Superior Molecular Formula Generation from Accurate-Mass Data

Authors

Ed Darland, Doug McIntyre, David Weil, Frank Kuhlmann, and Xiangdong Li
Agilent Technologies

Abstract

When using mass spectrometry to analyze samples containing unknowns, it is often necessary to derive elemental compositions (molecular formulas) for the unknowns based on the mass spectral data. For lower-mass compounds, accurate-mass measurements may be sufficient to produce a single, conclusive molecular formula or a small number of possible formulas. With increasing mass, however, the number of possible formulas increases dramatically. Additional constraints must be applied to reduce the list of candidate formulas to a manageable number. Agilent MassHunter Workstation software includes proprietary molecular formula generation (MFG) software. This software takes full advantage of the mass accuracy of the data, and then uses additional mass spectral information to logically narrow the list of possible formulas. It reduces ambiguity and delivers a list of candidate molecular formulas with scores based on the relative probability that each formula is the correct one. This significantly reduces data interpretation time and increases the value of accurate-mass analysis.

The challenge of molecular formula generation

Molecular formula generation from MS data is important for identification of endogenous and exogenous metabolites, drug impurities, synthesis products, pharmaceutical degradation products, and for many other applications. The task is difficult because the number of plausible elemental compositions grows very rapidly with increasing molecular mass, and with the number of possible elements. For example, at a mass of 509.978 u, with only C, H, N, O, S, Cl, and F allowed in the composition, there are more than 250 possible formulas within a ±2.5 ppm mass window. Such a large number of possibilities makes it very hard for researchers to ascertain the molecular formula using mass information alone.

A recent study[1] concluded that even 1-ppm mass accuracy is not, by itself, sufficient to narrow the list of possible molecular formulas. For molecules with higher masses, even expensive orbital trapping and Fourier transform mass spectrometers will fail to provide a concise list of possible formulas based solely on accurate mass. The same study, however, concluded that additional constraints such as isotope pattern matching can rule out most of the candidate formulas.

Mass accuracy defined

Mass accuracy refers to the difference between a measured mass and a theoretical or calculated mass. Mass accuracy (also called mass error) is typically expressed in parts-per-million (ppm). It is calculated as follows:

$$(m_{measured} - m_{calculated}) \times 10^6 / m_{calculated}$$

For example, if the theoretical mass is 500 u and the measured mass is 500.005 u, the mass error is:

$$(500.005 - 500) \times 10^6 / 500 = 10 \text{ ppm}$$

Overview of molecular formula generation

As noted previously, accurate-mass measurements are, by themselves, often insufficient to conclusively determine the elemental composition of unknown compounds. The unique Agilent molecular formula generation software, however, can incorporate additional information such as:

- Isotope masses, abundances, and spacing
- User knowledge of sample composition (allowed elements)
- Accuracy of the mass measurements
- Masses of MS/MS fragment ions and neutral losses (if available)

into its calculations (see Figure 1). By doing so, it can generate a much smaller and more relevant list of candidate molecular formulas. For each compound, the software scores the possible molecular formulas based on their relative probabilities and assigns each formula a relative rank.

MFG also works in concert with the MassHunter software's proprietary molecular feature extraction (MFE) algorithms. MFE does an outstanding job of identifying the relationship between multiple forms of a single compound (e.g. adducts). If multiple forms of the same compound are present, MFG combines their scores into a single "cross score."

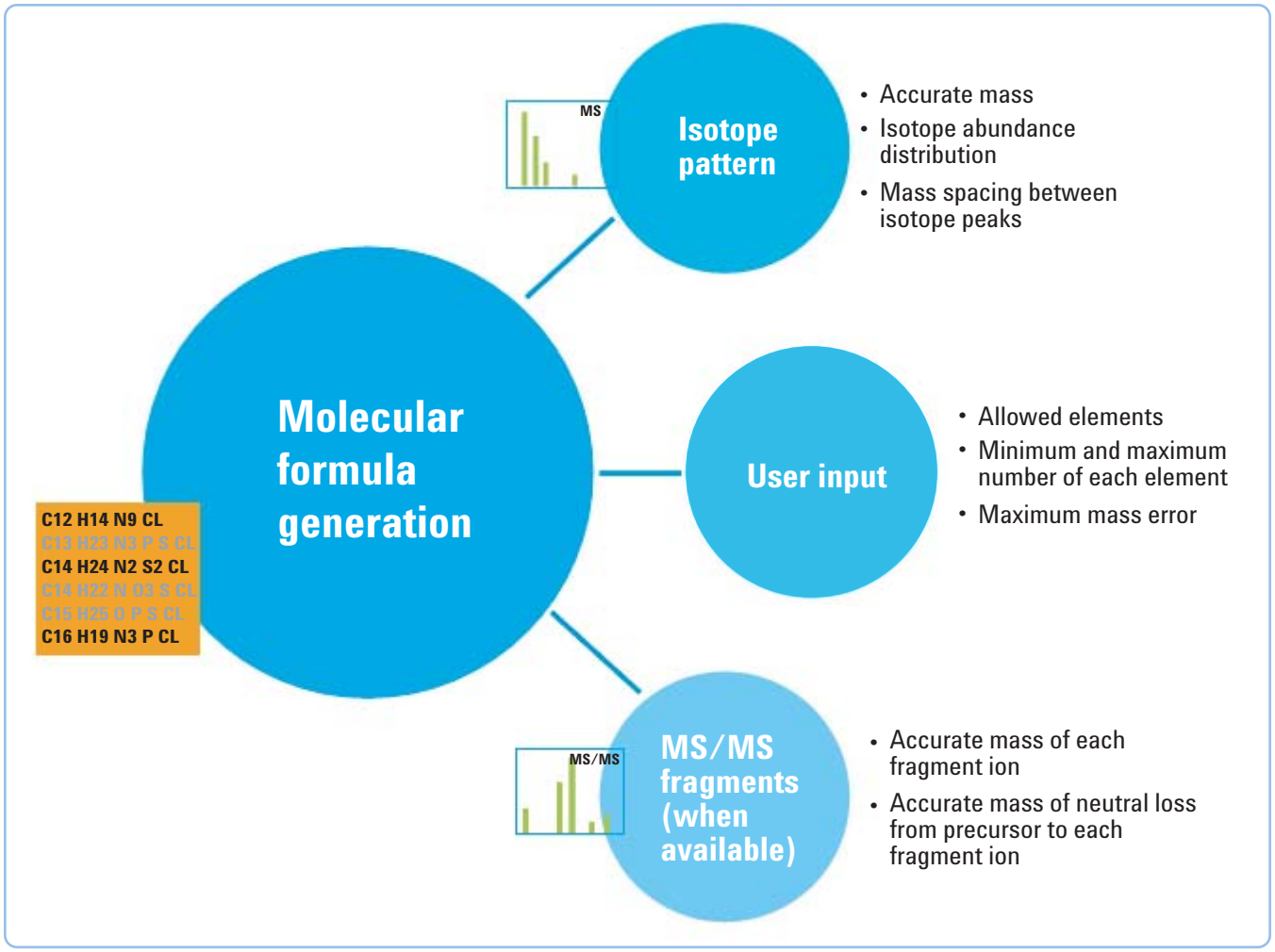

Figure 1. Agilent's molecular formula generation software uses multiple dimensions of information to generate and score lists of possible molecular formulas. It has been optimized for analysis of accurate-mass data from Agilent 6200 Series TOF and 6500 Series Q-TOF LC/MS systems.

By dramatically reducing the number of plausible formulas, and by using unique scoring methods, the MFG software reduces data interpretation time and increases the value of the accurate-mass analysis.

The MFG algorithms have been designed specifically to produce optimum results from the accurate-mass data generated by Agilent 6200 Series Time-of-Flight and 6500 Series Quadrupole Time-of-Flight LC/MS systems. Conversely, the superlative mass accuracy and consistency of the 6200 Series and 6500 Series data make it easier for the MFG algorithms to generate the best possible results.

MFG is applicable across a wide variety of applications that require identification of unknown compounds. Accordingly, Agilent has incorporated this sophisticated algorithm into the following Agilent software:

- MassHunter Workstation software (qualitative analysis module)
- MassHunter Metabolite ID software
- MassHunter Profiling software
- GeneSpring MS software

Processing of MS data

The MFG software starts by processing MS data, either from an MS-only analysis or from the precursor ions in MS spectra from an MS/MS analysis. After finding the mass, it calculates possible chemical compositions based on the accurate mass and isotope patterns, as well as any user-supplied constraints on the molecular formula. If molecular feature extraction (MFE) was used, the MFG software combines the scores of related features such as adduct ions to achieve a cross score.

Step 1: Determining the mass

Determining the mass is the most important step in obtaining a valid list of molecular formulas. If the wrong ion is chosen and the wrong mass assigned, none of the molecular formulas generated will be correct. Isotopic peaks and co-eluting species can complicate this process.

The MFG software examines clusters of ions. By applying common rules concerning isotopic abundance and the mass spacing of isotopes, it determines whether a group of ions are related and all part of the same isotope cluster, or whether two or more co-eluting compounds are present with their associated isotopes. The result is higher degree of confidence in the masses selected.

Beyond the obvious importance of mass accuracy in this process, mass resolution is also critical to identifying the correct masses. Without sufficient mass resolution, ions with very similar masses can be mistakenly identified as a single ion, leading to incorrect assignment of the mass and hence to incorrect molecular formulas. The enhanced mass resolution of the Agilent 6220 Accurate-Mass TOF and 6520 Accurate-Mass Q-TOF is of great benefit when determining masses.

Step 2: Applying user knowledge of the samples to limit possible formulas

Users can use their knowledge of the chemistry of their samples—for example, knowledge of which elements might logically be included and excluded—to limit the possible formulas. Figure 2 shows two tabs of the dialog box that permit users to specify constraints on the molecular formulas.

The Allowed Species tab lets users set the allowable elements, and the minimum and maximum number of each element. Elements can be added to the list which, by default, includes the elements most commonly found in organic compounds. A special syntax also permits naturally occurring isotopes (e.g. ^{13}C, ^{15}N) to be included. This is particularly useful for metabolomic studies using isotopic labeling. The Allowed Species tab also provides the option of applying the nitrogen rule[2,3] concerning even-electron and odd-electron species.

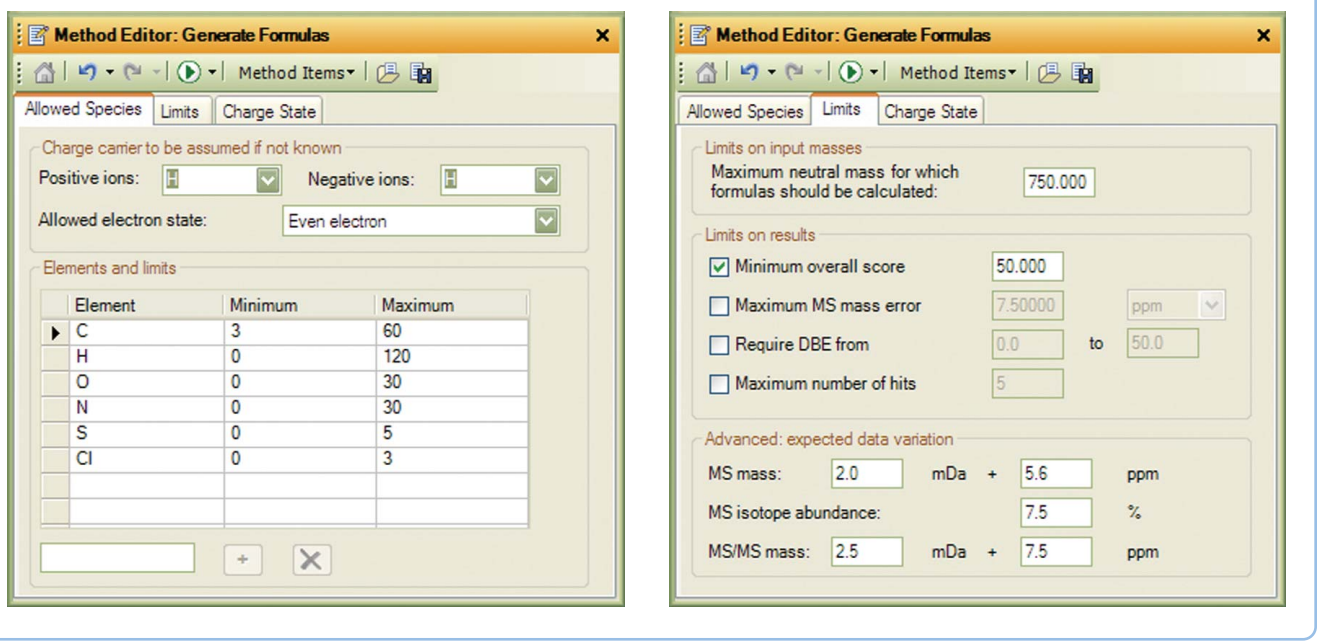

Figure 2. Users can eliminate unreasonable formulas by specifying constraints on the allowed elements, mass error, and other parameters.

The Limits tab allows users to specify the maximum mass error for their instrument, so that the software will reject chemical compositions that are outside of the mass error window. Users need to be familiar with the performance of their mass spectrometer to get the greatest benefit from this parameter. When an instrument provides consistent, excellent mass accuracy, the mass error window can be narrowed, significantly reducing the number of possible formulas generated. If an instrument has poor, or inconsistent mass accuracy, however, making the window too narrow could result in all of the generated formulas being incorrect. The excellent, consistent mass accuracy and consistency of the Agilent 6200 Series and 6500 Series instruments make it possible to set a narrower window than would be possible with some other instruments. The Limits tab also allows users to set limits on the number of double-bond equivalents that proposed formulas must have.

The MFG software uses all of this information to calculate all possible molecular formulas that meet user restrictions.

Step 3: Software rules out formulas that are chemically impossible

In addition to accepting user input, the MFG software has built-in logic to eliminate chemically implausible compositions. For example, it would rule out C_2H_8 because the ratio of carbon atoms to hydrogen atoms is impossibly low.

Step 4: Probability scores are calculated for the molecular formulas

At the MS level, MFG takes advantage of all available information from the molecular ion isotope cluster to calculate a probability score based on how well each of the generated molecular formulas fits the experimental data (Figure 3). To judge whether a proposed molecular formula is a good fit for the experimental data, the algorithm considers the following factors:

- How well does the candidate molecular formula explain the experimentally derived mass? In this calculation, the software incorporates the user-specified maximum mass error.
- How well do the calculated isotope abundance ratios for the candidate molecular formula match those of the experimental data?
- How well do the calculated mass intervals for the isotope peaks from the candidate formula match the experimental spectrum? This calculation also incorporates the user-specified mass error.

The use of the mass intervals between the peaks adds specificity to the generation of molecular formulas. When very-low-ppm mass accuracy is combined with this intelligent algorithm, the resulting list of formulas is shorter and the formula with the top score is more likely to be correct.

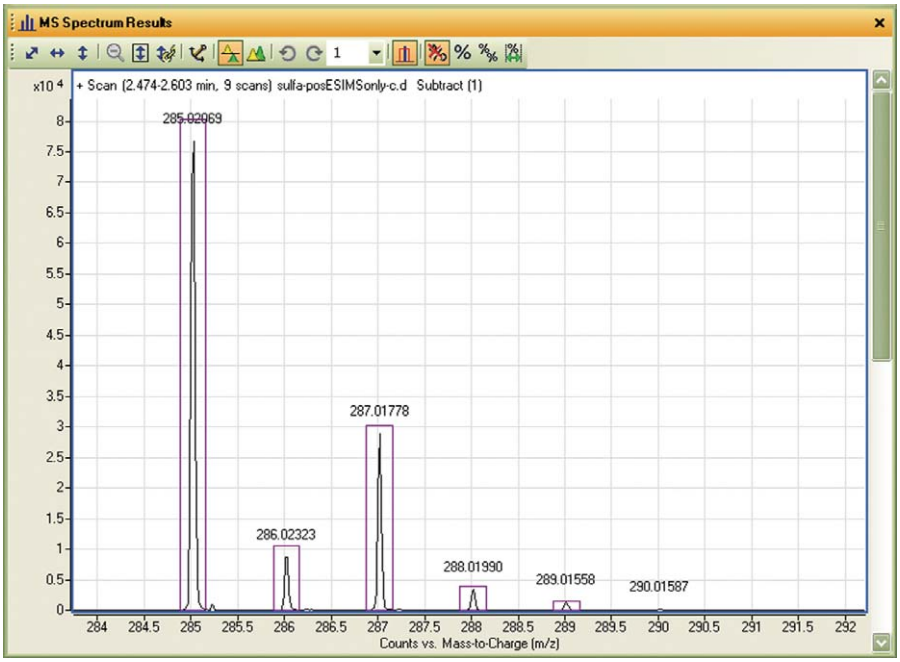

Figure 3. The molecular formula generation software uses isotope pattern matching to score tentative molecular formulas. The results show a theoretical isotope pattern (rectangles) overlaid on an experimental spectrum.

Step 5: Each of the candidate formulas is ranked based on its probability score

Finally, as shown in Figure 4, the MFG software assigns a relative rank to each candidate molecular formula based on is probability score. The software always assigns a score of 100 to the candidate elemental composition that best predicts the data. It calculates the rest of the scores as percentages of the best score, so lower scores represent less likely candidates. The relative scoring makes it easier to judge how significantly two predictions differ, and aids the user in selecting the correct molecular formula from the list of possibilities.

Molecular formula extraction

The molecular formula extraction (MFE) algorithm is a powerful compound-finding technique that locates individual sample components (molecular features), even when chromatograms are complex and compounds are not well resolved. Rather than designating compounds based exclusively on peak information, MFE locates ions that are covariant (rise and fall together in abundance) and that are logically related by charge-state envelope, isotopic distribution, and/or the presence of adducts and dimers. It assigns such ions to a single compound. Using this approach, the MFE algorithm can locate multiple compounds within a single chromatographic peak.

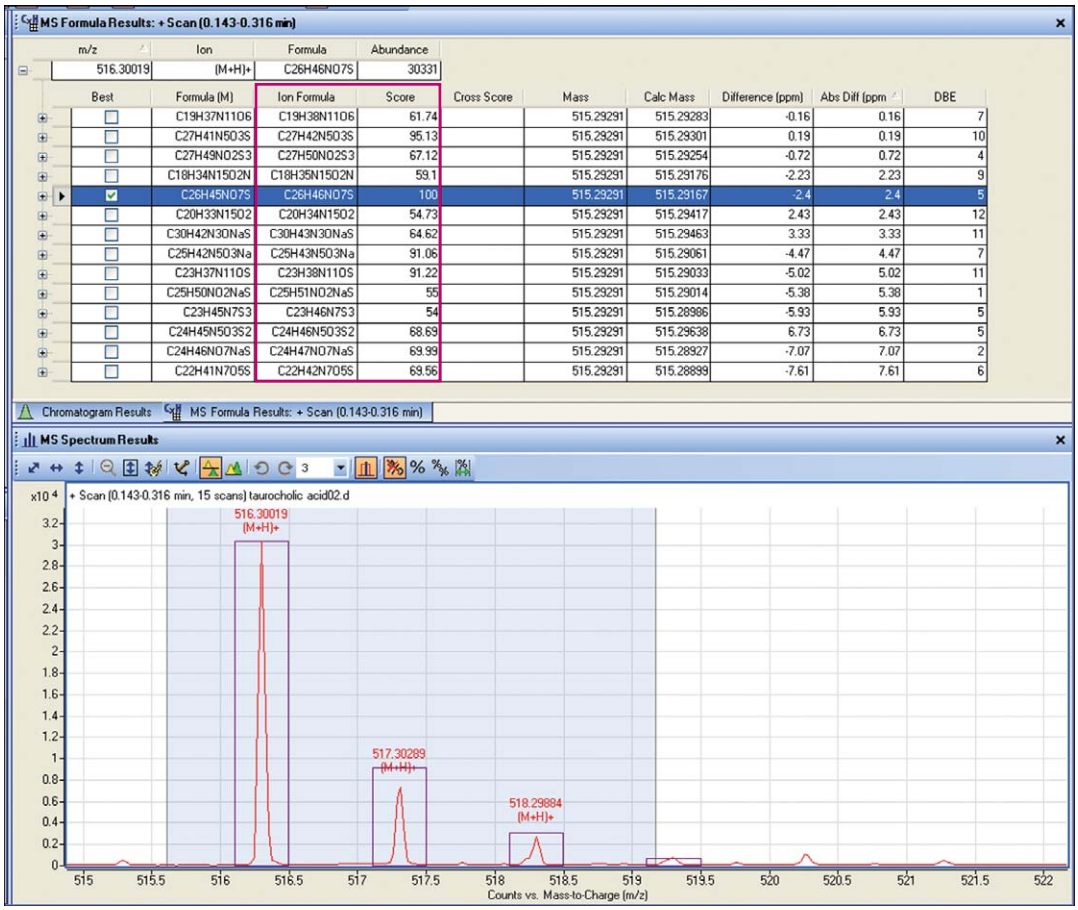

Figure 4. The final result of molecular formula generation for MS data is a ranked list of possible formulas, as shown in this example for taurocholic acid. Note that by mass error alone, the correct formula would have been fifth on the list, but with the MFG scoring technique, it is the top choice with a score of 100.

Unique cross score correlates data from multiple forms of the same compound

The MFG software uses a slightly different scoring system when it is used in conjunction with Agilent's proprietary molecular feature extraction (MFE) algorithm. The molecular feature extraction algorithm finds and correlates multiple species (ions) that are related to the same neutral molecule, (for example, ions representing multiple charge states or adducts of the same neutral molecule). When MFE information is available, the MFG software calculates an abundance-weighted, combined cross-species score for each molecular formula (Figure 5). This increases the accuracy of the MFG results.

The cross score is still a relative score that expresses how well the candidate molecular formula fits the experimental data. However, it is not scaled so that the top score is 100. The top score is 100 only if the same formula receives a score of 100 for each of the related species.

Excellent performance of the MFG scoring system for MS data

The MFG scoring system was tested by analyzing a 300-component pesticide mix at the 100-ppb level on the 6520 accurate-mass Q-TOF. By looking at the MS data only, MFE automatically found 98% of the manually detectable compounds. MFG found the correct molecular formula for 98% of the found compounds. 72% of the time, the correct formula was the one with the highest rank.

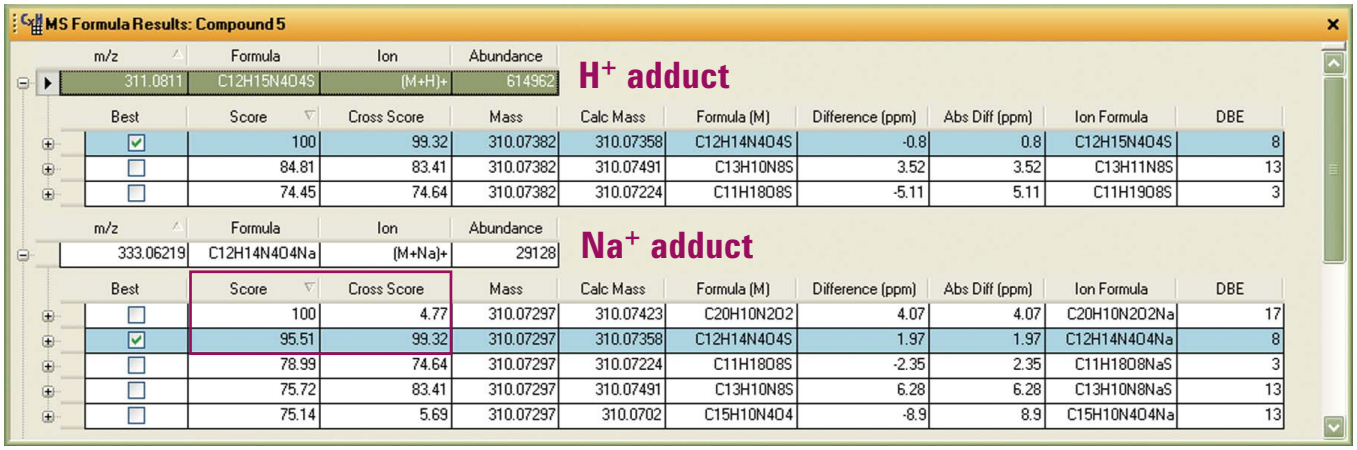

Figure 5. A cross score is a unique Agilent capability that combines results when the analysis shows multiple ions that are all related to the same neutral molecule. Note that for $[M+Na]^+$, the formula with the top MS score is not the correct answer, but the cross score correctly picks the overall best result.

Analysis of MS/MS data increases specificity and confidence

When MS/MS spectra are available, MFG can incorporate the MS/MS information to provide even greater confidence in the molecular formula generated for the precursor ion. With MS/MS data, MFG considers both the accurate mass of each fragment ion and the accurate mass of the difference (neutral loss) between the precursor and each MS/MS fragment.

Step 1: Software proposes formulas for each MS/MS fragment

To incorporate MS/MS data, the MFG software first proposes a list of elemental compositions that correlate with the accurate mass of each product ion. Because MS/MS fragments (product ions) are lower in molecular weight than their precursors, there are typically fewer possible molecular formulas for each fragment than for the precursor ion. As shown in the middle highlighted column in Figure 6, however, some proposed formulas are better than others because the mass deviations are lower.

MS/MS Formula Details: Compound 3 C10H9N4O2SCl

m/z	Calc m/z	Formula	Difference (mDa)	Loss Formula
92.04972	92.04948	C6H6N	-0.25	C4H4N3O2SCl
92.04972	92.05285	C3H10NS	3.12	C7N3O2Cl
108.0442	108.04439	C6H6NO	0.19	C4H4N3OSCl
156.01109	156.01138	C6H6NO2S	0.29	C4H4N3Cl
156.01109	156.008	C9H2NO2	-3.08	CH8N3SCl

Figure 6. MFG software assigns formulas to both the MS/MS product ions and the neutral losses.

Step 2: Precursor compositions unrelated to the product compositions are eliminated

For a precursor molecular formula to be valid, the product ions in the MS/MS spectrum must contain portions of that formula. In other words, to qualify a precursor composition, MFG must be able to find at least one subset composition for some of the product ions (see Figure 7). After the MFG software generates all possible elemental compositions for all product ions, it uses them to judge the goodness of the tentative precursor compositions. The software rules out a candidate precursor composition if the algorithm fails to establish the compositional relationship for a large proportion of the products. Thus, the MFG software uses the MS/MS information to trim the list of precursor molecular formulas, making the list much easier to review.

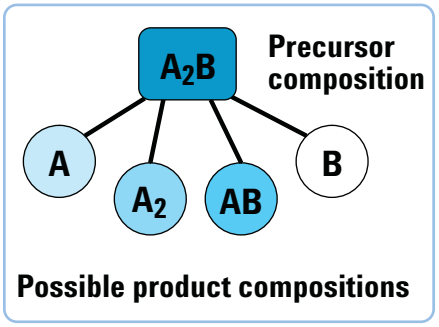

Figure 7. To make full use of MS/MS spectra, MFG searches for relationships between the tentative formulas for the precursor and product ions. If a precursor formula fails to explain a large number of the product ions, it is eliminated from the list.

Step 3: Proposed precursor formulas are scored and ranked based on their fit with the MS/MS results (coverage and mass error)

When MS/MS information is available, MFG scores the candidate molecular formulas based on their fit to the MS/MS data. MFG uses a coverage score to measure how much precursor-product relationship it can establish for a given precursor candidate composition. The coverage expresses how much of the total ion intensity in the MS/MS spectrum can be explained with the tentative molecular formula of the precursor ion. Coverage is a good indicator of how well the formula for the precursor ion fits the MS/MS spectrum. The coverage score is incorporated within the overall MS/MS score and it aids selection of the best molecular formula from the list of results.

The MS/MS score accounts for both the relative abundance of the product ions and how well each tentative product composition predicts the mass of the product ion. If the MFG algorithm cannot establish a precursor-product relationship for a more abundant product, it assigns a greater penalty than it does for a less abundant product ion. The algorithm also assigns a higher score when the tentative product composition better fits the experimental mass of the product ion. As with the MS data, the ranking for MS/MS data is relative. The candidate precursor composition with the best score receives a rank of 100, while the others are assigned proportionally lower rankings.

Step 4. A single overall score combines MS and MS/MS data

Finally, the MFG software combines the previously computed MS score and the MS/MS score into a single, overall score for each precursor molecular formula it reports. This combined score summarizes all the factors that went into the MFG calculations. The higher the score, the more likely a candidate composition is the correct formula. The overall result summary simplifies and accelerates the interpretation of results. Once the best formula has been selected, the software can automatically annotate the peaks in both the MS/MS and MS spectra with that formula.

Figure 8 shows the value that the MS/MS data adds to the calculation of molecular formulas. With the MS data alone (Figure 8a), the scores of the top two hits are less than two points apart, and the correct answer is second on the list. In this situation, it is difficult to unequivocally pick either the first or the second hit as the correct answer. With combined MS and MS/MS data (Figure 8b), the scores of the same top two hits are now 15 points apart, and the correct answer is first on the list. The addition of the MS/MS data provides another dimension of information that removes uncertainty and delivers better results.

a) MS data only—top two hits have similar scores and correct answer is second

m/z	Species	Formula	Abundance
588.30623	(M+NH4)+	C34H42N3O6	27374

Best	Formula	Score	Mass	Calc Mass	Difference (ppm)	Abs Diff (ppm)	DBE
☐	C30H34N8O4	100	570.2724	570.2703	-3.68	3.68	18
☑	C34H38N2O6	98.75	570.2724	570.27299	1.03	1.03	17
☐	C35H34N6O2	85.2	570.2724	570.27432	3.37	3.37	22

b) Combination of MS and MS/MS data—correct answer is first by a significant margin

m/z	Species	Formula	Abundance
588.30674	(M+NH4)+	C34H42N3O6	35749

Best	Formula	Score	Cross Score	Mass	Calc Mass	Difference (pp	Abs Diff (pp	MS Score	MS/MS Score	Coverage	DBE
☑	C34H38N2O6	100		570.27291	570.27299	0.13	0.13	97.86	100	100	17
☐	C30H34N8O4	85.96		570.27291	570.2703	-4.58	4.58	100	84.13	100	18
☐	C35H34N6O2	70.06		570.27291	570.27432	2.47	2.47	87.26	78.57	87.77	22

Figure 8. When both MS and MS/MS data are available, a single final score points to the best overall molecular formula. In this example, the addition of the MS/MS data (bottom) provides greater selectivity and brings the correct formula to the top of the list.

Conclusion—a superior approach to molecular formula generation

When using mass spectrometry to analyze samples containing unknowns, it is often necessary to derive elemental compositions (molecular formulas) for the unknowns based on the mass spectral data. As outlined in Figure 9, the Agilent molecular formula generation software uses a wide range of MS information, not just accurate-mass measurements, to produce a list of candidate molecular formulas that are ranked according to their relative probabilities. If molecular feature extraction locates multiple species from the same compound, the MFG software calculates a cross score that combines the scores of the related species for significantly greater accuracy. When MS/MS data are available, MFG uses that information to further reduce the list of potential molecular formulas and provides even greater confidence for high-scoring results. The MFG software saves users considerable time because it eliminates unlikely candidates and delivers relative ranking for the remaining candidates, which makes it easier to find the correct formulas.

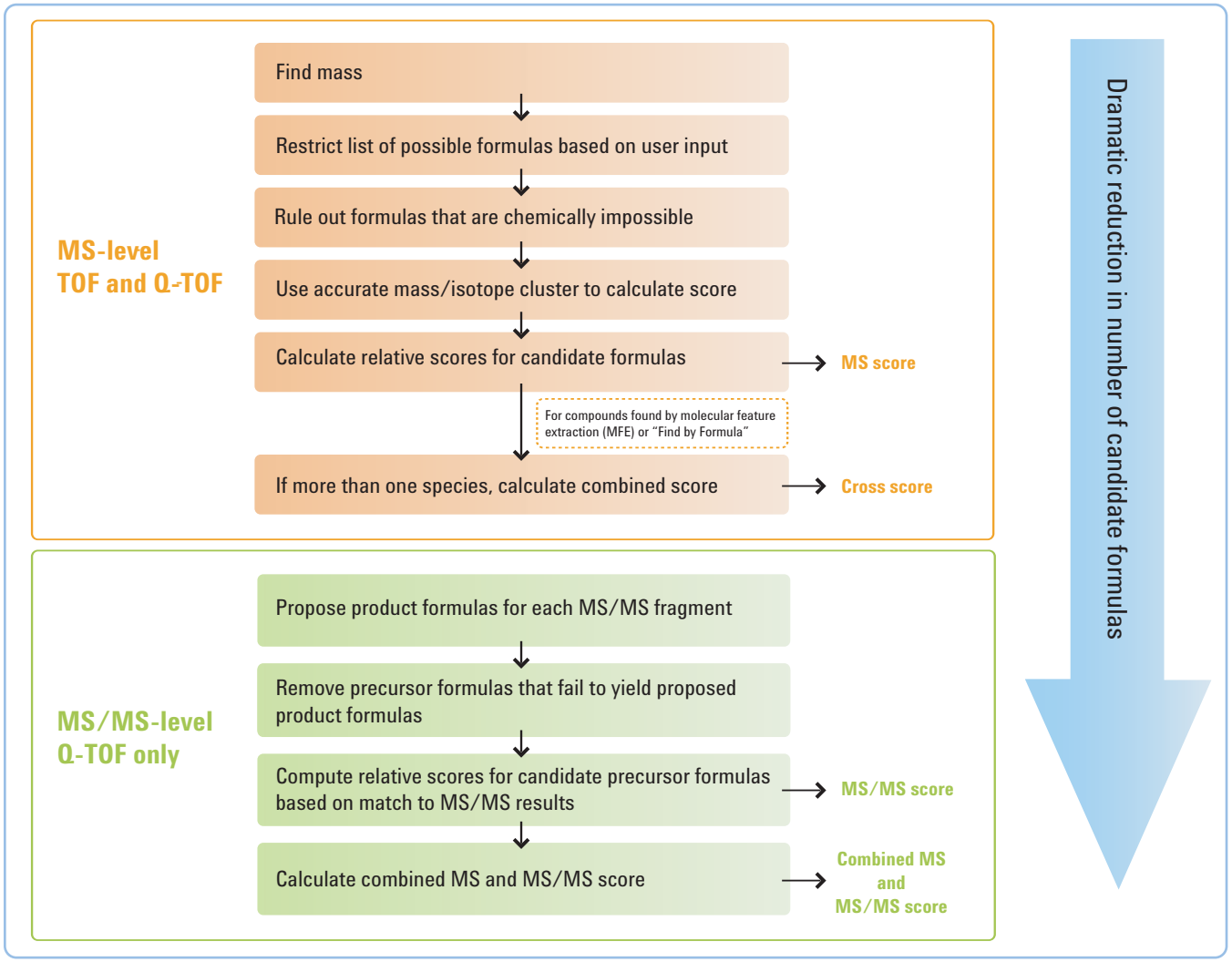

Figure 9. Each step of the MFG algorithm reduces the ambiguity in molecular formula assignment.

References

1. Kind, T. and Fiehn, O. "Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm" *BMC Bioinformatics* 7:234, **2006**.
2. Sparkman, O. D. 2006. *Mass Spec Desk Reference* (2nd Edition, p 64). Pittsburgh, Global View Publishing.
3. Kind, T. and Fiehn, O. 2007. "Nitrogen Rule": Metabolomics Fiehn Lab web site, University of California, Davis. http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/Nitrogen_Rule/

Authors

Ed Darland is a Senior Software Developer, Doug McIntyre is an Applications Chemist, Frank Kuhlmann is a Senior Scientist, and Xiangdong Li is a Software Developer, at Agilent Technologies in Santa Clara, California, U.S.A. David Weil is an Applications Scientist at Agilent Technologies in Schaumburg, Illinois, U.S.A

About Agilent Technologies

Agilent Technologies is a leading supplier of life science research systems that enable scientists to understand complex biological processes, determine disease mechanisms, and speed drug discovery. Engineered for sensitivity, reproducibility, and workflow productivity, Agilent's life science solutions include instrumentation, microfluidics, software, microarrays, consumables, and services for genomics, proteomics, and metabolomics applications.

Learn more:
www.agilent.com/chem/ms

Buy online:
www.agilent.com/chem/store

Find an Agilent customer center in your country:
www.agilent.com/chem/contactus

U.S. and Canada
1-800-227-9770

agilent_inquiries@agilent.com

Europe
info_agilent@agilent.com

Asia Pacific
adinquiry_aplsca@agilent.com

Printed in the U.S.A. January 4, 2008

5989-7409EN

Agilent Technologies