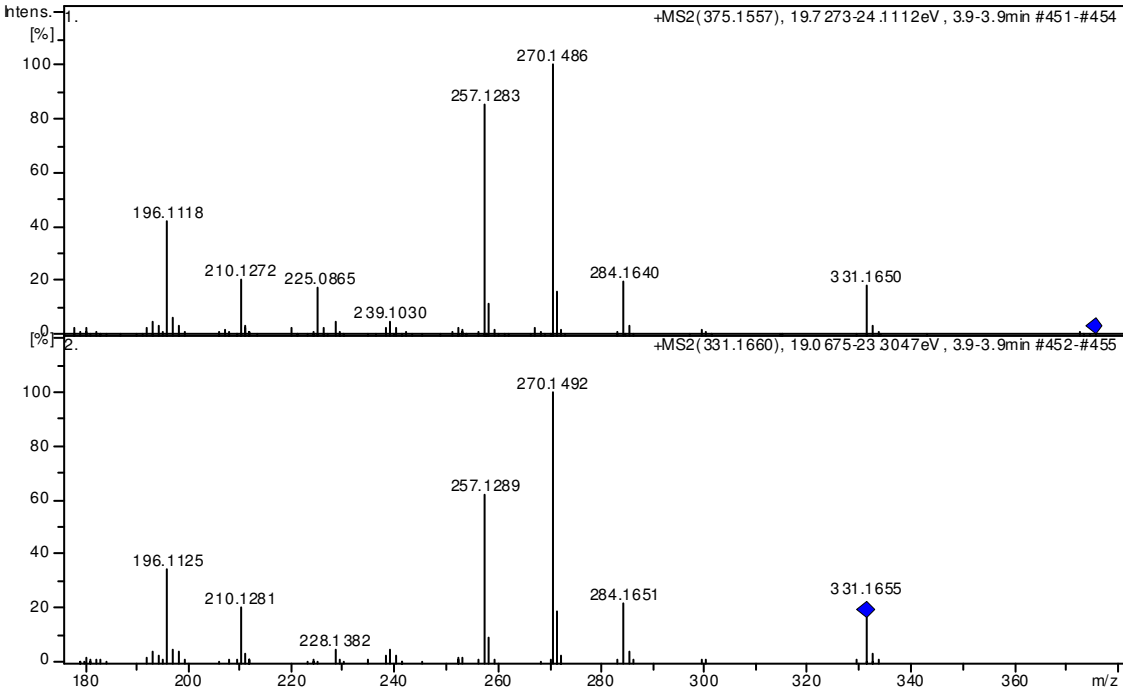


Fig. 6. The two MS/MS spectra, zoom into the detail of central region

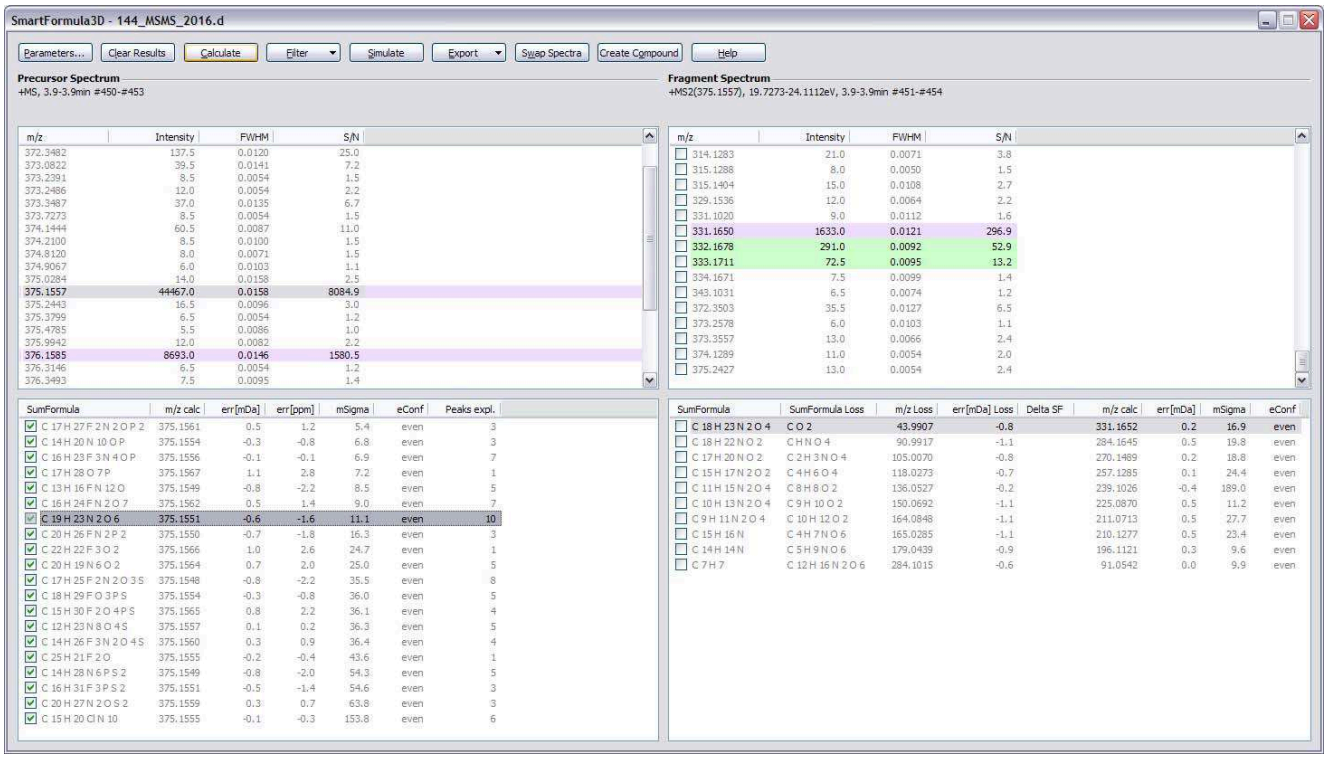


Fig. 7. SmartFormula3D results screen, allowing elements C, H, N, O, S, P, F, Br, Cl and minimum of one fragment ion match, 1 mDa mass tolerance and no isotope fit (mSigma) threshold applied. Top left: peak list, MS spectrum; Top right: peak list, MS/MS spectrum; bottom left: proposed precursor ion formulae; bottom right: proposed product ion formulae for the highlighted precursor ion candidate

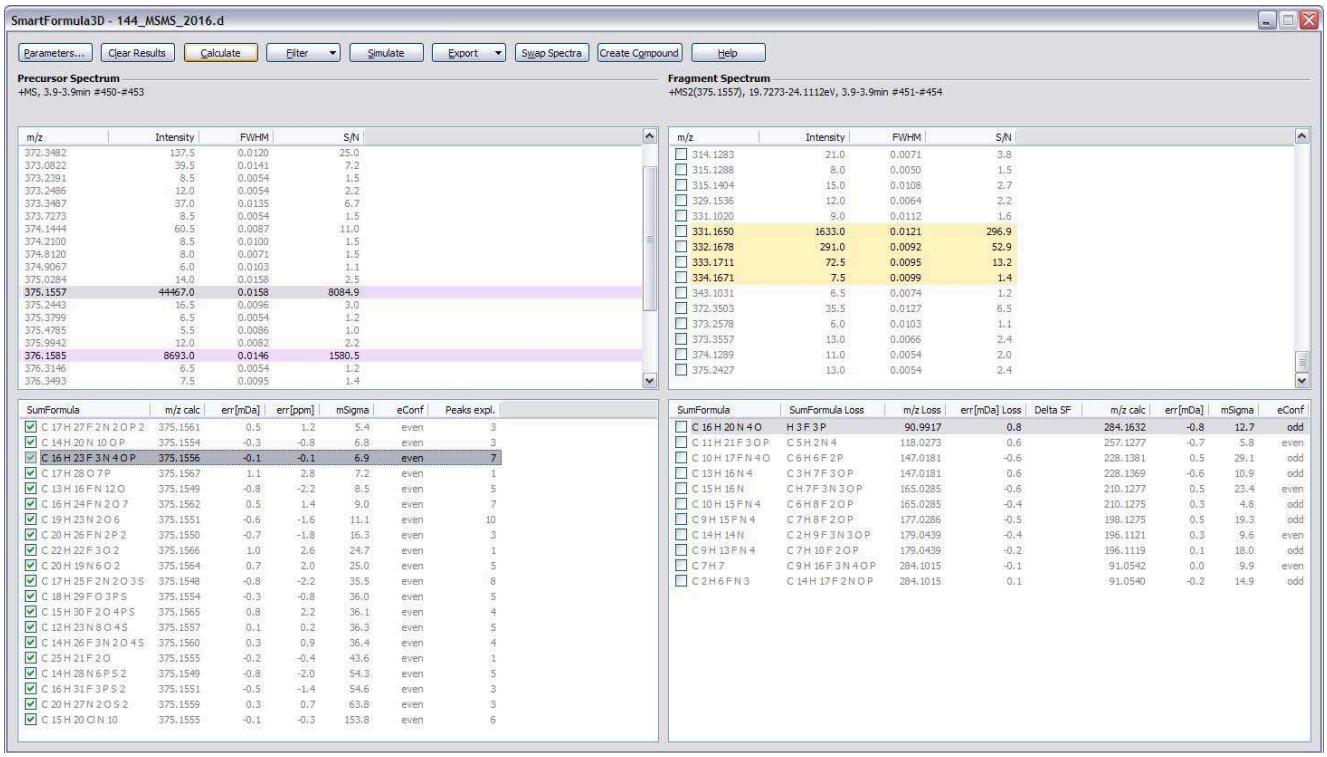


Fig. 8. The same data as for Fig. 7, but selecting a precursor candidate which cannot explain all fragments. The peach colour highlighting in the MS/MS mass list (top right) immediately shows an ion which cannot be explained by starting from the highlighted precursor. Therefore such precursor candidates can be filtered out

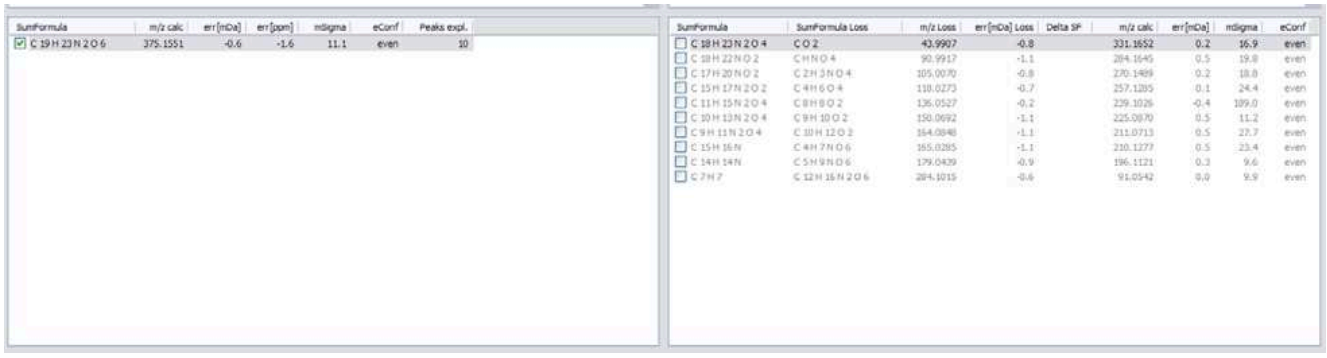
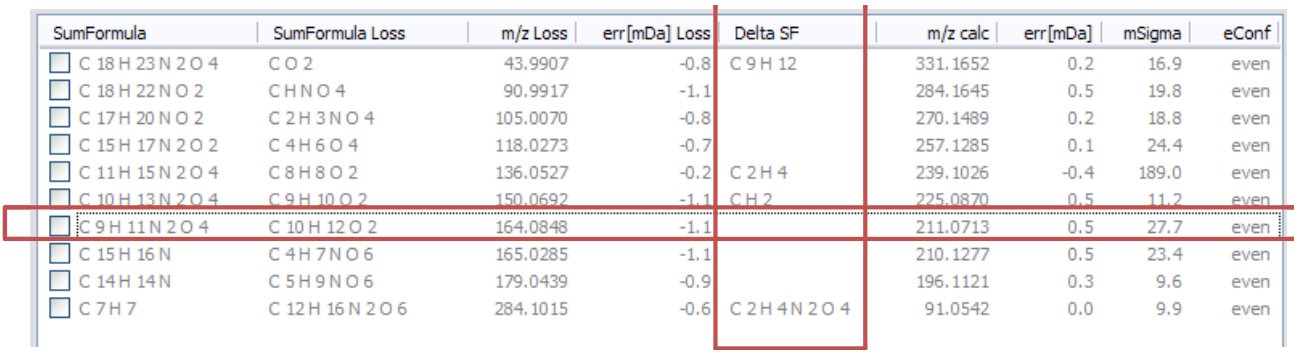


Fig. 9. Fully processed result from SmartFormula3D, filtered based on mass accuracy, isotopic fit and the correlation of the major fragment ions, showing only one possible formula for the precursor ion and product ions

the full data set is only explicable by one precursor ion candidate. Therefore, a definitive formula for the precursor of $C_{19}H_{22}N_2O_6$ is determined without needing to make any assumptions about the sample chemistry.

In addition, a full list of the product ion formulae that can be produced by this identified precursor has been obtained, plus their neutral losses. The software includes a further facility to filter the data based upon an expected neutral loss or fragment – this is useful if the compound is expected to be from a particular class with characteristic fragments.

If a formula in the product ions is selected, then the relationship between neighboring ions is shown by neutral differences in the "Delta Sum Formula" column – so possible MS/MS routes are highlighted (Figure 10).



SumFormula	SumFormula Loss	m/z Loss	err[mDa] Loss	Delta SF	m/z calc	err[mDa]	mSigma	eConf
C 18 H 23 N 2 O 4	C O 2	43.9907	-0.8	C 9 H 12	331.1652	0.2	16.9	even
C 18 H 22 N O 2	C H N O 4	90.9917	-1.1		284.1645	0.5	19.8	even
C 17 H 20 N O 2	C 2 H 3 N O 4	105.0070	-0.8		270.1489	0.2	18.8	even
C 15 H 17 N 2 O 2	C 4 H 6 O 4	118.0273	-0.7		257.1285	0.1	24.4	even
C 11 H 15 N 2 O 4	C 8 H 8 O 2	136.0527	-0.2	C 2 H 4	239.1026	-0.4	189.0	even
C 10 H 13 N 2 O 4	C 9 H 10 O 2	150.0692	-1.1	C H 2	225.0870	0.5	11.2	even
C 9 H 11 N 2 O 4	C 10 H 12 O 2	164.0848	-1.1		211.0713	0.5	27.7	even
C 15 H 16 N	C 4 H 7 N O 6	165.0285	-1.1		210.1277	0.5	23.4	even
C 14 H 14 N	C 5 H 9 N O 6	179.0439	-0.9		196.1121	0.3	9.6	even
C 7 H 7	C 12 H 16 N 2 O 6	284.1015	-0.6	C 2 H 4 N 2 O 4	91.0542	0.0	9.9	even

Fig. 10. Selecting a product ion formula automatically calculates the difference between that formula and others in the list, allowing the determination of fragment relationship

3. Experimental

3.1. LC-MS analysis

3.1.1. Sample preparation:

The samples were prepared for LC-MS analysis by adding 100 μl acetonitrile added to each dried sample well. The resulting sample solutions were diluted 1:200 by adding a 5 μl aliquot to 995 μl of 1:1 (v/v) water:acetonitrile. These samples were used for the LC-MS analysis.

3.1.2. LC-MS Method:

Mass spectrometry was performed using a maXis Ultra-High Resolution Time-of-Flight mass spectrometer (UHR-TOF) with Qq-TOF geometry [Bruker Daltonik, Germany]. Data were acquired over the mass range 50–1000 Da with an acquisition rate 2 spectra/second. All data were acquired using an automatic external calibration routine.

MS/MS spectra were acquired in Auto-MS/MS mode (data-dependent acquisition). Collision energy was estimated dynamically based on appropriate values for the mass and stepped across a +/– 10% magnitude range to ensure good quality fragmentation spectra.

Samples were introduced as 5 μl injections on a Ultimate3000RS UHPLC system [Dionex] with a Phenomemex Luna-HST C18(2) 2.5 μm 5 cm × 2.0 mm HPLC column. A gradient elution using solvent A: water + 0.2% formic (v/v) acid; solvent B: Acetonitrile + 0.2% formic acid (v/v) was applied as follows: start 20% B for 0.2 minutes; ramp to 100% B for 3.5 min; hold at 100% B for 2 min; return to starting conditions in 0.1 min and re-equilibrate for 2.5 min. The flow rate was 0.3 ml min^{-1} and the column was maintained at 40 °C. All solvents were LC-MS Chromasolv grade (Sigma-Aldrich Ltd., Poole, UK).

The resulting LC-MS data were processed in an automated software routine within the Compass DataAnalysis software [Bruker Daltonik, Germany] as follows: a base peak chromatogram was produced and integrated. Then the spectrum of each detected peak (relative threshold 5% peak area) was extracted. The formulae of all significant detected ions were automatically assigned using the SmartFormula tool. Based on the expected chemistry, elements carbon, hydrogen, oxygen, nitrogen, bromine, and iodine were permitted. Sodium was also included for the calculation of adduct masses. The number of nitrogens was limited to an upper threshold of ten. The number of rings plus double bonds was checked to be chemically meaningful (between 0 and 50). The nitrogen rule was enforced. Even-electron species only were permitted. A window of 1 mDa mass accuracy was allowed. A threshold was applied to the goodness of fit of the isotope pattern, where lower values of the calculated term Sigma indicate smaller differences between the patterns of the measured and proposed empirical formulae (set value mSigma value < 50).

We are grateful to Prof. Dr. F.-P. Montforts and Dr. M. Osmers (Institute of Organic Chemistry, University Bremen) for supplying the samples and fruitful discussions. We are grateful to Dr. Don Richards (Pfizer/UK) for his tremendous contribution to the concept of SmartFormula3D$^{\text{TM}}$.

1. S. Lacorte and A.R. Fernandez-Alba, Mass. Spec. Rev. (2006) Doi:10.1002/mas.200944.
2. S. Böcker, M.C. Letyel, Z. Lipták, and A. Pervukhin, Bioinform. **25**, 218 (2009).
3. S.G. Roussis and R. Proulx, Anal. Chem. **75**, 1470 (2003).

4. S. Ojanperä, A. Pelander, M. Pelzing, I. Krebs, E. Vuori, and I. Ojanperä, Rapid Commun. Mass. Spectr. **20**, 1161 (2006).
5. F. Moriya and Y. Hashimoto, Forensic Sci. Int. **131**, 108 (2003).
6. E. Liotta, R. Gottardo, A. Bertaso, and A. Polettini, J. Mass. Spectr. **45**, 261 (2009).
7. A.L. Rockwood, S.L. Van Orden, and R.D. Smith, Anal. Chem. **68**, 2027 (1996).
8. A.L. Rockwood, Haimi Perttu, J. Am. Soc. Mass. Spectr. **17**, 415 (2006).
9. A. Kaufmann, Rapid Commun. Mass. Spectr. **24**, 2035 (2010).
10. T. Kind and O. Fiehn, BMC Bioinform. **7**, 234 (2006).
11. T. Kind and O. Fien, BMC Bioinform. **8**, 105 (2007).
12. A.L. Rockwood, M.M. Kushnir, and G.J. Nelson, J. Am. Soc. Mass. Spectr. **14**, 311 (2003).

Received 21.01.11

ВИЗНАЧЕННЯ СТРУКТУРИ МАЛИХ МОЛЕКУЛ

Г. Тіле, Г. Маклеод, І. Кребс

Р е з ю м е

Висока точність мас-спектрометричних даних від нової генерації ESI-TOF- та ESI-Q-TOF мас-спектрометрів дозволяє знайти молекулярні ваги. Це застосовується для перевірки структури і чистоти речовини. Високе розділення ізотопічних профілей у спектрах дає нові можливості для визначення точних молекулярних формул.